CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

ORCA – Online Research @ Cardiff

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository:https://orca.cardiff.ac.uk/id/eprint/169210/

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Wang, Xueyi, Li, Shancang and Iqbal, M. 2024. Artificial intelligence enabled microgrid power generation prediction. Open Computer Science

Publishers page:

Please note:
Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See http://orca.cf.ac.uk/policies.html for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.

Artificial Intelligence Enabled Microgrid Power Generation Prediction

Xueyi Wang, Shancang Li, Muddesar Iqbal

Abstract—The rapidly increasing photovoltaic technology is one of the key renewable energies expected to mitigate the impact of climate change and the energy crisis, which has been widely installed in the past few years. However, the variability of PV power generation creates different negative impacts on the electric grid systems and a resilient and predictable PV power generation is crucial to stabilize and secure grid operation and promote large-scale PV power integration. This paper proposed machine learning based short-term PV power generation forecasting techniques by using XGBoost, SARIMA, and LSTM algorithms. The experimental results demonstrated that the proposed resilient LSTM solution can accurately predict (around 90% R^2 and 0.028 $RMSE$) PV power generation with minimum input data.

Index Terms—PV generation, Solar panel, Machine learning, SARIMA, XGBoost, LSTM, Resilient

1 Introduction

The use of renewable energies, including wind, water, and solar energy, plays a significant role in mitigating the energy crisis and archiving net-zero emissions. Among these renewable energies, solar energy is the most stable and efficient renewable energy to generate electricity [1]. According to the United Nations Development Programme (UNDP), solar energy resource has a worldwide potential of 1,600 to 49,800 exajoules (4.4×10^{14} to $1.4 \times 10^{16} kWh$) per year [2]. Considering the huge potential and advantages of solar energy, solar photovoltaic (PV) panels have been widely used and the worldwide annual installation capacity reached about 512 gigawatts (GW) in 2018 [3]. In 2021, at least 175 GW PV panels were put into use worldwide, which made the total PV installed capacity reach at least 942 GW [3]. A science study in 2017 estimated that 1845 GW of PV systems could generate around 2646 TWh (Terawatt-hour) of electricity all around the world by 2030 [4].

In smart production, considerable attention has been dedicated to the meticulous study of smart grids, encompassing facets such as energy prediction and cybersecurity. [5]. In particular, many electricity storage systems like PV systems have been used to support electricity in individual houses and autonomous devices also design active generators [1], [6]. Implementing a photovoltaic (PV) renewable system within a microgrid offers significant potential for enhancing current energy consumption patterns. Utilizing solar energy, such a system can supplement traditional fossil fuel-based power sources, thereby reducing reliance on non-renewable resources and lowering carbon emissions. Additionally, PV systems can contribute to greater energy independence and resilience, especially in remote or off-grid areas, by providing a reliable and sustainable source of electricity. This integration fosters a more sustainable and environmentally friendly energy landscape while promoting economic savings.

However, the fluctuating and uncertain output in a PV system is always an issue and the concern of research [7], [8]. Facing the situation that the usage of PV panels sharply increasing and intermittency problems of PV generation in smart grid, an effective and resilient method estimating the electricity generation of the PV panel needs to be researched and developed eagerly [8].

Depending on the predicting time span, PV panel generation prediction can be divided into short-term forecasting (under one day ahead), medium-term forecasting (1 week to 1 month ahead), and long-term forecasting (one month to one year ahead) [9]. Short-term PV generation forecasting can be used in optimal storage capacity and power smoothing; medium-term PV generation forecasting helps in power system management and scheduling; long-term PV generation forecasting provides references for grid devices distribution and electricity transmission [9].

As one of the solutions, deep learning has shown excellent performance in solving renewable energy difficulties compared to machine learning because of the complexity and massive data in smart grid [10], [11], especially solar energy's randomness and intermittency problems [5]. As a part of the EU Project SuSTAINABLE [12], in Évora in Portugal, a very short-term solar energy forecasting system has been deployed via gradient boosting algorithm to establish an automating smart grid [12], [13]. The main contributions of this work can be summarised as follows:

1) Utilizing the Pearson Correlation model and XGBoost algorithm to clarify the importance of various features exited in the PV generation prediction model.

2) A comparative study on time-series algorithms, using previously observed PV generation time-series data and weather data like solar irradiation and air temperature. High-performance and resilient short-term PV forecasting frameworks were studied, which require the minimum amount of data.

- *Wang and Li are with the School of Computer Science and Informatics at Cardiff University, Cardiff, CF24 4AG, UK.*
Muddesar Iqbal is with the Renewable Energy Lab, College of Engineering, Prince Sultan University, Riyadh, Saudi Arabia.
Corresponding author: Shancang Li (Email: shancang.li@ieee.org).

2 Related works

As a part of smart production of the concept Industry 4.0, the smart grid has been meticulously studied recently [5]. Previously, soft computing and AI have been used for energy storage and management [14], [15], [16]. In reports [11], [17], [18], researchers studied the scenarios of various deep learning algorithms in smart grids. Deep learning algorithms help energy forecasting, security detection, and optimization for smart grid operation management, high resiliency facing contingency, and customer requests [11].

The solar PV system, as one of the main renewable energy resources in the smart grid, many works have been presented in the mathematical and system models in PV panel generation. Ma *et al.* demonstrated and compared several PV mathematical and equivalent circuit models, depending on the PV panel's physical structure: an ideal model based on Shockley theory; R_s one-diode model ($4-p$ model); Rs and R_p one-diode model; two-diode model [19]. Binayak proposed a PV generation prediction mathematical model based on solar irradiance and ambient temperature [20]. Aranzazu also mentioned in his work that the temperature and irradiance decide DC power generation [6]. However, Kim used all known weather data, like irradiance, ambient temperature, wind speed, and relative humidity as input to predict the PV power generation [7]. From the research works on PV generation forecasting, there are mainly three different types of approaching methods: 1) Machine learning approaches utilize multi-variable weather data 2) Statistical time series methods based on uni-variable data 3) Physical models use Numerical Weather Prediction (NWP) or satellite images to predict PV generation [21], [22].

For the reports focusing on the prediction of PV electricity generation using multi-variable weather data via machine learning methods, most works used the ANN algorithms model to predict PV electricity generation [22]. Stanley and his team presented a short-term prediction [23], which is predicting 20 minutes ahead using the MLP model and has 82% to 95% PV generation prediction accuracy. In [24], four different models were used to conduct short-term prediction of PV power generation, including Multilayer Perceptron (MLP), Elman Recurrent Neural Network (ENN), Radial Basis Function neural network (RBF), and Time Delayed Neural Network (TDNN). The MLP model performance on short-term prediction on PV electricity generation has 0.62 error, which predicts 2866973.48 Wh (Watt-hour) electricity and the true value is 2,849,201Wh [24].

In [25], EMD and SVM methods were used to analyze PV power generation. The SVM is a supervised machine learning model which is good at generalized linear classification. In the report, the author summarized that ANN and SVM are the two mainly used prediction methods. What is most important is that this report mentioned that the daily temperature is one of the important weather factors that affect the PV panel electricity output.

Some of the reports used time-series algorithms to predict PV generation. Kardakos and his team [26] utilized the seasonal ARIMA time-series algorithm to predict short-term PV generation and improved it by applying solar radiation derived from the Numerical Weather Prediction (NWP) model to the SARIMA's output. In [27], Maria Malvoni tried to predict one day ahead PV generation via the time-series algorithm Group Least Square Support Vector Machine (GLSSVM) combined with Least Square Vector Machines (LS-SVM) and Group Method of Data Handling (GMDH) algorithms dealing with multiple weather data. In [22], the author proved that ARIMA has better performance than ANN models in short-term PV generation prediction. In another study [21], the author compared SARIMA, SARIMAX, modified SARIMA, and ANN algorithm performances on short-term PV generation prediction.

There are a few reports focusing on building a pure physical predicting model. In [28], Sun described a method that took instant photos around PV panels to detect the cloud movement that can infect PV electricity output and then used a Convolutional Neural Network (CNN) to predict the PV electricity generation based on analyzing sky images.

There are many other reports focusing on the missing data processing in PV generation prediction. In Taeyoung Kim's report [29], they tried four different missing data imputation methods: LI, MI, KNN imputation, and Multivariate Imputation by Chain Equations (MICE). They claimed that using the KNN imputation method to handle missing data situations has the best performance, especially when the dataset missing over 20% data rates.

3 Methodology

3.1 System Model

To have a better understanding of the PV electricity generation process, figuring out the PV system structure and setting up configurations of PV panels are very important.

Sandia National Laboratories, which operates under the U.S. Department of Energy has published one of the related mathematical models is: the Plane of Array (POA) Model [30], which figures out the mathematical relation between the solar energy that PV panels absorbed, and the solar radiation is necessary. POA represents the PV panel surface and the irradiance cast on POA can be calculated by Eq. (1) [30].

$$E_{POA} = E_b + E_g + E_d \tag{1}$$

in which three main components of the POA irradiance, E_b is the POA beam component, E_g is the POA ground reflect component, E_d is POA sky diffuse component [30]. The overall solar irradiance reflected on the solar panel is the summation of the irradiance from direct sun irradiance, irradiance reflected from the ground, and irradiance diffused from the sky.

Moreover, the POA beam component is decided by Direct Normal Irradiance and the angle between the sun rays and the PV panel, which is determined by not only the solar azimuth, and zenith angles, but also the tilt, azimuth angles of the PV panel [31], in which DNI denotes Direct Normal Irradiance, AOI denotes Angle of Incidence.

$$E_b = DNI \cdot \cos(AOI) \tag{2}$$

POA formula's second component ground reflected irradiance can be calculated as the following equation. The main affected variables include global horizontal irradiance GHI, the ground surface reflectivity, which is also called ground

albedo ρ, and the tilt angle of the PV panel mentioned in the PV structure section θ_T [32]. Here, ρ denotes albedo, which is highly dependent on the ground color and material. When the surface is dark, $\rho \approx 0$; when the surface is bright white, $\rho \approx 1$ [33].

$$E_g = GHI \cdot \rho \cdot \frac{1-\cos(\theta_T)}{2} \tag{3}$$

Sky diffuse irradiance has several different theory models, like Isotropic Sky Diffuse Model [34], Hay and Davies Sky Diffuse Model [35], Reindl Sky Diffuse Model [36], and Perez Sky Diffuse Model [37]. This report used the simple Isotropic Model to demonstrate sky diffuse [38] [34].

$$E_d = DHI \cdot \frac{1+\cos(\theta_T)}{2} \tag{4}$$

Also, the Ground Horizontal Irradiance (GHI) could be calculated by the Diffuse Horizontal Irradiance (DHI), Direct Normal Irradiance (DNI), and solar zenith angles θ_Z following the Eq. (5) [39]. In [40], the author mentioned that the DHI value is around 10-20% of GHI value on a sunny day, however, when encountering a cloudy day, the DHI value is almost equal to the GHI value.

$$GHI = DHI + DNI \cdot \cos(\theta_Z) \tag{5}$$

As a result, the irradiance reflected on the PV panel in total could be calculated by putting Eqs. (2), (3), (4) into Eq. (1)

$$\begin{aligned} E_{POA} =& E_b + E_g + E_d \\ =& DNI \cdot \cos(AOI) + GHI \cdot \rho \cdot \frac{1-\cos(\theta_T)}{2} \\ &+ DHI \cdot \frac{1+\cos(\theta_T)}{2} \end{aligned} \tag{6}$$

in which the DNI can be derived from an absolute cavity radiometer; POA can be obtained by a pyranometer; AOI is mainly determined by solar azimuth θ_A, solar zenith θ_Z, and surface tilt angle θ_T. The albedo parameter ρ as mentioned before, denotes the reflectivity of the ground surface [31], [32], [33]. In practice, parameters θ_A, and θ_Z are related to the sun's position which is changed by date (time-related features). The tilt angle of the surface of PV panel θ_T will be dynamic as well if it is a solar tracking rack rather than a steady PV panel.

To provide an intuitive feeling of three components: beam, ground reflect, and sky diffuse, the numeric values were provided to understand each component's contributions to the POA model. The daily average global, beam, and diffuse irradiance component measurements in Kimberly, Idaho are 413, 481, and 132 W/m^2, respectively [41]. In another research [42], the author provided DNI, GHI, and DHI measurement values in Doha, Qatar, which are 200.4, 225.2, and 94.7 W/m^2, respectively, among half-year records.

To sum up, despite solar irradiance related weather factors DNI, and DHI, there are some time-related features (θ_A, θ_Z), some weather-related features (cloud cover, temperature), and other external environment features (ρ, θ_T) that existed in the PV generation model.

3.2 Predictive Framework

The PV generation forecasting process in this paper will follow the steps in the flow chart. After obtaining the original PV generation data and related weather data. Two datasets will first be pre-processed to get rid of the missing and wrong data and aligned by their timestamps afterward. Then, the processed time, PV generation data, and weather data will be fed into the XGBoost model to decide on the most important and effective input features for the proposed resilient model. Lastly, use SARIMA and LSTM algorithms to predict PV generation in the short term, comparing models' performance under different input features. The predictive framework flowchart is illustrated in Figure 1.

3.3 XGBoost Model

CART tree algorithm, which is short for classical classification and regression tree, is the title for both trees: Classification Tree (the prediction results are types) and regression Tree (the prediction results are numeral). CART tree is a non-parametric decision tree algorithm [43]. CART algorithm first puts the input dataset in the root node; then splits sub-nodes from the root node according to the attribute, and decides the best homogeneity for the threshold; the splitting process keeps going until a pure subset or meets the maximum node depth, the final node called leaf node is the one who holds the decision [8], [43]. This whole iteration process will provide the relative best-fit model.

The eXtreme Gradient Boosting (XGBoost) is a supervised, improved gradient-boosted trees algorithm, which is integrated by many CART trees. The XGBoost's regularized objective function $L(\phi)$ includes training loss and regularization term [44].

$$L(\phi) = \sum_i L(\hat{y_i}, y_i) + \sum_k \Omega(f_k) \tag{7}$$

The differentiable convex training loss (L) measures the predictive ability of the model by comparing the difference between prediction value $\hat{y_i}$ and measurement value y_i. Where f_k denotes independent tree structure and leaf weights.

Regularization term (Ω) refers to the part that controls the model complexity that prevents the model from overfitting, which can be demonstrated as [44], [45].

$$\Omega(f) = \frac{1}{2} \cdot \lambda \cdot ||w||^2 + \gamma \cdot T \tag{8}$$

in which Ω defines the complexity of the tree f; λ determines the strength of the regularisation; γ is a penalise nodes constant, when it is greater than 1; T denotes the number of leaves in the tree.

XGBoost algorithm optimises the function by keep adding new trees to simulate the residuals from the last prediction rather than using methods in Euclidean space [43], [44]. The purpose of the model is to find f_p value to optimise the objective function by searching the smallest score [44]. In dataset $D = (x_i, y_i)$, (x_i denotes examples,

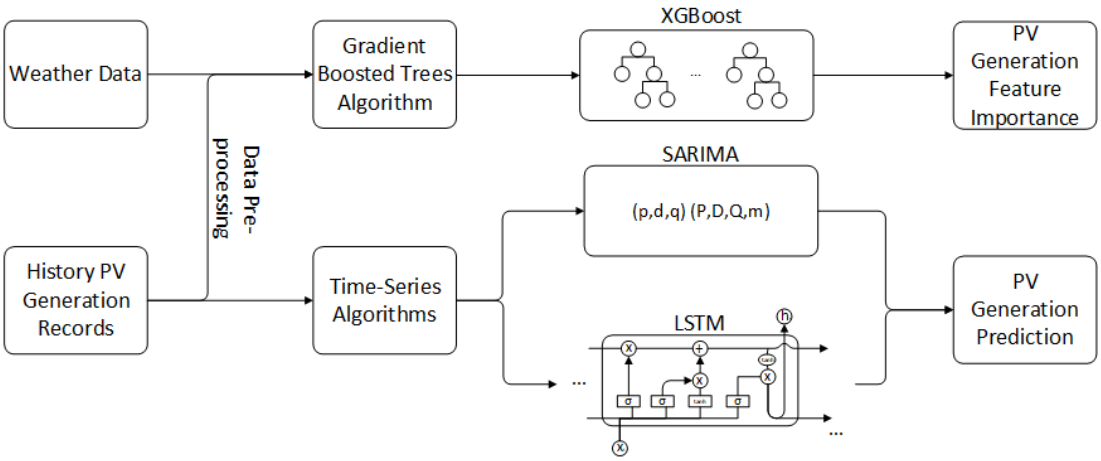

Fig. 1. PV Power Generation Prediction Framework Flowchart

y_i denotes features), the objective function $L^{(p)}$ can be presented as [43], [44].

$$\mathcal{L}(D, f_p) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(p-1)} + f_t(x_i)) + \Omega(f_p) \tag{9}$$

in which $\hat{y}_i^{(p)}$ refers to the prediction value on p-th iteration of i-th instance [44].

To simplify the objective function, the constant terms $\sum_{i=1}^{n} l(y_i, \hat{y}^{(p-1)})$ could be removed [44]. By applying one of the common training loss functions mean squared error (MSE) and regularisation term $\Omega(f_p)$, the objective function can be simplified as follows [43]

$$L^{(p)} = -\frac{1}{2}\sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma \cdot T \tag{10}$$

After the process of second-order approximation, this objective Eq. (9) sums the first and second input gradient statistics g_i and h_i up together as G_j and H_j; then calculates the score according to the formula, which indicates the quality of the tree structure [43].

This paper used the XGBoost model purely for time-series PV data predicting, which separates the various time factors (like months, weeks, days, and hours) from PV generation data and stores them in a tree. At last, predict future PV generation based on the summarised time features relationship.

3.4 Seasonal ARIMA Model

Auto-regressive Integrated Moving Average (ARIMA) is one of the effective univariate time series algorithms. As an extended version, Seasonal Autoregressive Integrated Moving Average (SARIMA) algorithm supports both autoregressive and moving average functions [46], which means SARIMA would identify seasonal changing input data and make better predictions compared to ARIMA. Seasonal changing data refers to the training data value changes due to seasonal factors. Accordingly, we use the SARIMA model to train and predict the PV generation value. The SARIMA model can be formed as $SARIMA((p,d,q),(P,D,Q,m))$ in which (p,d,q) represents the non-seasonal feature from the data; (P,D,Q,m) represents the seasonal feature. In more detail, p represents trend auto-regression order; d represents trend difference order; q moving average order; P represents seasonal auto-regressive order; D represents seasonal difference order; Q represents seasonal moving average order; m represents the number of time steps for a seasonal period [46]. The SARIMA model can be formulated into Eq. (11)

$$\begin{aligned} y_t =& c + \sum_{n=1}^{p} \phi_n y_{t-n} + \sum_{n=1}^{q} \theta_n \epsilon_{t-n} + \sum_{n=1}^{P} \eta_n y_{t-mn} \\ &+ \sum_{n=1}^{Q} \omega_n \epsilon_{t-mn} + \epsilon_t \end{aligned} \tag{11}$$

where y_t denotes the value of the time series at time t, ϕ denotes the coefficients of the auto-regressive, θ denotes coefficients of the auto regressive forecast errors; ϵ denotes the moving average forecast error; η denotes coefficients of seasonal forecast errors; ω denotes coefficients of seasonal auto regressive. The enumeration function was used to list all (p,d,q) and (P,D,Q,m) combinations to look for the best model fit, which is judged by the lowest Akaike Information Criterion (AIC) value [47].

3.5 Long Short-Term Memory Network (LSTM)

LSTM is one type of the RNN (Recurrent Neural Network) especially overcome exploding and vanishing gradient problems during long-term dependencies, which has recurrent neurons to process the input data through the activation and formulate an output to the next neuron [48]. The LSTM is good at solving the sequence problems because of the feedforward network from the last training [49], which no longer suffers from Simple Recurrent Networks [50]. A typical vanilla LSTM model architecture can be demonstrated as multiple memory blocks shown in Figure. 2 [48], [51].

In this special memory block, there are three essential multiplicative units: input gate, output gate, and forget gate, which always use sigmoid as their non-linear activation

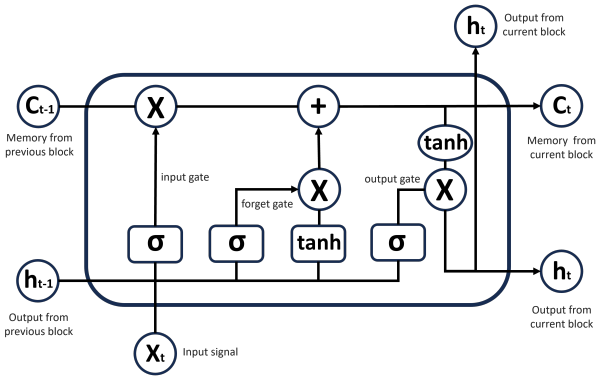

Fig. 2. LSTM memory block demonstration

function [48]. This memory block first takes previous memory c_{t-1}, output h_{t-1}, and input signal X_t as input to feed through input gate [48], [52], [53], [54].

$$i_t = \sigma(W_{xi}X_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (12)$$

Secondly in forget gate, the memory block decides what information will be forgotten based on the input signal X_t, memory c_{t-1} and output h_{t-1} from the previous block [48], [52], [53], [54].

$$f_t = \sigma(W_{xf}X_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \quad (13)$$

Third step is generating output combining input signal X_t, memory c_{t-1}, and output h_{t-1} from last iteration [48], [52], [53], [54].

$$o_t = \sigma(W_{xo}X_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \quad (14)$$

This memory block iteration will carry on when more input data has been fed through the LSTM model. LSTM networks there would be another loop in the model, which will provide the feedback value as an input vector from an output of the network to the input of the network [48].

4 Evaluation

4.1 Data Preparation

The quality of the input dataset may affect the training model's performance and accuracy [55], which means two data pre-processing steps: data cleaning and filtering missing data are necessary [56]. After the data pre-processing steps, the data normalised step is required to reduce the noise and normalise the dataset.

In this work, for the XGBoost, SARIMA, and LSTM algorithms, 1 year London area's PV electricity generation record from Sheffield open-source PV live data (London area) and weather data were from MIDAS UK open weather data [1]. The data were from Heathrow station, and both of the datasets (PV and weather) start from 2020/01/01 to 2020/12/27, recording in every 60 minutes. Due to the synchronisation failure, and misoperations, there are some repeated data or vacant data. In the data pre-processing step, we removed the repeat data according to the date and time, then left the vacant records (NaN) as noise. As a result, the PV generation dataset has in total of 8657 records after the pre-processing step.

The PV generation data itself is recorded in Megawatts, and hourly generation records can reach up to thousands of megawatts, sometimes even more than 5 thousand Megawatts. However, at night, due to lack of solar irradiance, the PV generation equals 0 mostly. The large periodical difference in the dataset will not benefit the learning process, which means that the data normalisation process is necessary as well. MinMaxScaler function was used to scale all the PV generation data under [0,1] scale. The weather data air temperature are stored in the degree $Celsius$; the global solar irradiance mount is stored in KJ/m^2.

The Pearson correlation coefficient model was used to analyse the associations behind the feature data. It is a normalisation evaluation method of the covariance of two features, which reviews the strength and direction of linear correlation between them [57]. On the one hand, the various weather data were studied including time-related features' correlations with PV electricity generation, which can be illustrated in figure 3.

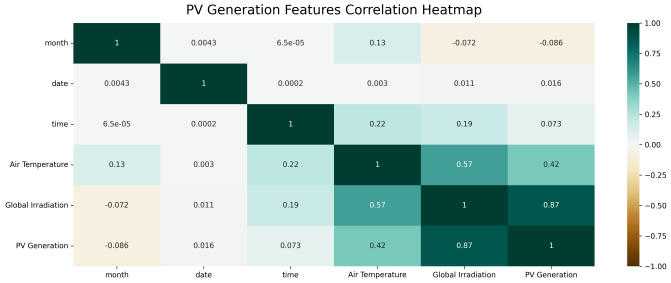

Fig. 3. Heatmap of PV Generation Features Correlation

Analysing the correlation between features from the dataset, we can see that pure time features like hours, date, and month do not have much correlation with hourly PV generation because their correlation values with PV generation are close to 0. However, previous records of global irradiance and air temperature have a very high correlation to the next hour's PV generation, which means they could be the main features that contribute to the prediction model.

On the other hand, the correlation between history PV generation (24 hours before) and current PV generation was also studied. Because the POA model is very complex and contains various times, weather, and other features we cannot easily have access to all of them. In contrast, the previous PV generation data are the accurate value simulated by all the factors. The correlation heatmap between history and current PV generation value is shown in Figure 4. Where P_24 refers to PV generation from 24 hours ago, P_0 refers to the PV generation in the future one hour, which needs to be predicted.

From Figure 4, P_1 to P_3 (previous 3 hours) and P_21 to P_24 (yesterday 3 hours after) PV generation has a high correlation (above 0.6) to the future one-hour PV generation prediction. The other hours of PV generation also have some associations with the future PV generation data.

1. catalogue.ceda.ac.uk/uuid/dbd451271eb04662beade68da43546e1

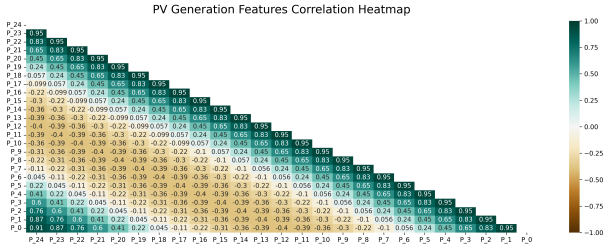

Fig. 4. Heatmap of PV Generation Features Correlation

To sum up, we used a total of 8752 records (from 2020/01/01 0:00 to 2020/12/21 23:00) to train the XGBoost model. The dataset is developed in 6 columns: Month, Date, Time, Air Temperature, Global Irradiance, and PV Generation. By applying different feature combinations to the algorithms to receive a high-performance model that requires minimum data. The testing dataset will be from 6 days of records (from 2020/12/22 0:00 to 2020/12/27 23:00).

4.2 XGBoost

In order to find the ideal prediction model, this work first used the XGBoost model to verify and analyse the hypothesis we concluded from the first Pearson correlation model that solar irradiance and air temperature are the core features to predict short-term PV generation. The XGBoost model predicts future PV generation based on the previous 1-hour weather data and the time data without the data normalisation process to keep the original data features. As a result, PV generation prediction uses megawatts as its measurement unit. PV generation data's relationships in the XGBoost model among different features can be illustrated in Figure 5.

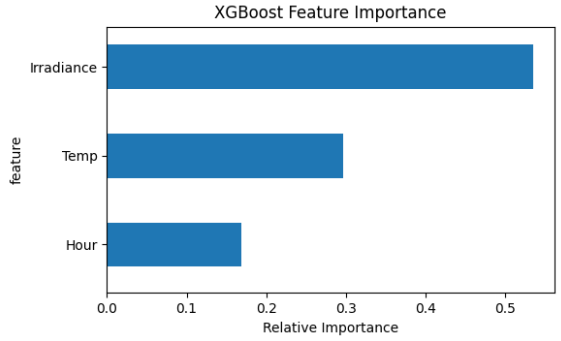

Fig. 5. XGBoost time features analysis in PV generation data

XGBoost feature importances function utilises F score that indicates each feature's percentage weight over the weights of all features. As figure 5 shows, in XGBoost solar irradiance has over 50% importance and air temperature has around 30% importance for this dataset.

For the hyperparameter optimization part, we still used the grid search method to automatically test the best fit of parameters for the XGBoost model. The XGBoost prediction model fit is shown in figure 6, which is an XGBoost one-hour ahead prediction model using previous hours, solar irradiance, and air temperature as the training dataset.

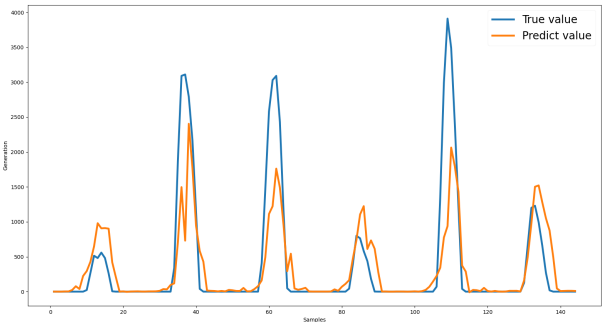

Fig. 6. XGBoost PV generation prediction

From the figure, we can see that the XGBoost model only using previous one-hour solar irradiance, air temperature, and hours time data shows low prediction accuracy on the summit PV generation of the day. The average R^2 score is around 0.7; the RMSE value is around 467 (megawatts). This model's performance is not very well for just using previous weather data and time data as its input. As a result, we decided to use previous PV generation data as input instead of using previous solar irradiance and air temperature. Because previous weather data are more related to previous PV generation instead of PV generation in the future. Also, previous PV generation is always accurate and more related to the predicting panel (containing more complex features) comparing to the local weather data.

4.3 SARIMA

To confirm the correlation between history and current PV generation, we utilized the time-series algorithm SARIMA to predict the current generation value using the previous generation value analysed by date and time. The auto-regression part of the SARIMA model measures the dependency by past observations, which will use the previous 8700 data records as the training dataset. Also, from observation of the PV generation value, the value repeats in a similar tendency loop for 24 hours of solar movement, which means the PV generation value can be set in a 24-hour (24 records) seasonal period in a SARIMA model. The test data will be 6 days starting from 2020/12/22 to 2020/12/27.

Our goal is to find the best-fit value for SARIMA $(p, d, q)(P, D, Q, m)$ to optimize the interest metric. We used the hyperparameter optimization method, Grid Search, to generate and test different combinations of variables p, d, q, and then evaluate the model accuracy by comparing their Akaike Information Criterion (AIC) value. AIC value is a common model selection theory, which will gradually minimize the mean squared error of prediction to find the best fit [58]. Normally, the lower AIC value is considered to be a better model fit.

After the grid search process, the best hyperparameters that fit the SARIMA model are $(1, 1, 1)(1, 1, 1, 24)$ with the lowest AIC score of 262.36. The SARIMA short-term PV prediction can be illustrated in Figure 7.

The SARIMA model's coefficients summary table is listed in Table 1:

in which ar.L1 and ma.L1 rows denote autoregressive (AR) and moving average (MA) coefficients for the non-

	coef	std err	z	$P > \|z\|$
ar.L1	0.4708	0.012	38.169	0.000
ma.L1	0.1012	0.013	7.847	0.000
ar.S.L24	0.1128	0.008	14.574	0.000
ma.S.L24	-0.8901	0.004	-217.983	0.000
sigma2	0.0599	0.001	116.467	0.000

TABLE 1
SARIMA model summary on PV generation prediction

seasonal component of the model, respectively. Similarly, "ar.S.L24" and "ma.S.L24" denote the AR and MA coefficients for the seasonal component of the model, where the seasonality is specified as 24 hours. The sigma2 (sigma square) column denotes the variance of residual values; the coef column denotes weights of each feature; the std err column denotes standard error. The significance of each coefficient is assessed using the z-test, with the associated p-values provided in the "$P > |z|$" column. A p-value less than the conventional threshold of 0.05 indicates that the corresponding coefficient is statistically significant, suggesting that it has a significant association with the predicted PV generation.

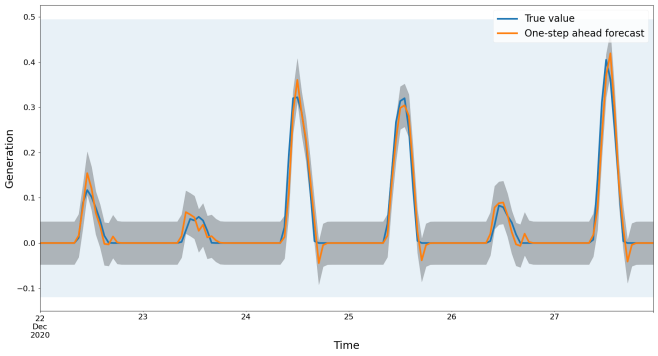

Fig. 7. SARIMA PV generation prediction

As Figure 7 shows, the SARIMA model has relatively high accuracy and confidence (average MAE is around 0.0097, average RMSE is around 0.0196) on short-term PV generation predicting. The grey shadow in the figure is the confidence bounds of the prediction model, which refers to the uncertainty of each step of prediction. However, we still can see there are imprecise predictions during the end of the day and the summit generation in a day.

4.4 LSTM

The LSTM model was first used to conduct one-step ahead PV generation prediction as a comparison group with the SARIMA model. We set the look-back window as 24 (last 1-day records) so that the LSTM algorithm will take previous 1-day records to predict the value. The LSTM model contains 50 neurons; mean square error as loss function; and the adam optimizer. LSTM model's average score (R^2) is around 0.96, average RMSE is around 0.2. The best LSTM one-step ahead model we conduct can be illustrated in Figure 8:

LSTM one hour ahead short-term prediction using 24 hours history records has a relatively good performance, even though sometimes it is still limited in predicting

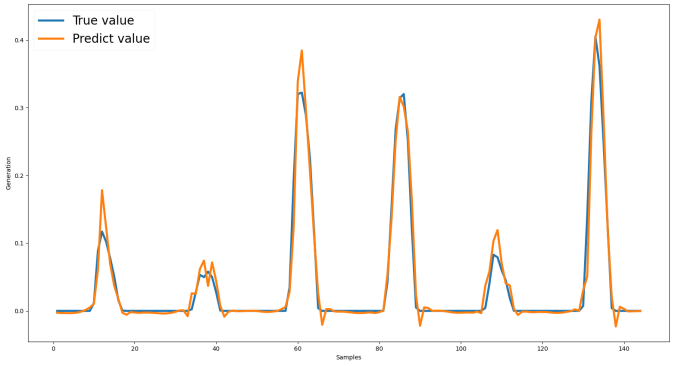

Fig. 8. LSTM PV generation prediction

the summit PV generation value of the day and some other turning points. Then, to find a resilient and high-performance model, we also added the weather data (solar irradiance and air temperature) and tested other look-back window sizes on the LSTM model to reduce the input feature amount. Different input feature combinations' requirements and performance are listed in table 1.

4.5 Result Comparison

We researched the time-series PV generation algorithms SARIMA and LSTM, which use previous PV generation data to predict future generations. We used 4 common evaluation methods R square score (R^2), mean absolute error (MAE), mean square error (MSE), and root mean squared error (RMSE).

R^2 score, which also means coefficient of determination, describes the accuracy of dependent variable changes according to the prediction of independent variables. In other words, it explains how well the data fit the model and where $R^2 = 1$ represents the perfect fit. Where y_i denotes the true value of the dataset, $\hat{y}_i$ denotes the predicted value, and $\overline{y}_i$ denotes the mean value.

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y}_i)^2} \tag{15}$$

MAE represents the average of absolute residuals between the prediction and true value, which can be calculated as:

$$R_{MAE} = \frac{1}{n} \sum_i |y_i - \hat{y}_i| \tag{16}$$

MSE measures the average square errors between predicted and true values. Compared to MAE, it measures the variance of residuals rather than the average of residuals. It can be calculated as:

$$R_{MSE} = \frac{1}{n} \sum_i (y_i - \hat{y}_i)^2 \tag{17}$$

RMSE is the square root of the MSE value, which measures the standard deviation of residuals and can be calculated as:

TABLE 2
Comparison Between XGBoost, SARIMA, and LSTM Models

	SARIMA*	LSTM*'24	LSTM*'3	LSTM*'2	LSTM* 24	LSTM* 3	LSTM* 2
R^2	0.9501	0.9563	0.9103	0.9179	0.9502	0.9022	0.8923
R_{MAE}	0.0097	0.0097	0.0155	0.0145	0.0103	0.0160	0.0185
R_{MSE}	0.0003	0.0003	0.0006	0.0006	0.0003	0.0007	0.0008
R_{RMSE}	0.0196	0.0183	0.0262	0.0251	0.0195	0.0278	0.0287

$$R_{RMSE} = \sqrt{\frac{1}{n}\sum_{i}(y_i - \hat{y}_i)^2} \tag{18}$$

In the algorithm results comparison table, * denotes the dataset with normalization; ' denotes to the dataset with weather data (solar irradiance and air temperature). The number after LSTM denotes the look-back number. For example, LSTM*' 5 refers to the LSTM model utilizing the normalization dataset and weather data, using the previous 5 hours of PV generation data as a look-back window. For each of the models, we used an average score among 20 times model fitting.

From the comparison, we can see that the LSTM model using 24 history PV generation records and weather data does have excellent performance (average R^2 score can reach 0.9563, and average RMSE is around 0.0195). However, various input features reduce the resilience and robustness of the model. When there are not that much PV generation data available or some of them are not accurate because of cyber attacks, the model will not perform or make an accurate prediction in that circumstance. Instead, we studied various LSTM models that require minimum input data to guarantee the model's resilience. In the table, LSTM models only use the previous 2 records and 3 records of PV generation data as input and have relatively high performance (with an average R^2 score of around 0.90 and an average RMSE score is around 0.026).

However, when previous solar irradiance, air temperature, or other weather data are available combined with previous PV generation data, using multi-variables LSTM algorithms will have a slightly better performance in short-term prediction (by comparing LSTM*' models with LSTM* models).

5 Conclusion

This research proposed resilient machine learning models for short-term photovoltaic (PV) power generation forecasting using XGBoost, SARIMA, and LSTM algorithms. The LSTM model leveraging 24 hours of historical PV data and weather information exhibited excellent performance, with an average R^2 of 0.9563 and $RMSE$ of 0.0183.

To enhance prediction resilience, LSTM models requiring only 2-3 hours of prior PV generation records were investigated. Though slightly underperforming the 24-hour model, they achieved competitive average R^2 scores of around 0.90 and RMSE values of approximately 0.026. When supplemented with weather data, these multi-variable LSTM models marginally improved. The time-series SARIMA model relying solely on historical PV data also demonstrated high accuracy (R^2 of 0.9501) and resilience.

The proposed models offer flexibility to choose the appropriate input feature based on available data, resilience, and accuracy requirements. The 24-hour LSTM model can maximize accuracy given abundant historical data, while the minimal input LSTM models prioritize resilience over marginal performance losses when data is limited or cyber-attacks are a concern. Overall, this work advances reliable, resilient PV forecasting to facilitate renewable energy integration.

Acknowledgements

This work was partially supported by the research grants funded by the Research, Development, and Innovation Authority (RDIA), Saudi Arabia, with grant number (13354-psu-2023-PSNU-R-3-1-EI-). The authors would like to acknowledge the support of Prince Sultan University for paying the Article Processing Charges (APC) of this publication.

References

[1] H. Kanchev, D. Lu, F. Colas, V. Lazarov, and B. Francois, "Energy management and operational planning of a microgrid with a pv-based active generator for smart grid applications," *IEEE Transactions on Industrial Electronics*, vol. 58, no. 10, pp. 4583–4592, 2011.

[2] U. N. D. Programme, *World Energy Assessment: Energy and the Challenge of Sustainability*, 1st ed. New York, NY 10017: United Nations Development Programme, 2000.

[3] A. Detollenaere and G. Masson, "Snapshot of global pv markets 2022 task 1 strategic pv analysis and outreach," pp. 8–10, 2022.

[4] N. M. Haegel, R. Margolis, T. Buonassisi, D. Feldman, A. Froitzheim, R. Garabedian, M. Green, S. Glunz, H.-M. Henning, B. Holder, I. Kaizuka, B. Kroposki, K. Matsubara, S. Niki, K. Sakurai, R. A. Schindler, W. Tumas, E. R. Weber, G. Wilson, M. Woodhouse, and S. Kurtz, "Terawatt-scale photovoltaics: Trajectories and challenges," *Science*, vol. 356, no. 6334, pp. 141–143, 2017.

[5] R. Cioffi, M. Travaglioni, G. Piscitelli, A. Petrillo, and F. De Felice, "Artificial intelligence and machine learning applications in smart production: Progress, trends, and directions," *Sustainability*, vol. 12, no. 2, 2020. [Online]. Available: https://www.mdpi.com/2071-1050/12/2/492

[6] A. D. Martin, J. M. Cano, J. F. A. Silva, and J. R. Vázquez, "Backstepping control of smart grid-connected distributed photovoltaic power supplies for telecom equipment," *IEEE Transactions on Energy Conversion*, vol. 30, no. 4, pp. 1496–1504, 2015.

[7] G. G. Kim, J. H. Choi, S. Y. Park, B. G. Bhang, W. J. Nam, H. L. Cha, N. Park, and H.-K. Ahn, "Prediction model for pv performance with correlation analysis of environmental variables," *IEEE Journal of Photovoltaics*, vol. 9, no. 3, pp. 832–841, 2019.

[8] S. Ferlito, G. Adinolfi, and G. Graditi, "Comparative analysis of data-driven methods online and offline trained to the forecasting of grid-connected photovoltaic plant production," *Applied Energy*, vol. 205, pp. 116–129, 2017.

[9] M. N. Akhter, S. Mekhilef, H. Mokhlis, and N. Mohamed Shah, "Review on forecasting of photovoltaic power generation based on machine learning and metaheuristic techniques," *IET Renewable Power Generation*, vol. 13, no. 7, pp. 1009–1023, 2019.

[10] 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. 175 871–175 880, 2020.

[11] M. Massaoudi, H. Abu-Rub, S. S. Refaat, I. Chihi, and F. S. Oueslati, "Deep learning in smart grid technology: A review of recent advancements and future prospects," *IEEE Access*, vol. 9, pp. 54 558–54 578, 2021.

[12] R. Bessa, A. Trindade, C. S. Silva, and V. Miranda, "Probabilistic solar power forecasting in smart grids using distributed information," *International Journal of Electrical Power and Energy Systems*, vol. 72, pp. 16–23, 2015.

[13] C. Gouveia, D. Rua, F. Soares, C. Moreira, P. Matos, and J. P. Lopes, "Development and implementation of portuguese smart distribution system," *Electric Power Systems Research*, vol. 120, pp. 150–162, 2015.

[14] N. Gowtham, V. Prema, M. F. Elmorshedy, M. Bhaskar, and D. J. Almakhles, "A power aware long short-term memory with deep brief network based microgrid framework to maintain sustainable energy management and load balancing," *Distributed Generation & Alternative Energy Journal*, pp. 1–26, 2024.

[15] R. Abdulkader, H. M. Ghanimi, P. Dadheech, M. Alharbi, W. El-Shafai, M. M. Fouda, M. H. Aly, D. Swaminathan, and S. Sengan, "Soft computing in smart grid with decentralized generation and renewable energy storage system planning," *Energies*, vol. 16, no. 6, p. 2655, 2023.

[16] A. Rajagopalan, D. Swaminathan, M. Alharbi, S. Sengan, O. D. Montoya, W. El-Shafai, M. M. Fouda, and M. H. Aly, "Modernized planning of smart grid based on distributed power generations and energy storage systems using soft computing methods," *Energies*, vol. 15, no. 23, 2022. [Online]. Available: https://www.mdpi.com/1996-1073/15/23/8889

[17] O. A. Omitaomu and H. Niu, "Artificial intelligence techniques in smart grid: A survey," *Smart Cities*, vol. 4, no. 2, 4 2021.

[18] M. Pérez-Ortiz, S. Jiménez-Fernández, P. A. Gutiérrez, E. Alexandre, C. Hervás-Martínez, and S. Salcedo-Sanz, "A review of classification problems and algorithms in renewable energy applications," *Energies*, vol. 9, no. 8, 2016. [Online]. Available: https://www.mdpi.com/1996-1073/9/8/607

[19] T. Ma, H. Yang, and L. Lu, "Solar photovoltaic system modeling and performance prediction," *Renewable and Sustainable Energy Reviews*, vol. 36, pp. 304–315, 2014.

[20] B. Bhandari, S. R. Poudel, K.-T. Lee, and S.-H. Ahn, "Mathematical modeling of hybrid renewable energy system: A review on small hydro-solar-wind power generation." *Int. J. of Precis. Eng. and Manuf.-Green Tech*, vol. 1, p. 157–173, 2014.

[21] S. I. Vagropoulos, G. I. Chouliaras, E. G. Kardakos, C. K. Simoglou, and A. G. Bakirtzis, "Comparison of sarimax, sarima, modified sarima and ann-based models for short-term pv generation forecasting," in *2016 IEEE International Energy Conference (ENERGYCON)*, 2016, pp. 1–6.

[22] A. Álvarez Gallegos, L. Fara, A. Diaconu, D. Craciunescu, and S. Fara, "Forecasting of energy production for photovoltaic systems based on arima and ann advanced models," in *International Journal of Photoenergy*, 2021.

[23] S. K. Chow, E. W. Lee, and D. H. Li, "Short-term prediction of photovoltaic energy generation by intelligent approach," *Energy and Buildings*, vol. 55, pp. 660–667, 2012.

[24] L. A. Fernandez-Jimenez, A. Muñoz-Jimenez, A. Falces, M. Mendoza-Villena, E. Garcia-Garrido, P. M. Lara-Santillan, E. Zorzano-Alba, and P. J. Zorzano-Santamaria, "Short-term power forecasting system for photovoltaic plants," *Renewable Energy*, vol. 44, pp. 311–317, 2012.

[25] M. Mao, W. Gong, and L. Chang, "Short-term photovoltaic output forecasting model for economic dispatch of power system incorporating large-scale photovoltaic plant," *2013 IEEE Energy Conversion Congress and Exposition, ECCE 2013*, pp. 4540–4545, 09 2013.

[26] E. G. Kardakos, M. C. Alexiadis, S. I. Vagropoulos, C. K. Simoglou, P. N. Biskas, and A. G. Bakirtzis, "Application of time series and artificial neural network models in short-term forecasting of pv power generation," in *2013 48th International Universities' Power Engineering Conference (UPEC)*, 2013, pp. 1–6.

[27] M. Malvoni, M. G. De Giorgi, and P. M. Congedo, "Forecasting of pv power generation using weather input data-preprocessing techniques," *Energy Procedia*, vol. 126, pp. 651–658, 2017.

[28] Y. Sun, V. Venugopal, and A. R. Brandt, "Convolutional neural network for short-term solar panel output prediction," in *2018 IEEE 7th World Conference on Photovoltaic Energy Conversion (WCPEC) (A Joint Conference of 45th IEEE PVSC, 28th PVSEC and 34th EU PVSEC)*, 2018, pp. 2357–2361.

[29] T. Kim, W. Ko, and J. Kim, "Analysis and impact evaluation of missing data imputation in day-ahead pv generation forecasting," *Applied Sciences*, vol. 9, no. 1, 2019.

[30] Q. Sun, K. Lin, C. Si, Y. Xu, S. Li, and P. Gope, "A secure and anonymous communicate scheme over the internet of things," *ACM Transactions on Sensor Networks (TOSN)*, vol. 18, no. 3, pp. 1–21, 2022.

[31] X. Wang, S. Li, and M. Iqbal, "Live power generation predictions via ai-driven resilient systems in smart microgrids," *IEEE Transactions on Consumer Electronics*, 2024.

[32] Sandia National Laboratories, "Poa ground reflected." [Online]. Available: https://pvpmc.sandia.gov/modeling-steps/1-weather-design-inputs/plane-of-array-poa-irradiance/calculating-poa-irradiance/poa-ground-reflected/

[33] Y. Liu and S. Li, "A review of hybrid cyber threats modelling and detection using artificial intelligence in iiot," *Computer Modeling in Engineering & Sciences*, 2023.

[34] P. Loutzenhiser, H. Manz, C. Felsmann, P. Strachan, T. Frank, and G. Maxwell, "Empirical validation of models to compute solar irradiance on inclined surfaces for building energy simulation," *Solar Energy*, vol. 81, no. 2, pp. 254–267, 2007.

[35] J. E. Hay, "Calculating solar radiation for inclined surfaces: Practical approaches," *Renewable Energy*, vol. 3, no. 4, pp. 373–380, 1993.

[36] D. Reindl, W. Beckman, and J. Duffie, "Evaluation of hourly tilted surface radiation models," *Solar Energy*, vol. 45, no. 1, pp. 9–17, 1990.

[37] R. Perez, P. Ineichen, R. Seals, J. Michalsky, and R. Stewart, "Modeling daylight availability and irradiance components from direct and global irradiance," *Solar Energy*, vol. 44, no. 5, pp. 271–289, 1990.

[38] M. Lave, W. Hayes, A. Pohl, and C. W. Hansen, "Evaluation of global horizontal irradiance to plane-of-array irradiance models at locations across the united states," *IEEE Journal of Photovoltaics*, vol. 5, no. 2, pp. 597–606, 2015.

[39] Sandia National Laboratories, "Global horizontal irradiance." [Online]. Available: https://pvpmc.sandia.gov/modeling-steps/1-weather-design-inputs/irradiance-and-insolation-2/global-horizontal-irradiance/

[40] F. Vignola, "Ghi correlations with dhi and dni and the effects of cloudiness on one-minute data," 06 2012.

[41] F. Vignola, P. Harlan, R. Perez, and M. Kmiecik, "Analysis of satellite derived beam and global solar radiation data," *Solar Energy*, vol. 81, no. 6, pp. 768–772, 2007.

[42] "Dni, ghi and dhi ground measurements in doha, qatar," *Energy Procedia*, vol. 49, pp. 2398–2404, 2014, proceedings of the SolarPACES 2013 International Conference.

[43] B. Dutta, "A classification and regression tree (cart) algorithm," analyticssteps, July. 27 2021. [Online]. Available: https://www.analyticssteps.com/blogs/classification-and-regression-tree-cart-algorithm

[44] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 785–794. [Online]. Available: https://doi.org/10.1145/2939672.2939785

[45] xgboost developers, "Xgboost tutorials," 2021. [Online]. Available: https://xgboost.readthedocs.io/en/stable/tutorials/model.html

[46] J. Brownlee, "A gentle introduction to sarima for time series forecasting in python," Aug, 21, 2019. [Online]. Available: https://machinelearningmastery.com/sarima-for-time-series-forecasting-in-python/

[47] R. J. Hyndman and G. Athanasopoulos, "Forecasting: Principles and practice, 2nd edition," Melbourne, Australia. OTexts.com/fpp2, 2018. [Online]. Available: https://otexts.com/fpp2/seasonal-arima.html

[48] G. Van Houdt, C. Mosquera, and G. Nápoles, "A review on the long short-term memory model," *Artificial Intelligence Review*, vol. 53, no. 8, pp. 1573–7462, 2020.

[49] J. Brownlee, "Long short-term memory networks with python," English PDF format EBook, 2017 [Online].
[50] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "Lstm: A search space odyssey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2222–2232, 2017.
[51] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm," *Neural Computation*, vol. 12, no. 10, pp. 2451–2471, 2000.
[52] M. Abdel-Nasser and K. Mahmoud, "Accurate photovoltaic power forecasting models using deep lstm-rnn," *Neural Computing and Applications*, vol. 31, no. 7, pp. 1433–3058, 2019.
[53] 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 Conversion and Management*, vol. 212, p. 112766, 2020.
[54] K. Wang, X. Qi, and H. Liu, "Photovoltaic power forecasting based lstm-convolutional network," *Energy*, vol. 189, p. 116225, 2019.
[55] R. Ahmad, Pratyush, and R. Kumar, "Very short-term photovoltaic (pv) power forecasting using deep learning (lstms)," in *2021 International Conference on Intelligent Technologies (CONIT)*, 2021, pp. 1–6.
[56] 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)*, 2021, pp. 1–6.
[57] J. Benesty, J. Chen, and Y. Huang, "On the importance of the pearson correlation coefficient in noise reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 757–765, 2008.
[58] S. I. Vrieze, "Model selection and psychological theory: A discussion of the differences between the akaike information criterion (aic) and the bayesian information criterion (bic)." *Vrieze, Scott I.*, vol. 17, no. 2, pp. 228–243, 2012.