processes

MDPI

Article

Intelligent and Data-Driven Fault Detection of Photovoltaic Plants

Siya Yao [1,2], Qi Kang [1,2], Mengchu Zhou [3,4,*], Abdullah Abusorrah [4] and Yusuf Al-Turki [4,5]

1 Department of Control Science and Engineering, Tongji University, Shanghai 201804, China; yaosiya@tongji.edu.cn (S.Y.); qkang@tongji.edu.cn (Q.K.)
2 Shanghai Institute of Intelligent Science and Technology, Tongji University, Shanghai 201804, China
3 Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ 07102, USA
4 Center of Research Excellence in Renewable Energy and Power Systems, King Abdulaziz University, Jeddah 21589, Saudi Arabia; aabusorrah@kau.edu.sa (A.A.); yaturki@yahoo.com (Y.A.-T.)
5 Department of Electrical and Computer Engineering, Faculty of Engineering, and K. A. CARE Energy Research and Innovation Center, King Abdulaziz University, Jeddah 21589, Saudi Arabia
* Correspondence: zhou@njit.edu

Abstract: Most photovoltaic (PV) plants conduct operation and maintenance (O&M) by periodical inspection and cleaning. Such O&M is costly and inefficient. It fails to detect system faults in time, thus causing heavy loss. To ensure their operations are at an ideal state, this work proposes an unsupervised method for intelligent performance evaluation and data-driven fault detection, which enables engineers to check PV panels in time and implement timely maintenance. It classifies monitoring data into three subsets: ideal period A, transition period S, and downturn period B. Based on A and B datasets, we build two non-continuous regression prediction models, which are based on a tree ensemble algorithm and then modified to fit the non-continuous characteristic of PV data. We compare real-time measured power with both upper and lower reference baselines derived from two predictive models. By calculating their threshold ranges, the proposed method achieves the instantaneous performance monitoring of PV power generation and provides failure identification and O&M suggestions to engineers. It has been assessed on a 6.95 MW PV plant. Its evaluation results indicate that it is able to accurately determine different functioning states and detect both direct and indirect faults in a PV system, thereby achieving intelligent data-driven maintenance.

Keywords: fault detection; performance evaluation; PV monitoring system; tree-based regression; unsupervised learning method

Citation: Yao, S.; Kang, Q.; Zhou, M.; Abusorrah, A.; Al-Turki, Y. Intelligent and Data-Driven Fault Detection of Photovoltaic Plants. *Processes* **2021**, *9*, 1711. https://doi.org/10.3390/pr9101711

Academic Editor: Sara Pescetelli

Received: 5 August 2021
Accepted: 10 September 2021
Published: 24 September 2021

1. Introduction

Nowadays, PV energy represents the third-largest source of renewable energy after wind and hydro [1,2]. Many countries are developing PV projects to utilize such renewable sources. For example, a massive solar park with 7.2 million PV panels has been built in Egypt to increase its generation capacity [3], and an Iowa farm [4] in the USA uses solar power to generate fuel and fertilizer on-site. In order to increase the efficiency of generating power, PV power plants have shifted their focus from large-scale development to large-scale operation and maintenance (O&M) [5,6]. Under this circumstance, intelligent O&M methods are being widely researched. By employing them, PV power stations are capable of analyzing their operation process automatically, coping with faulty situations timely, thus greatly improving the overall efficiency of maintenance and management.

Most intelligent O&M methods are based on a video/image analysis or monitoring database. Video/image-based methods [6–8] utilize UAVs, satellite or 24 h cameras to get videos or images of PV stations and then train deep learning models to detect potential anomalies. However, to obtain a reliable and accurate model requires a large number of labelled samples (anomalies in the PV panels, e.g., short circuit and cell cracking). Yet,

in most cases PV plants are operated normally and thus it is impractical and difficult to get sufficient labelled samples. Therefore, monitoring-based methods [9,10] are playing an increasingly important part in intelligent O&M. Most existing PV plants have been equipped with sensors and monitoring systems. We are able to record detailed historical data (PV power generation data and meteorological data, e.g., solar irradiance, environment temperature, relative humidity, wind velocity, and wind direction) and technical parameters of every piece of equipment. Unlike video/image-based methods which rely on additional UAVs or high-resolution cameras, the monitoring-based methods are directly applicable to conventional PV systems and often cost far less in equipment costs compared to video/image-based ones.

However, there are two shortcomings in existing monitoring systems. First, the interpretation of stored data requires technical background. Query of the database is usually allowed, and data can be organized in plots and tables; however, interpretation is left to the users of such monitoring systems. The curves from the monitoring system are simply values of current and voltage and do not give much information by themselves. Only experienced engineers who are proficient in the process of photoelectric conversion can tell from it whether photocells are healthy or not. Second, a large number of curves are produced every day. Many PV plants record operational data every 10 min or even every minute [11]. It is challenging and time consuming for engineers to distinguish useful and critical information from such huge data. Consequently, intelligent detection methods and instantaneous warnings are in desperate need for present PV monitoring systems.

This work proposes an intelligent O&M method to analyze historical data automatically in a monitoring system and detect an anomalous status, including both direct and indirect faults. Our main contribution is to build two highly effective unsupervised predictive models and refine their predictions with the weather scale factors and power data clustering results. Consequently, we can reveal the upper and lower reference values used for faults detection in PV stations. Note that the predictive models do not provide fault detection results by themselves but are used to offer references and support such detection. Our proposed model not only fits well with the non-continuous characteristic of PV generation, but it also is more sensitive to the indirect faults among PV panels.

This paper consists of the following sections: Section 2 reviews the related works of intelligent O&M approaches and presents the advantages of the proposed method. Then it is detailly illustrated in Section 3. Section 4 describes the experiment results. Finally, the paper is concluded in Section 5.

2. Related Work

Methods of intelligent O&M and fault diagnosis using monitoring data can be categorized into PR (performance ratio) methods, I-V (current-voltage) curve methods, statistical methods, and prediction (machine learning-based) methods.

In PR methods [12,13], the yield of a PV system is evaluated by the ratio between the measured power and the nominal power (which requires precise formulas to calculate, and the formulas are based on the theory of photoelectric conversion) of a system. A low value is an indicator of potential anomalies. In I-V approaches [14–17], the I-V curve of a normally operating PV panel is considered as standard characteristic. The mismatch between the standard and real-time ones is used as the judgement of failures. Similarly, the dI/dV-V curves [18] can also be used to detect failures in PV panels, where dI/dV values are the gradients of I-V curves. Statistical methods do not require any model training [19], but identify abnormal operation based on the statistical characteristics of individual PV feature (e.g., current and voltage) [20]. Statistical methods use the 3σ rule, Hampel identifier, box-plot, and so on; or are based on statistical tests [21], such as ANOVA test and Kruskal-Wallis test.

Prediction methods mainly train a machine learning-based model to directly predict whether PV panels are normal or not. Wang et al. [22], Hussain et al. [23] and Aziz et al. [24] take neural networks for fault diagnosis and O&M with data cloud ac-

quisition. Huang et al. [25] utilizes the AdaBoost algorithm to establish a fault diagnostic model. Momeni et al. [26] uses a graph-based semi-supervised learning (GBSSL) algorithm to identify, classify, locate, and correct faults. Ma et al. [27] focuses on a partial shading scenario, and apply a multiple-output support vector regression (M-SVR) to estimate the shading strength. Chen at al. [28] proposes a random forest (RF) based fault diagnosis model and takes the real-time operating voltage and string currents of the PV arrays as fault features. Compared to the above-mentioned PR, I-V and statistical methods, the prediction methods are data-driven, which learn the diagnosis knowledge based on historical data and is free of the expertise domain background. Moreover, such machine learning-based models can detect faults in real-time and classify their specific type with high prediction accuracy. However, these methods require data to be collected from both normal and faulty conditions. Our proposed method is an unsupervised prediction method, and we do not directly predict which fault occurs. Instead, we predict the expected ideal and worst power generation and then make two comparisons between them, and real-time ones. Hence, we are able to evaluate its real-time performance and identify faults. Our work is novel and advances the area of intelligent O&M in the following aspects:

(1) Applying unsupervised detection method. PV panels' performance depends on meteorological conditions, and a large number of faults may appear. It is difficult to get a dataset covering all possible fault scenarios. Thus, some methods must artificially produce labeled anomalous data by intentionally making some open circuit or short circuit to PV panels [29,30]. This undermines the total power generation of PV stations and declines their operation efficiency. On the contrary, our method is unsupervised without relying on the labelled faulty samples, and simply makes use of the existing monitoring data to evaluate operating status and detect anomalies.

(2) Building non-continuous regression models. Considering the special characteristic of non-continuity in PV generation (Non-continuity is caused by the current-limiting nature in photoelectric conversion [1,29,31], where data values are not continuous in the whole scope and some ranges are meaningless and thus no data values are found there in practice.), we build non-continuous regression models. First, we deploy the ensemble tree-based regression algorithm that adjusts a tree structure according to the data characteristics and thus handles non-linear functions better [32]. Moreover, we implement clustering-based modification to the regression predictions so as to ensure that there is no inexistent value in such non-continuous regression tasks. To the best of our knowledge, our method is the first to deal with such non-continuous issues in PV predictive models.

(3) Detecting indirect faults sensitively. Unlike direct faults that lead to conspicuous performance loss and can be identified easily, indirect faults (caused by dust, module degradation and so on) result in such a gradual PV generation loss that many methods fail to detect it [33,34]. Instead, by comparing the real-time measured value with both the upper and lower references, our method can accurately distinguish different states of PV panels, and hence detect indirect faults sensitively and also provide instantaneous alarm of degradation.

3. Proposed Framework

The main objective of the proposed O&M framework is to enable PV system production to reach its expected level of efficiency intelligently [19]. Therefore, the proposed approach aims at PV system failure detection, performance evaluation, and O&M planning. The notations frequently used in this paper and their descriptions are summarized in Table 1. The concrete steps of the proposed method are detailed in the following sections.

Table 1. Notations and descriptions.

Notation	Description
A	Ideal period
S	Transition period
B	Downturn period
p	PV power generation (KWh)
r	Solar irradiance (W/m^2)
$\widetilde{r}$	Differential value between two adjacent r (W/m^2)
r_L	Logarithmic value of r (W/m^2)
τ	Temperature of a PV panel (°C)
h	Relative humidity (%)
h_L	Logarithmic value of h (%)
v	Wind velocity (m/s)
d	Wind direction (0°–360°)
D	Original dataset after data preprocessing
D_A	Subset of selecting A data from D
D_B	Subset of selecting B data from D
$\mu_i,\ i = 1,\ 2,\ \ldots,\ k$	Cluster centroids, k centroids in total
$C = \{C_i,\ i = 1,\ 2,\ \ldots,\ k\}$	Clustered dataset, cluster C_i represents the i-th class, k classes in total
$Max_i,\ i = 1,\ 2, \ldots, k$	The maximum value of cluster C_i
$Min_i,\ i = 1,\ 2, \ldots, k$	The minimum value of cluster C_i
$l_i\ \epsilon\ \{1,\ 2,\ 3, \ldots,\ k\},\ i = 1,\ 2,\ \ldots,\ m$	Class label of the i-th sample in D
$f_A,\ f_B$	Upper and lower regression models
$x = [r, \tau, h, v, d, T, \widetilde{r}, r_L, h_L]$	Real-time feature vector
$p_A,\ p_B$	Predictions from the upper and lower model
$\hat{p}_A,\ \hat{p}_B$	Modified upper and lower predictions, also simplified as a and b
$\alpha_1,\ \alpha_2,\ \beta_1,\ \beta_2$	Coefficients that divide up a baseline range
w	Weather scale factor

A general framework of the proposed method is presented in Figure 1. First, we apply data preprocessing (including outlier detection, feature analysis and data pre-classification) to both historical meteorological and PV power datasets. Then, historical data are pre-classified into three subsets that represent different operational statuses, namely, ideal period dataset D_A, transition period dataset D_S and downturn period dataset D_B. We apply an XGBoost-based regression algorithm to datasets D_A and D_B, so as to train upper and lower baseline models of a PV plant's power output. Moreover, since PV power data are noncontinuous, we deploy k-means [35] to cluster hierarchical PV power data and use the statistical results of every cluster to modify the prediction values. Furthermore, due to very low PV generation in bad weather (e.g., rainstorm, blizzard, hail, and sandstorm), we consider its corresponding weather scale factors to revise both references. Thus, by integrating the results of upper and lower reference models, clustering results and weather scale factors, we acquire the final upper and lower reference curves. Comparing the measured power with two reference curves, we can evaluate their performance, detect faults, and carry out intelligent data-driven O&M. It is noted that our method does not use the information related to a PV system's components like inverters, which means that it is a generic data-based method and not limited to a certain type of PV systems.

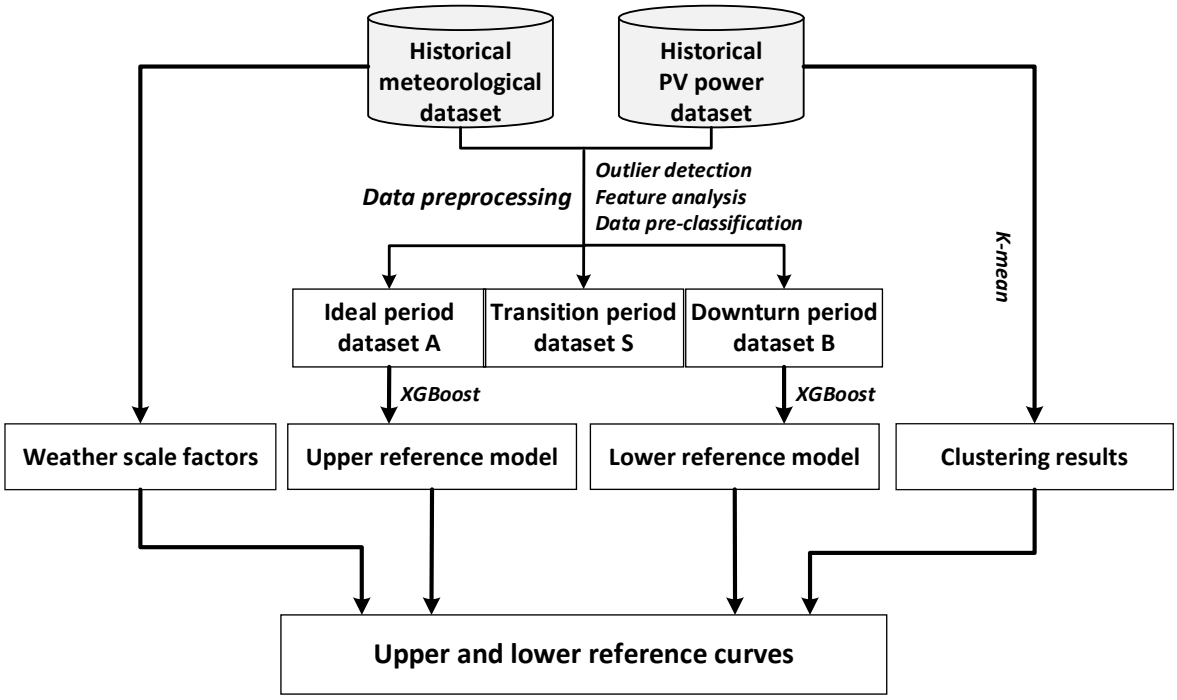


Figure 1. The general framework of the proposed method.

We intend to evaluate the different operating statuses of real-time power generation so as to better implement O&M. Thus, we propose the five stages (determined by two references) and present corresponding definitions, which are detailed in Section 3.3. If only using a simple threshold to identify how close the actual power value is to the expected value, there is only one reference indicating the expected generation, so it would be easy and clear to identify whether the PV panels are in the expected state or not. However, it is not qualified to answer the following important questions: which operating situation (ideal or malfunction period) the PV panels are in when the actual generation values are higher than the expected ones, and how to distinguish the downturn period when covered by light-barriers and the malfunction periods when suffering from short-circuit (their power outputs in such cases are far below the expected reference values)? Using only one threshold makes it difficult to make more fine-grained performance evaluations. Therefore, we propose to use two references indicating the expected best and worst generation. With two references, the above questions can be easily answered. Moreover, it is more accurate to determine the operating status and evaluate real-time generation efficiency.

3.1. Data Preprocessing

Before a prediction model is applied, the first step is to conduct data preprocessing, including outlier detection, feature analysis, and data pre-classification.

3.1.1. Outlier Detection

Due to the error or failure of sensor data transmission, there are various anomalies in raw PV monitoring data, such as missing, negative, and duplicated values. Apart from conducting basic preprocessing towards such obvious outliers, we pay attention to detecting others, e.g., extreme and unmatched values in the original dataset so as to thoroughly clean the data.

First, we apply classical statistical methods, i.e., box-plot and 3σ criterion, on each single feature and try to detect outliers that deviate far away from most data. Note that, such statistical methods achieve local detection that only identify extreme values in a single feature. As for global detection, we concentrate on unmatched values. For example, in one PV plant, when irradiation is more than 1000 W/m^2, the corresponding PV generation

should also be quite large, e.g., 1000 KWh. However, there is a record with 1000 W/m^2 irradiation but very low generated power, e.g., 20 KWh. In the proposed framework, such unmatched structural outliers are removed by an unsupervised machine learning algorithm, i.e., DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [36], which is able to identify hidden outliers from the global perspective. DBSCAN has two major parameters: radius ε determines the scope of a cluster, and minimum number of points N means the minimum number of members in a cluster. It can be regarded as a simple binary classification (normal data vs. outliers) method. Although it is an excellent anomaly detection method, it cannot be directly used in a monitoring system for classifying different operating status and different faults. Due to its design, it is sensitive to data distribution and depends heavily on the manual off-line setting of parameters, which is not suitable for online detection. Especially, under a situation where data are recorded every minute, DBSCAN would gradually turns to be unstable and inaccurate without timely human supervision. The biggest motivation of intelligent PV fault detection is to identify faults instantly and warn engineers of anomalous situations in time, so that they don't have to keep their eyes on these monitoring data but still can notice anomalies at the first time. Using only DBSCAN is hard to perform such goal, so we propose the methods depicted next. After detecting and removing outliers in raw data, we obtain our dataset D.

3.1.2. Feature Analysis

Since we try to predict PV generation, it is necessary to carry out detailed analysis of PV power data (p). First, it is greatly affected by the fluctuation and uncertainty of meteorological factors, and hence exhibits variability and volatility. Particularly, under the nonstationary and low illumination intensity in cloudy and rainy days, PV power data are prone to fluctuating violently [2,26,31,37]. Second, due to the current-limiting nature, a PV system has non-continuous output characteristics of power generation [2,29,37]. However, traditional methods, e.g., linear regression and support vector regression, fit continuous data and output continuous value. The special characteristics of PV power data increase difficulties and challenges for the accurate prediction of PV output. We deal with such non-continuous regression tasks by clustering-based modification which is presented in detail in the next sections.

Besides analyzing PV power data, we pay attention to the factors that affect or contribute to PV output. The most directly related factors are meteorological one. Commonly used factors include solar irradiance (r), temperature of PV panel (τ), relative humidity (h), wind velocity (v), and wind direction (d). In order to capture the non-linear relationship between meteorological factors and PV power, as many feature engineering methods do, we construct two additional features: r_L and h_L, which are the logarithmic values of r and h. To capture the changing trend (increasing or decreasing) of solar irradiance, we add feature $\widetilde{r}$ which denotes the differential irradiance between two adjacent r values.

3.1.3. Data Pre-Classification

In the proposed method, a critical part is to build two models: upper and lower reference models. It is of great importance to select suitable valid data from the original data and use them to train two models. We dig out the original data D and manually select the ideal period dataset D_A and the downturn period dataset D_B for the upper and lower model, respectively.

As shown in Figure 2, we propose to pre-classify the original data into three periods. For most PV plants, as time goes by, if there are no faults during running, the PV panels degrade due to dust or module deterioration. Thus, the power generation presents a declining tendency as shown in Figure 2. The states of PV panels are divided into the following three periods:

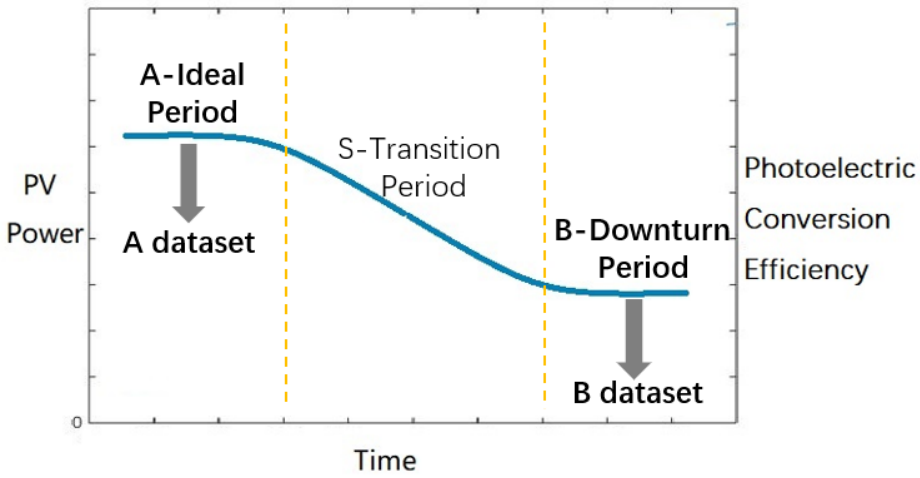


Figure 2. Three typical states of PV power.

(1) Ideal period A: The first time when the panels are brought into operation or after maintenance (e.g., cleaning and washing), the PV panel is in a healthy and clean state without any light barriers. At this time, the efficiency of photoelectric conversion is comparatively high. The power generation in a PV plant is also at an ideal state, namely, relatively high and stable.

(2) Transition period S: Under a natural state and without any interference, PV panels gradually accumulate dust and some light barriers (e.g., bird dropping, leaves, snow and plastic bags). Under this circumstance, the conversion efficiency slows down, and power generation gradually declines too. The total PV power generation makes a gradual transition from ideal state to a lower state.

(3) Downturn period B: When there is visible dust or too much light barriers on PV panels, PV cells receive little solar irradiation, or when they are aging, the photoelectric conversion efficiency reaches its lowest limit, and the generated power continues to be sluggish.

Among these operating periods, we pay special attention to A and B periods. We collect data from these two periods to construct the ideal period dataset D_A and downturn period dataset D_B. Note that in this paper we manually classify the dataset D and select suitable data for D_A and D_B. According to the definitions of operating periods and our preliminary investigation, we are able to select data for D_A and D_B based on weather and maintenance records. Based on experience, these two factors are the most related ones in regard to generation efficiency. Besides, it is convenient to get access to its own historical operating records and the weather-related data online. Our method can be easily applied in other similar tasks. In the future, we consider labeling historical data D with different period labels and then train a classification model, thus avoiding manual selection of data.

3.2. Non-Continuous Regression Models

The proposed method builds the upper and lower baseline models with D_A and D_B, respectively. The training procedure of these two models are similar, and the only difference lies in a different PV dataset we input for training.

As mentioned above, PV power data have special characteristics of variability and non-continuity, which motivates us to deploy an ensemble trees-based regression method called extreme gradient boosting (XGBoost) [38]. It assembles a number of CART (Classification and Regression Tree) as base learners, which can deliver more accurate prediction. It inherits the advantages of a decision tree algorithm and handles well non-continuous functions, which exactly suits the prediction task on non-continuous PV power data. Hence, we deploy it as our regression algorithm.

The power generation p is the output of our prediction model whose inputs are the combination of meteorological features (r, τ, h, v, d), time-related features (T), and additional features $(\widetilde{r}, r_L, h_L)$. Then, our regression prediction model is denoted as:

$$p = f(r, \tau, h, v, d, T, \widetilde{r}, r_L, h_L) \tag{1}$$

where f is an XGBoost-based prediction function. Note that we cannot obtain explicit expressions in a tree-based regression method. Hence, f is a simplified notation of a tree structure and corresponding parameters. Using D_A and D_B as training datasets, we can obtain two prediction models f_A and f_B.

By inputting a real-time feature vector:

$$x = [r, \tau, h, v, d, T, \widetilde{r}, r_L, h_L] \tag{2}$$

into f_A and f_B, we can conduct PV generation prediction and acquire the upper and lower references, i.e.,

$$p_A = f_A(x) \tag{3}$$

$$p_B = f_B(x) \tag{4}$$

Although the proposed XGBoost-based regression model is suitable for fitting non-continuous PV data, it is still a regression method and sometimes obtain outputs that do not exist in a real PV system. Considering the non-continuous characteristic of PV generation, it is necessary to implement further modification to refine (3) and (4), i.e., modifying with the weather scale factors and power data clustering results, which is detailed as follows. Due to the above mentioned current-limiting principle, PV power data are of obviously hierarchical discreteness. Power values belong to several particular groups where they are continuous. Between two adjacent groups, there is a blank gap with no data. We propose to cluster the original power data by using k-means algorithm [35]. In k-means, there is only one key parameter: the number of clusters denoted as k. After clustering, we calculate the minimum and maximum values for each cluster. Hence, we can know the ranges to which actual values belong. For upper or lower predictions located outside existing ranges, we propose to modify them with the maximum or minimum values of their closest cluster.

For the upper prediction value p_A, if it does not belong to any cluster, then the principle of proximity is adopted to correct it. We replace it with the maximum value of the closest cluster. The modified prediction value is

$$\hat{p}_A = \underset{Max_j}{argmin}\left(\left|p_A - Max_j\right|\right),\ j = 1,\ 2, \ldots, k \tag{5}$$

where Max_j is the maximum value of the j-th cluster.

Similarly, for p_B located beyond any existing cluster, we modify it with the closest minimum value, as follows:

$$\hat{p}_B = \underset{Min_j}{argmin}\left(\left|p_B - Min_j\right|\right),\ j = 1,\ 2, \ldots, k \tag{6}$$

where Min_j is the minimum value of the j-th cluster.

Considering the variability of PV generation under different weather conditions, we propose the weather scale factors so as to make our prediction more robust. When predicting expected PV generation in bad weather (here, bad weather refers to the case of greatly unstable irradiance or extremely low irradiance, e.g., rainstorms, blizzards, hail, and sandstorms), we multiply them by weather scale factors. In the proposed method, a weather scale factor w is defined as the percentage of reduction of power generation in bad weather. It can be computed as the ratio of average power output from a normal day to that of a bad weather day, which can be derived from the historical monitoring data. Then,

(5) and (6) are modified by $\hat{p}_A = w\hat{p}_A$ and $\hat{p}_B = w\hat{p}_B$. Prediction modification is realized in Algorithm 1.

Algorithm 1: Non-continuous Regression Prediction

Input: Dataset D_A, dataset D_B, original dataset D, number of clusters k, real-time feature vector x and its corresponding w.
Output: Upper and lower prediction values $\hat{p}_A$ and $\hat{p}_B$

1. Train XGBoost algorithm on D_A and obtain the upper prediction model f_A
2. Train XGBoost algorithm on D_B and obtain the lower prediction model f_B
3. From D extract power data series $P = \{p_1, p_2, \ldots, p_m\}$
4. Initialize cluster centroids $\mu_1, \mu_2, \ldots,$ and μ_k randomly
5. **For** each sample p_i in P
6. classify it into the closest category:
7. Let $l_i = argmin_{1 \le j \le k} \| p_i - \mu_j \|$
8. Moving each cluster centroid μ_j to the mean of the points assigned to it:

$$\mu_j = \left(\sum_{i \in c_j} p_i \right) / \left| c_j \right|$$

9. **Repeat** above for-loop until the change of centroids is less than a certain threshold
10. Obtain the clustered data $C = \{C_1, C_2, \ldots, C_k\}$
11. **For** each cluster C_i in C
12. Let $Max_i = \max\left\{ p_j \middle| p_j \in C_i \right\}$, $Min_i = \min\left\{ p_j \middle| p_j \in C_i \right\}$
13. Input x to f_A and obtain upper prediction p_A
14. **If** $p_A \in \cup_{j=1}^{k} \left[Min_j, Max_j \right]$
15. Let $\hat{p}_A = p_A$
16. **Else** do Equation (5)
17. Input x to f_B and obtain lower prediction p_B
18. **If** $p_B \in \cup_{j=1}^{k} \left[Min_j, Max_j \right]$
19. Let $\hat{p}_B = p_B$
20. **Else** do Equation (6)
21. Match weather scale factor w
22. **Let** $\hat{p}_A = w\hat{p}_A$, $\hat{p}_B = w\hat{p}_B$
23. **Return** upper and lower prediction values $\hat{p}_A$ and $\hat{p}_B$

3.3. Performance Evaluation, Fault Detection, and O&M Planning

Generally, a PV system can be affected by different types of faults that result in the significant loss of power [20,39,40]. According to the factors causing PV faults, two types of faults can be distinguished: direct and indirect faults. Some direct faults such as cell cracking, nonconnected module, open circuit and short circuit in a PV system, and broken fuse or cable, cause conspicuous performance loss. Indirect factors, such as shading due to dust or light barriers, encapsulation degradation due to ultraviolet and yellowing EVA (ethylene vinyl acetate), module degradation due to light or heat, and rust due to water infiltration, lead to the gradual deterioration of PV panels, and hence result in the gradual power loss [34]. Using the monitored data, a PV monitoring system has to decide whether there is degradation in its generation performance [41].

In the proposed approach, apart from the real-time PV power-versus-time curve displayed in the monitoring system, there are also two reference curves (A and B) from regression models exhibited in the same figure. For each real-time record (including a feature vector x and its corresponding power generation p), we obtain its expected ideal and worst PV generations $\hat{p}_A$ and $\hat{p}_B$ by inputting its feature vector x into Algorithm 1. To simplify, we set $a = \hat{p}_A$ and $b = \hat{p}_B$. Our method evaluates the PV panels' real-time status by comparing real-time PV power p with a and b. As in Table 2, generally, according to different conditions, there are five different stages. They are classified into four typical operating periods: malfunction period in Stages 1 and 5; ideal period in Stage 2; transition period in Stage 3; and downturn period in Stage 4.

Table 2. Five conditions of power data and corresponding status.

Stage	Condition	Period
Stage 1	Far above $\hat{B}$: $p > (1+\alpha_1)a$	Malfunction period
Stage 2	Fluctuating near $\hat{B}$: $(1-\alpha_2)a < p < (1+\alpha_2)a$	Ideal period
Stage 3	Between $\hat{B}$ and $\check{B}$: $(1+\beta_2)b < p < (1-\alpha_2)a$	Transition period
Stage 4	Fluctuating near $\check{B}$: $(1-\beta_2)b < p < (1+\beta_2)b$	Downturn period
Stage 5	Far below $\check{B}$ baseline : $p < (1-\beta_1)b$	Malfunction period

If real-time power exhibits more than a given percentage of generation losses, they are classified into a downturn or malfunction period. In our method, we set α_1, α_2, β_1, and β_2 ($\alpha_1 > \alpha_2$, $\beta_1 > \beta_2$ and α_1, α_2, β_1, $\beta_2 \in [0,1]$) to divide up the warning ranges. Users receive an alarm if a PV panel produces more than α_1 of the expected ideal baseline $\hat{B}$, i.e., $p > (1+\alpha_1)a$, or less than β_1 of the expected worst baseline $\check{B}$, i.e., $p < (1-\beta_1)b$, which means that the sensors or PV panels may break down or the data transmission is incorrect. If p is fluctuating near $\check{B}$, i.e., $(1-\beta_2)b < p < (1+\beta_2)b$, the PV panels are of low efficiency, and they need timely maintenance, such as manual cleaning and equipment repair. For α_1 and β_1, smaller values mean stricter alert and more sensitive detection; larger values mean looser limitation, but help reduce false alarms. On the contrary, a larger β_2 means larger range of Stage 4, which may result in more observations classified into the downturn period, hence bring in more false alarms.

If p is near $\hat{B}$, i.e., $(1-\alpha_2)a < p < (1+\alpha_2)a$, it indicates that the PV power generation is running in an ideal state. There is no need to implement any maintenance. Furthermore, there is no warning or alarm when p is between $\hat{B}$ and $\check{B}$, i.e., $(1+\beta_2)b < p < (1-\alpha_2)a$, it is in the transition period, and we consider it as a normal life cycle of PV panels. Note that, if α_2 is too large, there are more data classified into the ideal period, leading to the risk of misclassification of potential faults.

4. Experimental Results

4.1. Data Description

We have conducted experiments based on the proposed method and used it in a 6.95 MW PV plant. Apart from zero-records (at night or missing), available effective monitoring data consist of 5936 records (In our experiments, the monitoring system records sensor data every 15 min). Although the time range of the collected dataset covers less than one year, our data included all kinds of weather situations, especially some extreme weather, e.g., snow, high temperature and rainstorms. Examples of the collected data are shown in in Supplementary File. As mentioned in Section 3, we collect power data p and features $x = [r, \tau, h, v, d, T, \widetilde{r}, r_L, h_L]$ in (2).

4.2. Experimental Results

4.2.1. Results of Data Preprocessing

We use DBSCAN [36] to detect outliers and the parameters are set as ε = 46 and n = 25. To explore raw data intuitively, we plot PV data of a week in Figure 3. Only daytime hours (from 6:00 to 18:00) are considered in our PV forecasting application. As in Figure 3, PV power experiences violent fluctuation within a day as well as drastic variation among days. Meanwhile, we plot all measured PV power data versus time, as shown in Figure 4. It is noticeable that PV power data is non-continuous. Looking at the picture from the bottom to top, there are many blank gaps between two adjacent data dot clusters. Obviously, power data belong to different groups. In sum, the characteristics of raw data: intensive fluctuation and variability, and hierarchical non-continuity, motivate us to apply an XGBoost algorithm that suits most for non-continuous PV power regression task.

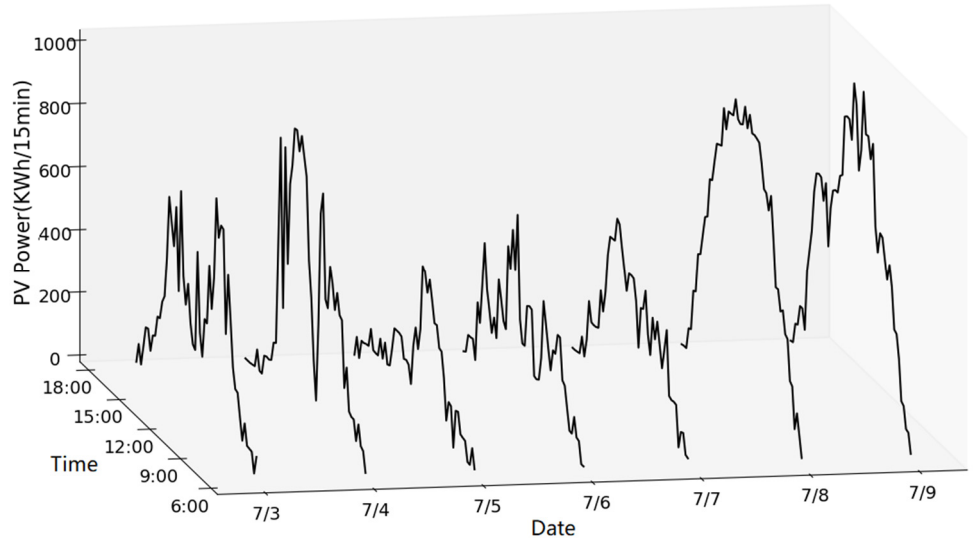


Figure 3. Day-curves of PV power from partial data.

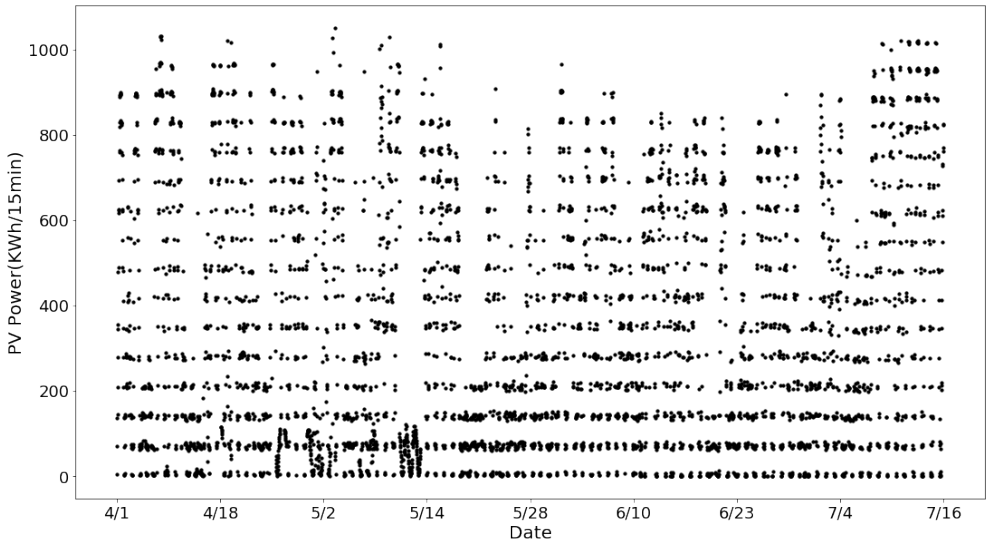


Figure 4. All measured PV power data.

We aim to select appropriate data for two datasets, D_A and D_B, reflecting the ideal and downturn periods, respectively. First, we get D_A by considering maintenance records of a studied PV plant. The scheduled maintenance plan for year 2018 is to do the cleaning work every month, and each last for nearly half a month (from 16th to the end of that month). Obviously, the cleaning work is quite frequent. Under this circumstance, PV panels always stay in a comparatively clean status, namely, the ideal period. (Since we need the downturn period data, we stop the monthly cleaning from May.) Moreover, we consider that PV panels are perfectly clean two days after cleaning, so we take data from 18th to the end of the month from January to April as an ideal period dataset. For example, in January, data are collected into D_A. In order to obtain valid D_B, we suspend the monthly cleaning from May. Without cleaning, PV panels depend on the rain to wash away dust or other light barriers. So, it is important to find data originated from continuous sunny days, where PV panels may be covered with dust or light barriers, and hence in the downturn period. By checking the historical weather, which is freely available online, we are able to select suitable valid data for D_B from May to July. Take July as an example, it is rainy in the first week. Thus the data are not appropriate for dataset D_B. But since 8 July 2018, it has been

cloudy, overcast or sunny, which suits our selection principle of dataset D_B. So, we take data from 9 July 2018 to 15 July 2018 into dataset D_B.

To conclude, by data preprocessing, we determine the XGBoost-based regression algorithm according to the special characteristics of PV power data. Moreover, we present how to construct datasets D_A and D_B. More detailed analyses about data preprocessing can be viewed in Supplementary File.

4.2.2. Results of Non-Continuous Regression Models

We compare XGBoost with seven universal regression methods, and each of them has achieved good performance in existing research. This includes multivariable linear regression (MLR) attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. ElaticNet [42] is a regularized regression method that linearly combines the L1 and L2 penalties of the lasso and ridge methods. Support vector regression [27] is a version of support vector machine (SVM) for regression. Here, we apply kernel-based SVM. Support vector regression with linear kernel is denoted as SVR, and the one with radial basis function kernel is denoted as SVR-RBF. Decision tree regression (DTR) [43] uses tree structure to predict the continuous output on the basis of input or situation described by a set of properties. Random forest regression (RFR) [28] is an ensemble learning method, which constructs a multitude of decision trees at training time and outputs the mean prediction of the individual trees. Gradient boosting decision tree (GBDT) [44] builds the ensemble model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function. Note that we compare their basic regression performance and do not implement the proposed modifications (clustering-based one and weather scale factors-based one). The compared algorithms are listed:

(1) Multivariable linear regression (MLR)
(2) ElasticNet
(3) Support vector regression with linear kernel (SVR)
(4) Support vector regression with radial basis function kernel (SVR-RBF)
(5) Decision tree regression (DTR)
(6) Random forest regression (RFR)
(7) Grmadient boosting decision tree (GBDT)

The above algorithms are available in scikit-learn [45] which is a free software machine learning library for the Python programming language. Among them, RFR and GBDT are similar to XGBoost. They are tree-based ensemble methods but others use a single model.

To validate the generalization performance, we extract four datasets from the dataset D: April, May, June, and July. They are under different meteorological conditions. Then, for each dataset, we split 50% for training regression models and the rest for testing. We apply five-fold cross validation to search the optimal parameters that show the highest accuracy.

We use three performance metrics on test data. They are the ratio of root mean squared error to the mean value denoted as $\widetilde{E}$, the mean absolute error denoted as $\overline{E}$, and the goodness of fit denoted as R^2.

$$\widetilde{E} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{\overline{y}_i} \times 100\% \tag{7}$$

$$\overline{E} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{8}$$

$$R^2 = \left(1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2}\right) \times 100\% \tag{9}$$

where y_i is the ground truth of p, $\hat{y}_i$ is the prediction value, $\overline{y}_i$ is the mean of all y_i, and n is the number of test samples.

All experiments are carried out in Python 3.7 with an Intel Core i5-8250 CPU and 8G memory. Table 3 shows the performance results on each dataset, and the best one is highlighted in bold. From Table 3, compared to other methods, XGBoost generally achieves much better performance. Even if it does not rank the best on dataset May, it achieves the second-best with a small gap to the best one, i.e., RFR. Table 4 presents the average evaluation metrics on the four datasets. XGBoost outperforms its peers with the best average performance metrics. In particular, XGBoost achieves better performance than RFR and GBDT, both of which are the state-of-the-art ensemble regression methods. In sum, compared with its seven peers, XGBoost achieves higher average accuracy and more generalized performance under different meteorological conditions.

Table 3. Accuracy of algorithms.

Algorithms	April			May			June			July		
	$\widetilde{E}$ (%)	$\overline{E}$	R^2 (%)	$\widetilde{E}$ (%)	$\overline{E}$	R^2 (%)	$\widetilde{E}$ (%)	$\overline{E}$	R^2 (%)	$\widetilde{E}$ (%)	$\overline{E}$	R^2 (%)
MLR	10.87	72.20	88.25	9.71	76.61	86.26	5.29	38.76	95.89	5.79	41.31	96.06
ElasticNet	10.91	72.60	88.17	9.76	76.74	86.10	5.32	39.03	95.85	5.92	42.78	95.87
SVR	16.08	71.50	74.31	14.89	87.32	67.66	7.58	53.33	91.58	7.23	51.53	93.83
SVR-RBF	16.11	76.72	74.20	15.34	95.42	65.70	8.24	56.83	90.03	8.88	64.22	90.72
DTR	5.16	37.10	97.35	6.04	46.05	94.67	6.28	46.09	94.22	7.24	51.53	93.82
RTR	4.45	31.58	98.03	**5.25**	**38.78**	**95.98**	5.39	38.79	95.74	6.23	42.42	95.43
GBDT	4.86	34.96	97.65	5.40	39.73	95.74	5.44	38.09	95.66	6.44	43.35	95.12
XGBoost	**4.31**	**30.96**	**98.16**	5.66	41.97	95.32	**5.19**	**37.27**	**98.02**	**5.61**	**38.07**	**96.29**

Table 4. Average accuracy of algorithms.

Algorithms	MLR	ElasticNet	SVR	SVR-RBF	DTR	RTR	GBDT	XGBoost
$\widetilde{E}$(%)	7.91	7.98	11.44	12.14	6.18	5.33	5.53	**5.19**
$\overline{E}$	57.22	57.78	65.92	73.29	45.19	37.89	39.03	**37.06**
R^2 (%)	91.62	91.50	81.85	80.16	95.02	96.30	96.04	**96.95**

Then, we apply XGBoost to train upper and lower reference models. We randomly split 50% of selected D_A and D_B to train regression models. The rest of D_A and D_B are for testing. The optimal XGBoost parameters for upper models are as follows: maximum depth is 3, learning rate is 0.1 and the number of estimators is 125. As for the lower model, maximum depth is 3, learning rate is 0.06, and the number of estimators is 300. For both upper and lower models, other not-mentioned parameters are set as default values. As discussed in Section 3, we deploy a k-means clustering algorithm to classify original PV power data and then take the clustering results to modify the prediction values. In the k-means algorithm, we set parameter $k = 16$ to ensure that our data are exactly classified into 16 classes (As in Figure 4, there are 16 classes). The clustering results are presented in the Supplementary File section. We calculate the maximum value (Max) and the minimum value (Min) of every cluster. Then, it is necessary to take the weather scale factors into consideration to avoid incorrect near-zero PV prediction. The near-zero values usually are regarded as data from a malfunction period, but for extreme weather, it is reasonable to get zero PV output. Some methods may produce false alarm under this situation. The weather scale factors are used to modify the predictions under the case of greatly unstable irradiance or extremely low irradiance, thus effectively avoiding the abovementioned false alarm. The weather scale factors are computed as the ratio of average power output from a normal day to that of a bad-weather day. The studied PV plant is operated under normal weather conditions, i.e., the collected datasets does not include data under extreme weather conditions (e.g., blizzards, hail, and sandstorms), and we thus set $w = 1$. Our future work plans to find more datasets that cover all types of weather

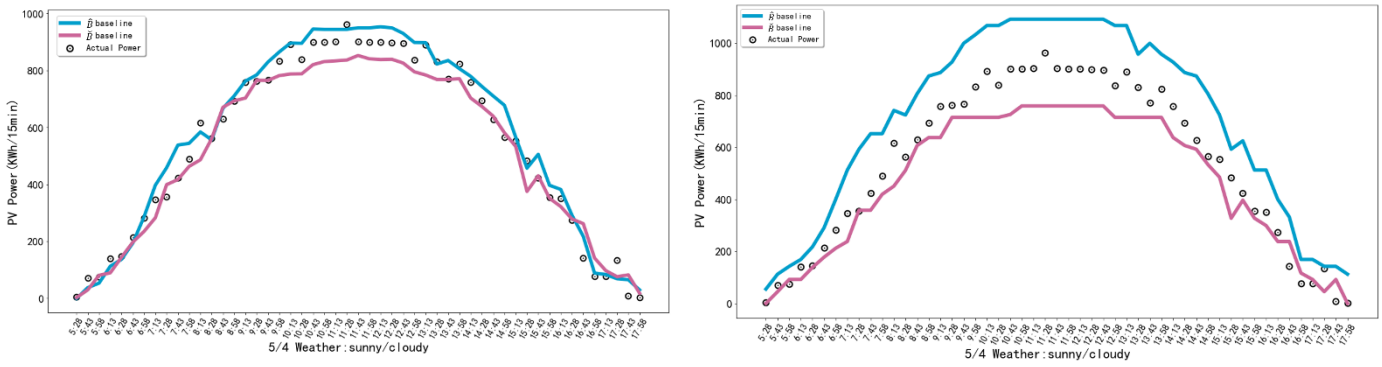
conditions and conduct related analysis to show how weather scale factors can be well-used to improve model performance. Following Algorithm 1, we obtain the final predictions for the test samples.

Table 5 presents model performance on training and test datasets. $\widetilde{E}$ and $\overline{E}$ are low on two test datasets, R^2 is very close to 1 on both models, indicating that our two prediction models are highly accurate and reliable. Note that, the performance in Table 5 is not as good as those in Tables 3 and 4. This is because our selected datasets D_A and D_B consist of records from different months, and hence the training and test datasets are less stable than those in above experiments.

Table 5. Performance metrics of two prediction models.

Performance Metric	Upper Model	Lower Model
$\widetilde{E}$ (%)	8.34	10.07
$\overline{E}$	23.72	16.80
R^2 (%)	98.97	99.09

To better present the superiority of our proposed modification methods, i.e., the clustering-based modification and weather scale factor-based one as shown in Algorithm 1, Figure 5a,b show an example of the monitoring system before and after modifica-tion, respectively. Two trained models are applied to the monitoring data of a day in May which are presented as black hollow circles. The blue curve is the upper reference from f_A, and the purple one is the lower reference deriving from f_B. In Figure 5, it is noticeable that purple one is sometimes above the blue one, e.g., curves during 5:28–6:43, 8:28–9:28, and 16:28–17:58. Moreover, there are many observations below or above the reference lines. Instead, in Figure 5, there is no overlap, and the purple reference line is below the blue one. Furthermore, the modifies references are more consistent with the changing trend of actual PV power values, and there are also less observations located outside two references.



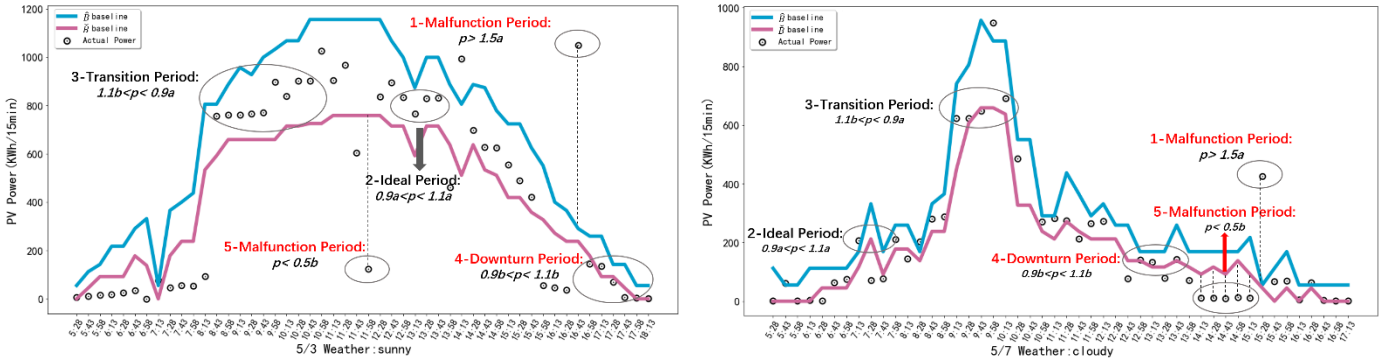
(**a**) (**b**)

Figure 5. (**a**) Baselines and actual power data of a day in May; (**b**) Modified baselines and actual power data.

4.2.3. Results of Performance Evaluation, Fault Detection, and O&M Planning

Based on two modified prediction models, we obtain the corresponding upper and lower references of power generation. We compare them with the measured power value, and assess PV panel status according to their distributions in the reference range. After different tests about selecting suitable parameters, we recommend to set $\alpha_1 = \beta_1 = 0.5$ and $\alpha_2 = \beta_2 = 0.1$, which show satisfactory results in most experiments. To make a comprehensive comparison, we show pictures of different weather conditions, i.e., a sunny day in Figure 6a, and a cloudy dayin Figure 6b. There are five kinds of typical distributions. As discussed before, data in Stages 1, 4, and 5 can provide early warning for the engineers in a PV station. In Stages 1 and 5, actual values deviate far from the references. This is the

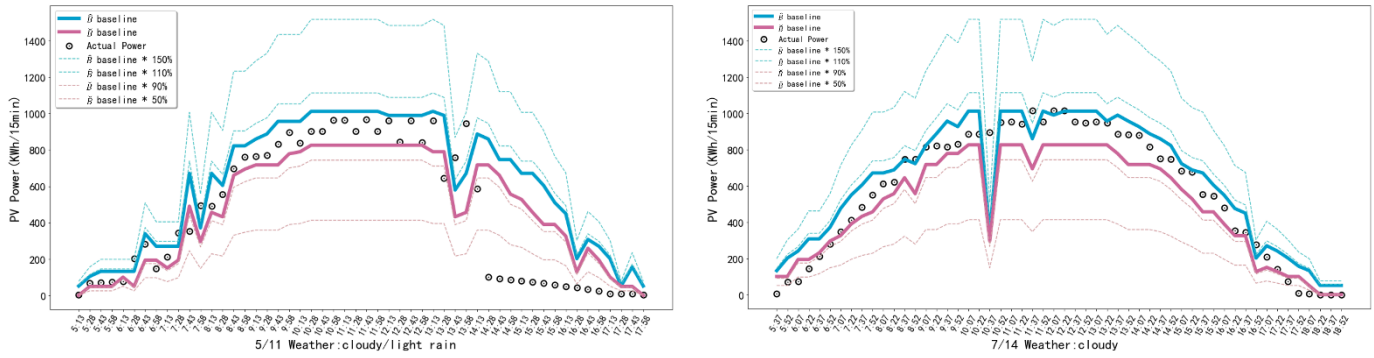
malfunction period, where the sensors may break down and the data transmission may be incorrect. Specially for Stage 5 where the power is relatively low, chances are high that there are open circuits or short circuits in PV panels. In Stage 4, the generated power is comparatively low. PV panels are in a downturn period, and may be covered by dust and need cleaning. It is worth noting that some disturbance in the power grid also leads to the decrease of output power. It may fall below the lower reference curve, but it is not due to a failure in the PV plant. After such warning, on-site engineers need to conduct further inspection and corresponding O&M plans. Stages 2 and 3 do not trigger warning, because they correspond to an ideal period and transition period. According to our previous data analysis and definitions of stages, the transition period (Stage 3) and the downturn period (Stage 4) are the most common ones. Generally, a normal operating PV station does not frequently break down with severe faults or always yield an ideal power generation with high operational efficiency. Hence, a malfunction period (Stages 1 and 5) and ideal period (Stage 2) are the relatively less common ones.



(**a**) (**b**)

Figure 6. Five kinds of distribution on (**a**) a sunny day and (**b**) a cloudy day.

In order to explicitly present the performance evaluation, we plot the warning boundary lines in Figure 7a,b. In the monitoring system, the engineers are capable of directly distinguishing the PV panel statuses and getting suggestions about how to carry out proper O&M plans. As in Figure 7a, the PV power generation experiences an abrupt decline and drops greatly at 14:28 in 2018/5/11, which indicates that the PV station is in a malfunction period and maintenance is required. According to the abrupt decline and long-lasting Stage 5, we consider that there is direct fault in the PV plant, e.g., nonconnected modules and short/open circuits. Such direct faults are relatively easy to notice in a monitoring system. They usually happen in Stage 5, accompanied by an obvious and long-term decline of PV generation. In this case, further O&M plans lie in checking detailed PV records about each panel and then locating the faulty one(s). As in Figure 7b, a cloudy day in summer, the solar irradiance is strong, so the curves of p, a, and b are nearly sinusoids shape. At 10:37, both A and B baselines fall greatly, whereas the actual p stays in the original trend. There is a high chance that meteorological sensor errors or transmission mistakes appear. The wrong data are input to our prediction models, so we get wrong results. We suggest that further O&M plans attach importance to check the original database and repair or replace faulty sensors.



(**a**) (**b**)

Figure 7. Performance Evaluation of (**a**) 2018/5/11 and (**b**) 2018/7/14.

As for the validation, as there is no label about which performance stage the PV system is in and which fault it suffers (which makes it difficult to conduct detailed verification and give specific classification metrics), we have manually labelled the data and conducted classification experiments to verify the performance of our method compared with other advanced machine learning classification algorithms. The input of classification models are the monitoring records that include both meteorological data as listed in (2) and corresponding generated PV power data. The output of classification models indicates which performance period the PV system is in, i.e., malfunction, ideal, transition, or downturn periods. We compare our method with several widely-used and powerful algorithms under their classification implementations, i.e., support vector machine classification [46–48] with linear kernel (SVC-Linear), support vector machine classification with RBF (SVC-RBF), decision tree classification (DTC) [49], random forest classification (RTC) [50], gradient boosting decision trees classification (GBDTC) [51] and extreme gradient boosting trees classification (XGBC) [52]. The above algorithms are available in scikit-learn [45]. The classification performance metrics are shown in Figure 8a–g and Tables 6–12.

Table 6. Performance of DTC.

DTC	Precision	Recall	f1-Score
0	0.88	0.85	0.87
1	0.56	0.70	0.62
2	0.63	0.61	0.62
3	0.74	0.65	0.70
Macro-Average	0.71	0.70	0.70

Table 7. Performance of SVC-Linear.

SVC-Linear	Precision	Recall	f1-Score
0	0.96	0.64	0.77
1	0.75	0.84	0.79
2	0.63	0.70	0.66
3	0.75	0.81	0.78
Macro-Average	0.77	0.75	0.75

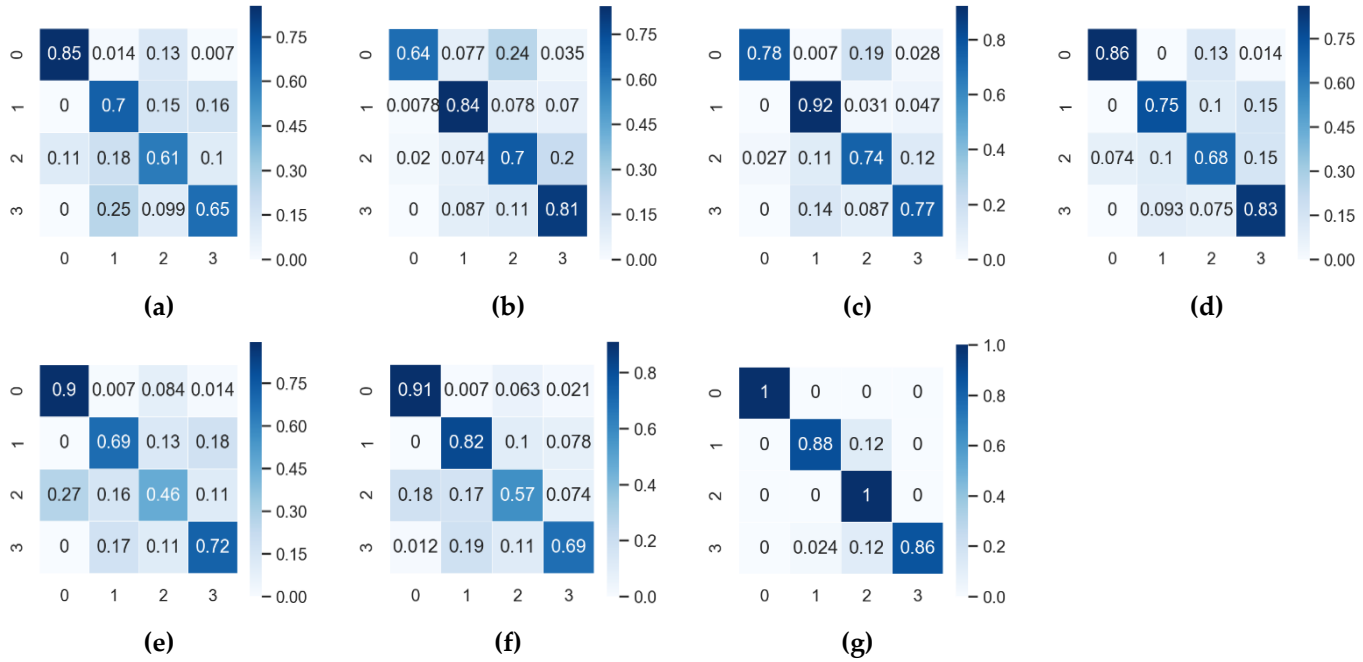


Figure 8. CM after (**a**)DTC; (**b**) SVC-Linear; (**c**) SVC-RBF; (**d**)RTC; (**e**)GBDTC; (**f**)XGBC; (**g**) Our method.

Table 8. Performance of SVC-RBF.

SVC-RBF	Precision	Recall	f1-Score
0	0.97	0.78	0.86
1	0.75	0.92	0.83
2	0.71	0.74	0.73
3	0.82	0.77	0.79
Macro-Average	0.81	0.80	0.80

Table 9. Performance of RTC.

RTC	Precision	Recall	f1-Score
0	0.92	0.86	0.89
1	0.76	0.75	−0.76
2	0.70	0.68	0.69
3	0.76	0.83	0.79
Macro-Average	0.78	0.78	0.78

Table 10. Performance of GBDTC.

GBDTC	Precision	Recall	f1-Score
0	0.76	0.90	0.82
1	0.63	0.69	0.66
2	0.60	0.46	0.52
3	0.73	0.72	0.73
Macro-Average	0.68	0.69	0.68

Table 11. Performance of XGBC.

XGBC	Precision	Recall	f1-Score
0	0.82	0.91	0.86
1	0.65	0.82	0.73
2	0.68	0.57	0.62
3	0.82	0.69	0.75
Macro-Average	0.74	0.75	0.74

Table 12. Performance of Our Method.

Ourmethod	Precision	Recall	f1-Score
0	1.00	1.00	1.00
1	0.97	0.88	0.92
2	0.80	1.00	0.89
3	1.00	0.86	0.92
Macro-Average	0.94	0.93	0.93

Based on whether PV generation is matched with real-time meteorological records, we manually classify the original data into 4 classes, i.e., malfunction, ideal, transition and downturn periods. With meteorological records, it is possible to calculate the nominal power generation by formulas of photoelectric conversions. Specifically, the generated PV power in an ideal period is supposed to be close to the nominal power generation; the generation in a transition period is slightly lower than the nominal one; the generation in a downturn period is relatively low but reasonable (due to too-much light barriers or aging panels); and the generation values in a malfunction period are extremely larger or lower than the nominal values. Such manual divisions are conducted based on expert knowledge and prior experience. We label the malfunction, ideal, transition and downturn periods with the class indices 0, 1, 2, and 3, respectively. Then, we split 75% for training classification models and the rest for testing. We apply five-fold cross validation to search the optimal parameters that show the highest performance. Figure 8 shows the confusion matrix (CM) of all classes in test dataset. Tables 6–12 details the performance metrics (precision, recall, f1-score) of each compared algorithm and our method. From Figure 8a–g and Tables 6–12, we can conclude that our method achieves the best classification performance with the highest averaged precision 0.94, recall 0.93 and f1-score 0.93. Moreover, the other compared methods are far behind, which validates the superiority of our method.

In addition, we assess model performance by consulting engineers and judge whether the proposed method gives right performance evaluation and accurate fault alarm. We apply the proposed method to the monitoring system of our studied PV plant. According to the feedback from their on-site engineers, our method achieves accurate performance evaluation and fast fault detection. First, our method is able to present instantaneous evaluation for each real-time observation. With the assistance of our method, the O&M engineers do not have to keep their eyes on the curves, and they only check the database when Stages 1 and 4 appear. Second, our method is able to detect both direct and indirect faults in a PV system. It presents an accurate classification and seldom misses potential anomalous situation, which greatly enhances the operation safety and maintenance efficiency. More results and analyses are presented in the Supplementary File section.

Although our proposed method runs well in a practical PV plant, from Figure 7a,b we can tell that there are still a few unusual observations in early morning and late afternoon when illumination intensity is quite weak. We plan to improve the robustness of our prediction models as future research, so that they can make more accurate prediction even when the power output is pretty low.

4.3. Discussion

In practical scenarios of PV stations, direct faults, like open circuit and transmission errors, are comparatively easy to notice in the monitoring system. There is an abrupt shift from previous trend. Among indirect factors, encapsulation or module degradation is common in the life cycle of PV panels, which is unavoidable. Therefore, the difficult task of O&M in a PV plant is to intelligently implement panel cleaning, including dust removal and anti-blocking. Compared to direct faults, shade reduces a small amount of output power, which is hard to be detected. In the past, the cleaning O&M of PV panels was mainly periodically manual or robotic cleaning, such as once a month or once a week. Now with the proposed method, which evaluates the state of PV panels and provides instantaneous alarm of degradation, the cleaning maintenance is triggered only when needed. Furthermore, the proposed framework can easily detect direct PV faults and offer timely O&M suggestions.

5. Conclusions

This paper presents an O&M framework consisting of an intelligent detection structure which can enhance the O&M efficiency in the PV monitoring system and reduce the burden on monitoring staff. Our method evaluates operating performance and identifies anomalies by comparing to two reference baselines, which is an unsupervised way and exerts no dependence on labeled faulty data. Moreover, considering the special characteristic of non-continuity in PV generation, we build corresponding non-continuous regression models, which are based on XGBoost algorithm and refined by the results of k-means clustering. Last, by comparing the real-time measured value with both the upper and lower references, our method is sensitive to indirect faults and can provide instantaneous alarm of degradation. Results on a 6.95 MW PV plant indicate that the proposed method is able to evaluate different operating statuses and provide faults identification and O&M suggestions to engineers.

Our work focuses on performance monitoring, fault detection and diagnosis, and O&M optimization in large complex systems. With proper data of ideal and downturn periods, the proposed method can be easily applied to other similar engineering scenarios, such as the assessment of workshop equipment and fault detection in wind power plants. Moreover, our method can be transferred to the application of RUL (remaining useful life) [53] prediction and equipment's PHM (prognostic and health management) [54]. In future studies, we plan to concentrate on the classification refinement of detected faults and predictive maintenance based on the proposed method.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/pr9101711/s1. More details about the proposed method are available online at https://www.dropbox.com/s/a4kz5sjns5au4l9/SupplementaryFile-Intelligent%20and%20Data-driven%20Fault%20Detection%20of%20Photovoltaic%20Plants.pdf?dl=0 (accessed on 13 September 2021) or https://pan.baidu.com/s/1u2oX7XutiDDlE_fURqLVYQ (code: fozu) (accessed on 13 September 2021)

Author Contributions: Conceptualization, S.Y. and Q.K.; Methodology, S.Y.; Software, S.Y.; Validation, S.Y.; Formal analysis, Q.K.; Investigation, S.Y.; Resources, Q.K.; Data curation, S.Y.; Writing—Original Draft Preparation, S.Y.; Writing—Review & Editing, M.Z., A.A. and Y.A.-T.; Visualization, S.Y.; Supervision, Q.K. and M.Z.; Funding acquisition, Q.K., A.A. and Y.A.-T. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the China Scholarship Council Scholarship, in part by the National Natural Science Foundation of China under Grant 51775385, 61703279 and 71371142, the Strategy Research Project of Artificial Intelligence Algorithms of Ministry of Education of China, and in part by Innovation Program of Shanghai Municipal Education Commission under Grant 202101070007E00098, and in part the Deanship of Scientific Research (DSR) at King Abdulaziz University under grant no. RG-22-135-41.

Acknowledgments: We sincerely acknowledge the previous researchers for their excellent work, which greatly assisted our academic study. We are also grateful for the efforts from our colleagues in the Sino-German Center of Intelligent Systems, Tongji University.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Benedek, J.; Sebestyen, T.; Bartók, B. Evaluation of renewable energy sources in peripheral areas and renewable energy-based rural development. *Renew. Sustain. Energy Rev.* **2018**, *90*, 516–535. [CrossRef]
2. Harrou, F.; Sun, Y. *Advanced Statistical Modeling, Forecasting, and Fault Detection in Renewable Energy Systems*; InTechOpen: London, UK, 2020. [CrossRef]
3. Nordrum, A. At last, a massive solar park for Egypt: A 1.8-GW, $4 billion solar power plant is coming on line in the Sahara-[News]. *IEEE Spectr.* **2019**, *56*, 8–9. [CrossRef]
4. Schmuecker, J. The carbon-free farm. *IEEE Spectr.* **2019**, *56*, 30–35. [CrossRef]
5. Yi, Z.; Etemadi, A.H. Line-to-Line fault detection for photovoltaic arrays based on multiresolution signal decomposition and two-stage support vector machine. *IEEE Trans. Ind. Electron.* **2017**, *64*, 8546–8556. [CrossRef]
6. Gallardo-Saavedra, S.; Hernandez-Callejo, L.; Duque-Perez, O.; Hermandez, L. Image resolution influence in aerial thermographic inspections of photovoltaic plants. *IEEE Trans. Ind. Inform.* **2018**, *14*, 5678–5686. [CrossRef]
7. Dotenco, S.; Dalsass, M.; Winkler, L.; Wurzner, T.; Brabec, C.J.; Maier, A.; Gallwitz, F. Automatic detection and analysis of photovoltaic modules in aerial infrared imagery. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 March 2016; pp. 1–9.
8. Tadj, M.; Benmouiza, K.; Cheknane, A.; Silvestre, S. Improving the performance of PV systems by faults detection using GISTEL approach. *Energy Convers. Manag.* **2014**, *80*, 298–304. [CrossRef]
9. Heinrich, M.; Meunier, S.; Samé, A.; Quéval, L.; Darga, A.; Oukhellou, L.; Multon, B. Detection of cleaning interventions on photovoltaic modules with machine learning. *Appl. Energy* **2020**, *263*, 114642. [CrossRef]
10. Leloux, J.; Narvarte, L.; Desportes, A.; Trebosc, D. Performance to Peers (P2P): A benchmark approach to fault detections applied to photovoltaic system fleets. *Sol. Energy* **2020**, *202*, 522–539. [CrossRef]
11. Habault, G.; Lefrancois, M.; Lemercier, F.; Montavont, N.; Chatzimisios, P.; Papadopoulos, G.Z. Monitoring traffic optimization in a smart grid. *IEEE Trans. Ind. Inform.* **2017**, *13*, 3246–3255. [CrossRef]
12. Platon, R.; Martel, J.; Woodruff, N.; Chau, T.Y. Online fault detection in PV systems. *IEEE Trans. Sustain. Energy* **2015**, *6*, 1200–1207. [CrossRef]
13. Hariharan, R.; Chakkarapani, M.; Ilango, G.S.; Nagamani, C. A method to detect photovoltaic array faults and partial shading in PV systems. *IEEE J. Photovolt.* **2016**, *6*, 1278–1285. [CrossRef]
14. Sarikh, S.; Raoufi, M.; Bennouna, A.; Benlarabi, A.; Ikken, B. Implementation of a plug and play I-V curve tracer dedicated to characterization and diagnosis of PV modules under real operating conditions. *Energy Convers. Manag.* **2020**, *209*, 112613. [CrossRef]
15. Gokmen, N.; Karatepe, E.; Silvestre, S.; Celik, B.; Ortega, P. An efficient fault diagnosis method for PV systems based on operating voltage-window. *Energy Convers. Manag.* **2013**, *73*, 350–360. [CrossRef]
16. Fadhel, S.; Delpha, C.; Diallo, D.; Bahri, I.; Migan, A.; Trabelsi, M.; Mimouni, M. PV shading fault detection and classification based on I-V curve using principal component analysis: Application to isolated PV system. *Sol. Energy* **2019**, *179*, 1–10. [CrossRef]
17. Li, C.; Yang, Y.; Zhang, K.; Zhu, C.; Wei, H. A fast MPPT-based anomaly detection and accurate fault diagnosis technique for PV arrays. *Energy Convers. Manag.* **2021**, *234*, 113950. [CrossRef]
18. Ali, M.H.; Rabhi, A.; El Hajjaji, A.; Tina, G.M. Real time fault detection in photovoltaic systems. *Energy Procedia* **2017**, *111*, 914–923. [CrossRef]
19. Triki-Lahiani, A.; Abdelghani, A.B.-B.; Slama-Belkhodja, I. Fault detection and monitoring systems for photovoltaic installations: A review. *Renew. Sustain. Energy Rev.* **2018**, *82*, 2680–2692. [CrossRef]
20. Garoudja, E.; Harrou, F.; Sun, Y.; Kara, K.; Chouder, A.; Silvestre, S. Statistical fault detection in photovoltaic systems. *Sol. Energy* **2017**, *150*, 485–499. [CrossRef]
21. Vergura, S.; Acciani, G.; Amoruso, V.; Patrono, G.E.; Vacca, F. Descriptive and inferential statistics for supervising and monitoring the operation of PV plants. *IEEE Trans. Ind. Electron.* **2008**, *56*, 4456–4464. [CrossRef]
22. Wang, Z.; Li, L.; Yang, X.; Guan, M.; Li, Y.; Zhou, B. Fault diagnosis and operation and maintenance of PV components based on BP neural network with data cloud acquisition. *IOP Conf. Ser. Earth Environ. Sci.* **2019**, *227*, 052063. [CrossRef]
23. Hussain, M.; Dhimish, M.; Titarenko, S.; Mather, P. Artificial neural network based photovoltaic fault detection algorithm integrating two bi-directional input parameters. *Renew. Energy* **2020**, *155*, 1272–1292. [CrossRef]
24. Aziz, F.; Haq, A.U.; Ahmad, S.; Mahmoud, Y.; Jalal, M.; Ali, U. A novel convolutional neural network-based approach for fault classification in photovoltaic arrays. *IEEE Access* **2020**, *8*, 41889–41904. [CrossRef]
25. Huang, J.-M.; Wai, R.-J.; Gao, W. Newly-Designed fault diagnostic method for solar photovoltaic generation system based on IV-Curve measurement. *IEEE Access* **2019**, *7*, 70919–70932. [CrossRef]
26. Momeni, H.; Sadoogi, N.; Farrokhifar, M.; Gharibeh, H.F. Fault diagnosis in photovoltaic arrays using GBSSL method and proposing a fault correction system. *IEEE Trans. Ind. Inform.* **2020**, *16*, 5300–5308. [CrossRef]

27. Ma, J.; Pan, X.; Man, K.L.; Li, X.; Wen, H.; Ting, T.O. Detection and assessment of partial shading scenarios on photovoltaic strings. *IEEE Trans. Ind. Appl.* **2018**, *54*, 6279–6289. [CrossRef]
28. Chen, Z.; Han, F.; Wu, L.; Yu, J.; Cheng, S.; Lin, P.; Chen, H. Random forest based intelligent fault diagnosis for PV arrays using array voltage and string currents. *Energy Convers. Manag.* **2018**, *178*, 250–264. [CrossRef]
29. Zhao, Y.; Yang, L.; Lehman, B.; de Palma, J.-F.; Mosesian, J.; Lyons, R. Decision tree-based fault detection and classification in solar photovoltaic arrays. In Proceedings of the 2012 Twenty-Seventh Annual IEEE Applied Power Electronics Conference and Exposition (APEC), Orlando, FL, USA, 5–9 February 2012; pp. 93–99.
30. Guerriero, P.; Di Napoli, F.; Vallone, G.; D'Alessandro, V.; Daliento, S. Monitoring and diagnostics of PV plants by a wireless self-powered sensor for individual panels. *IEEE J. Photovolt.* **2015**, *6*, 286–294. [CrossRef]
31. Abusorrah, A.M.; Al-Turki, Y.A.; Kang, Q.; Yao, S.; Zhou, M. Monitoring and Fault Detection Method and System for Photovoltaic Plants. U.S. Patent No 10,826,428, 3 November 2020.
32. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [CrossRef] [PubMed]
33. Satpathy, P.R.; Sharma, R. Reliability and losses investigation of photovoltaic power generators during partial shading. *Energy Convers. Manag.* **2020**, *223*, 113480. [CrossRef]
34. Zhao, Y.; Li, D.; Lu, T.; Lv, Q.; Gu, N.; Shang, L. Collaborative fault detection for large-scale photovoltaic systems. *IEEE Trans. Sustain. Energy* **2020**, *11*, 2745–2754. [CrossRef]
35. Lloyd, S. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **1982**, *28*, 129–137. [CrossRef]
36. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining KDD-96, Portland, OR, USA, 2–4 August 1996; pp. 226–231.
37. Yao, S.; Pan, L.; Yu, Z.; Kang, Q.; Zhou, M. Hierarchically non-continuous regression prediction for short-term photovoltaic power output. In Proceedings of the 2019 IEEE 16th International Conference on Networking, Sensing and Control (ICNSC), Banff, AB, Canada, 9–11 May 2019; pp. 379–384.
38. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [CrossRef]
39. Khalil, I.U.; Ul-Haq, A.; Mahmoud, Y.; Jalal, M.; Aamir, M.; Ahsan, M.U.; Mehmood, K. Comparative analysis of photovoltaic faults and performance evaluation of its detection techniques. *IEEE Access* **2020**, *8*, 26676–26700. [CrossRef]
40. Bastidas-Rodriguez, J.D.; Franco, E.; Petrone, G.; Ramos-Paja, C.A.; Spagnuolo, G. Model-Based degradation analysis of photovoltaic modules through series resistance estimation. *IEEE Trans. Ind. Electron.* **2015**, *62*, 7256–7265. [CrossRef]
41. Samara, S.; Natsheh, E. Intelligent real-time photovoltaic panel monitoring system using artificial neural networks. *IEEE Access* **2019**, *7*, 50287–50299. [CrossRef]
42. Yang, X.; Wen, W. Ridge and lasso regression models for cross-version defect prediction. *IEEE Trans. Reliab.* **2018**, *67*, 885–896. [CrossRef]
43. Millan-Castillo, R.S.; Morgado, E.; Goya-Esteban, R. On the use of decision tree regression for predicting vibration frequency response of handheld probes. *IEEE Sens. J.* **2019**, *20*, 4120–4130. [CrossRef]
44. Pan, C.; Tan, J.; Feng, D. Identification of power quality disturbance sources using gradient boosting decision tree. In Proceedings of the 2018 Chinese Automation Congress (CAC), Xi'an, China, 30 November–2 December 2018; pp. 2589–2592.
45. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
46. Kang, Q.; Shi, L.; Zhou, M.; Wang, X.; Wu, Q.; Wei, Z. A distance-based weighted undersampling scheme for support vector machines and its application to imbalanced classification. *IEEE Trans. Neural Networks Learn. Syst.* **2018**, *29*, 4152–4165. [CrossRef]
47. Zhang, H.; Li, Y.; Lv, Z.; Sangaiah, A.K.; Huang, T. A real-time and ubiquitous network attack detection based on deep belief network and support vector machine. *IEEE/CAA J. Autom. Sin.* **2020**, *7*, 790–799. [CrossRef]
48. Zhang, P.; Shu, S.; Zhou, M. An online fault detection model and strategies based on SVM-grid in clouds. *IEEE/CAA J. Autom. Sin.* **2018**, *5*, 445–456. [CrossRef]
49. Shen, Z.; Elibol, A.; Chong, N.Y. Understanding nonverbal communication cues of human personality traits in human-robot interaction. *IEEE/CAA J. Autom. Sin.* **2020**, *7*, 1465–1477. [CrossRef]
50. Tao, G.; Zheng, Z.; Guo, Z.; Lyu, M.R. MalPat: Mining patterns of malicious and benign android apps via permission-related APIs. *IEEE Trans. Reliab.* **2018**, *67*, 355–369. [CrossRef]
51. Punmiya, R.; Choe, S. Energy theft detection using gradient boosting theft detector with feature engineering-based preprocessing. *IEEE Trans. Smart Grid* **2019**, *10*, 2326–2329. [CrossRef]
52. Zhang, D.; Qian, L.; Mao, B.; Huang, C.; Huang, B.; Si, Y. A data-driven design for fault detection of wind turbines using random forests and XGboost. *IEEE Access* **2018**, *6*, 21020–21031. [CrossRef]
53. Wang, B.; Lei, Y.; Li, N.; Li, N. A hybrid prognostics approach for estimating remaining useful life of rolling element bearings. *IEEE Trans. Reliab.* **2020**, *69*, 401–412. [CrossRef]
54. Guo, J.; Li, Z.; Li, M. A review on prognostics methods for engineering systems. *IEEE Trans. Reliab.* **2020**, *69*, 1110–1129. [CrossRef]