An Enhanced Pre-processing and Nonlinear Regression Based Approach for Failure Detection of PV System

Chung-Chian Hsu[1](✉), Jia-Long Li[1], Arthur Chang[1],
and Yu-Sheng Chen[2]

[1] Big Data Research Center, National Yunlin University
of Science and Technology, Yunlin, Taiwan
hsucc@yuntech.edu.tw
[2] Reforecast Co., Ltd., Taipei, Taiwan
anson@reforecast.com.tw

Abstract. The solar energy is getting popular due to the awareness of the environmental issues. Multiple module strings are set up in a solar-power plant to increase power production which is sold to electricity company via connected grid. Inevitably, devices can break, leading to loss of power production. To minimize the loss, it is important to be able to detect faulty devices as soon as possible for maintenance. In this paper, an approach relying on careful data pre-processing is proposed and compares with an existing approach.

Keywords: Solar energy · Fault detection · Machine learning · *k*-nearest neighbors

1 Introduction

The photovoltaic (PV) power plants have grown rapidly in the last decade due to large demand on solar energy, which becomes an indispensable resource of human culture [1, 2]. All around the world, corporations as well as governments are exploiting solar energy market [3]. In order to respond to ever increased demand on energy consumption, many countries began to address the problem of energy production. As a result, the green energy has attracted attention from many governments. Green energy includes the wind power, the tidal power, solar energy, etc.

Various aspects of PV systems have been explored as discussed in a number of review papers [4–6, 10], including fault detection, diagnosis, prediction and degradation. Despite PV arrays have protective arrangement, there may still undetected faults occurring on devices [9]. To prevent from continuous loss of power production, it is important to detect faulty devices as soon as possible so that maintenance can take place.

Many methods have been proposed to detect faulty devices, and yet they usually took lots of time and required high-quality data. For instance, a fault detection and diagnosis of a grid-connected PV system approach based on PNN classifier was presented by Garoudja et al. in [4]. The fault detection of PV systems based on local outlier factor was proposed by Ding et al. in [5]. They preprocessed data and identified

C.-Y. Chang et al. (Eds.): ICS 2018, CCIS 1013, pp. 260–269, 2019.
https://doi.org/10.1007/978-981-13-9190-3_27

normal and faulty data before constructing models. Various approaches based on the K-Nearest-Neighbor technique have been proposed in recent papers like [6, 7], which compared real-time data read from the device in the solar-power plant with similar historical data. This approach can fail when historical data themselves are abnormal due to long-time faulty devices or rainy days.

PR is an important indicator of production performance of PV modules [8], which measures efficiency of solar power production within a certain time period. In contrast, array ratio (RA) is an instant indicator which measures production performance in real-time. In this study, we intend to detect faulty devices as soon as possible. Therefore, RA is adopted. For normal devices, RA values shall be stable. However, several factors such as module brand, module degradation, bad weather, faulty modules in a string, cloud, etc. can affect the values more or less. As a result, the RA value threshold setting for judging whether a solar device is faulty or not becomes a non-trivial issue.

In this paper, we propose a fault detection approach which exploits statistics median and nonlinear regression based on the plant-owned data collected from individual solar-power plants. In such a way, each plant has its own threshold which is adapted to the plant's conditions. In particular, the RA values of all devices are calculated and sorted. To avoid affecting by faulty devices, only a certain range around the median of the RAs is considered. Then, we use nonlinear regression on training, historical data to obtain the upper and the lower boundary of the RA for a device to be considered normal. Devices continuously have RA values out of the range are claimed faulty once detected.

2 Data Acquisition of PV Systems

The PV data were collected from several solar power plants located in central Taiwan. Data used in the analysis of this study include date, time, irradiance, power, current, voltage, and capacity of individual module strings. Irradiance, current, voltage, and power were sampled every five minutes. One pyrheliometer was installed in each plant for measuring irradiance. The angle of the pyrheliometer is horizontal in some plants and the same with the angle of panels along the roof in the others. The period of historical data analyzed in this study was between 2018/08/01 and 2018/09/10. The capacity of individual strings varies. Therefore, when calculating power production efficiency of the strings, the capacity shall be taken into consideration.

3 Method

The proposed approach is based on the idea that normal RA values shall fall in a certain range under normal conditions. However, many unexpected factors can lead to abnormal RA values which can prevent us from estimating the proper range of normal RAs. To tackle the problems, enhanced preprocessing is suggested and the nonlinear regression algorithm is used to estimate the RA range. The devices which RAs exceed the range are considered faulty. The architecture of the proposed method is showed in Fig. 1.

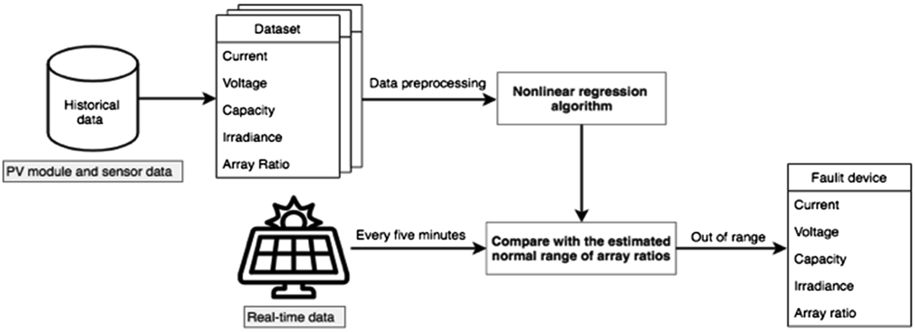

Fig. 1. Architecture of the fault detection system.

3.1 PV Module Formula

Many studies used one diode model (ODM) as shown in Fig. 2 to simulate the power production of the PV system. In this model, the output current is given by Eq. (1). As shown, the output estimation requires several parameters which may not be obtained easily.

$$I_{PV} = I_{ph} - \underbrace{I_0\left(exp\left(\frac{q\left(V_{pv}+R_S I_{pv}\right)}{nK_B T}\right)-1\right)}_{I_d} - \underbrace{\frac{V_{pv}+R_S I_{pv}}{R_{sh}}}_{I_{sh}} \tag{1}$$

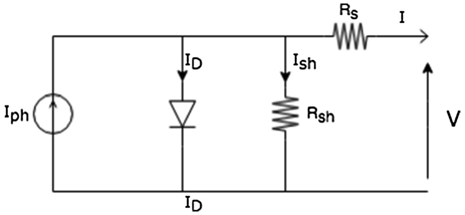

Fig. 2. Include new data and old data about Irradiance distribution of a day.

In the study, we evaluate the performance of PV modules by array ratio or RA, which is expressed as in Eq. (2).

$$Array\,Ratio = \frac{P_{DC}/P_0}{G_1/G_0} \tag{2}$$

where P_{DC} and P_0 divides to real-time of power and system rated of power; G_1 and G_0 divides to project plane of radiance and standard strong of irradiance. In this work, array ratio is for references to evaluate device that is normal or faulty device.

3.2 Data Pre-processing

The RA value is very sensitive to dramatic change of irradiance. Therefore, the preprocessing includes two steps: First, the removal of data points which have a dramatic change of irradiance at the times. Second, the removal of data points which production efficiency is abnormal, much higher or lower than most of the rest devices. We divide the data points into those collected in the morning and those in the afternoon, respectively.

As can be seen in Fig. 3, an abrupt drop in irradiance can yield extremely high RA. To avoid this, we replace the points with smooth points.

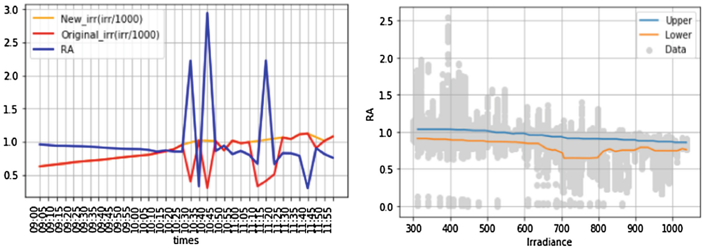

Fig. 3. Abnormal irradiances in the left diagram are smoothed by neighbor normal points and RA outliers shown in the right diagram can occur occasionally.

In addition, RA outliers can occur due to malfunction of solar devices. To address this problem, we sort RA values of all the devices and consider only the second and the third quartile of the RAs. Not only eliminating abnormal data, it might eliminate normal data at the same time. In the next section, we use nonlinear regression method with some tolerance to avoid deleting too many normal data points.

3.3 Nonlinear Regression Algorithm

To find out the range for detecting faulty devices, the regression on the RAs with respect to various irradiances is used. The polynomial regression algorithm is one of the regression analysis methods in which independent variable x and the dependent variable y are modelled as an nth degree polynomial in x, as expressed in Eq. (3). In our case, the dependent variable is the RA and the independent is the irradiance.

$$y_i = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \ldots + \alpha_n x^n + \varepsilon \tag{3}$$

To avoid eliminating too many normal data points, tolerance shall be added to the regression result. According to empirical results, the new upper and lower bounds is determined by the equations shown in (4) and (5).

$$New(upper_i) = upper_i + (Upper_i - y_i) \tag{4}$$

$$New(lower_i) = lower_i - (y_i - lower_i) \tag{5}$$

3.4 Evaluation

Confusion matrix is used to evaluate the results by comparing to the K-nearest neighbor algorithm. The matrix has four indicators of a true positive (TP), a false negative (FN), a false positive (FP) and a true negative (TN), respectively. Classification accuracy, precision, and recall can be calculated based on those four indicators as shown in Eq. (6)–(8).

$$Classfication\,accuracy = (TP + TN)/(TP + FN + FP + TN) \tag{6}$$

$$Precision = TP/(TP + FP) \tag{7}$$

$$Recall = TP/(TP + FN) \tag{8}$$

4 Experiments

We first present the result of data preprocessing and the result of model construction. Then we compare the proposed approach with the K-Nearest-Neighbor approach [7] on accuracy of diagnosing PV string fault.

4.1 Results of the Data Pre-processing

Table 1. The number of training and test data in the three PV plants

		PV plant 1	PV plant 2	PV plant 3
Training data	Morning	129,652	106,370	41,158
	Afternoon	131,546	98,098	39,068
Test data	Morning	2,548	2,100	1,000
	Afternoon	3,774	3,108	1,200

The training data came from data acquisition of PV between 2018/08/17 and 2018/10/08. The test data were collected on 2018/10/03 as shown in Table 1. As can be seen in Fig. 4, malfunction of the devices can result in outliers which have extreme values, even become zero.

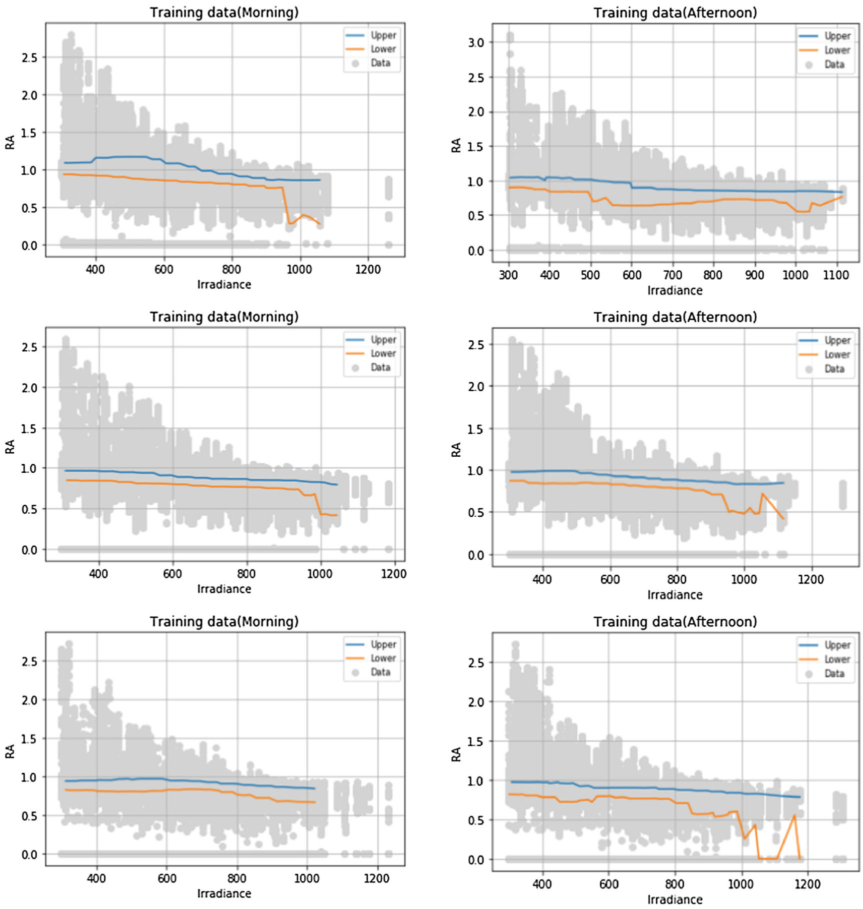

Fig. 4. Distribution of RA with respect to irradiances in the three solar power plants.

4.2 Results of Nonlinear Regression Algorithm

Figure 5 shows the regression result of the training data. According to data distribution, the quadratic regression was chosen. In each diagram, the green and the red dash line indicate the range of RAs considered acceptable. Those outside the boundaries are considered abnormal.

Figure 6 shows the distribution from the test data. The results demonstrated the most of the abnormal data points can be detected based on the boundaries estimated from the training data. Some points which might be normal were out of the boundary, such as those in the second row (Plant 2). However, the later stage which require three consecutive abnormal points for a device to be taken as faulty can avoid misclassification in those cases.

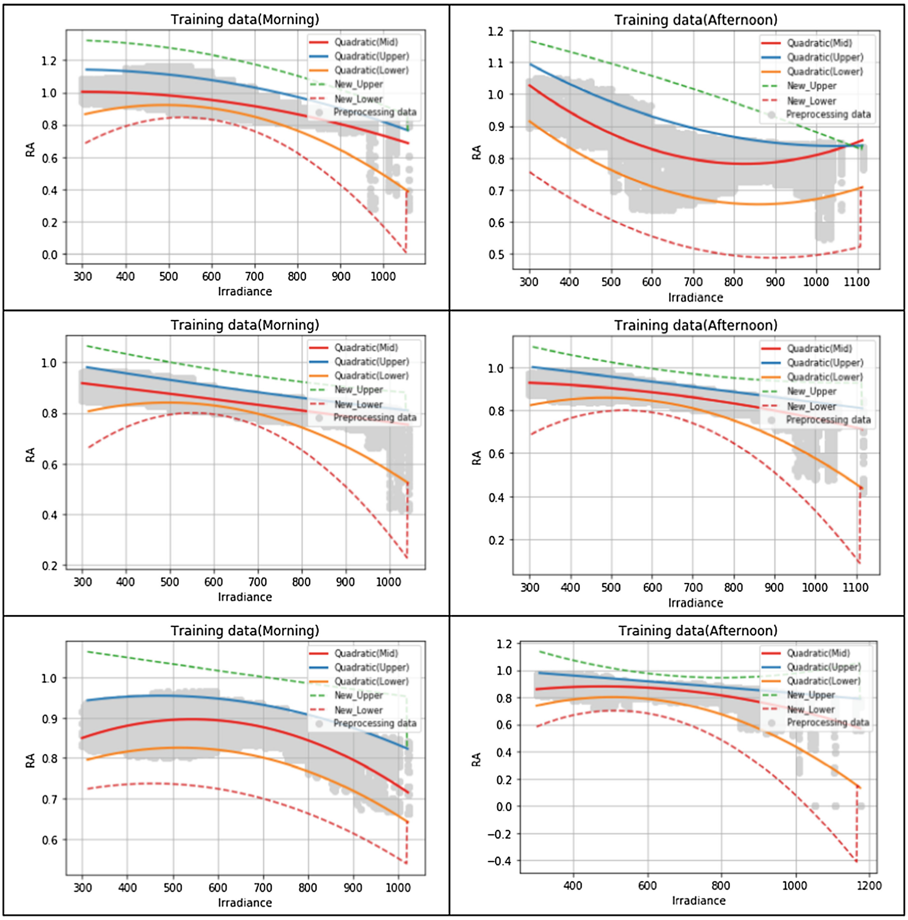

Fig. 5. The upper and lower bounds were estimated from the training data.

4.3 Results of Fault Detection and Diagnosis

To verify the performance, we compare the proposed method with the KNN approach [7] which estimated the power production by the average of the ten nearest neighbors found from the data collected in the past seven days. If the measured power is lower than the estimated by the threshold, 25% used in this study according to experimental results for three consecutive points, the device is thus considered faulty.

Table 2 shows the numbers of individual abnormal data points detected by the nonlinear regression approach is much larger than those by the KNN. Furthermore, the KNN took 2.14 min while the NLR took 0.42 min. We require the number of the consecutive abnormal points must at least be three to avoid false positive alarms. The right part of Table 2 shows the number of the points, representing the number of faulty devices, after merging consecutive points and eliminating the merged points which did not contain more than two consecutive points.

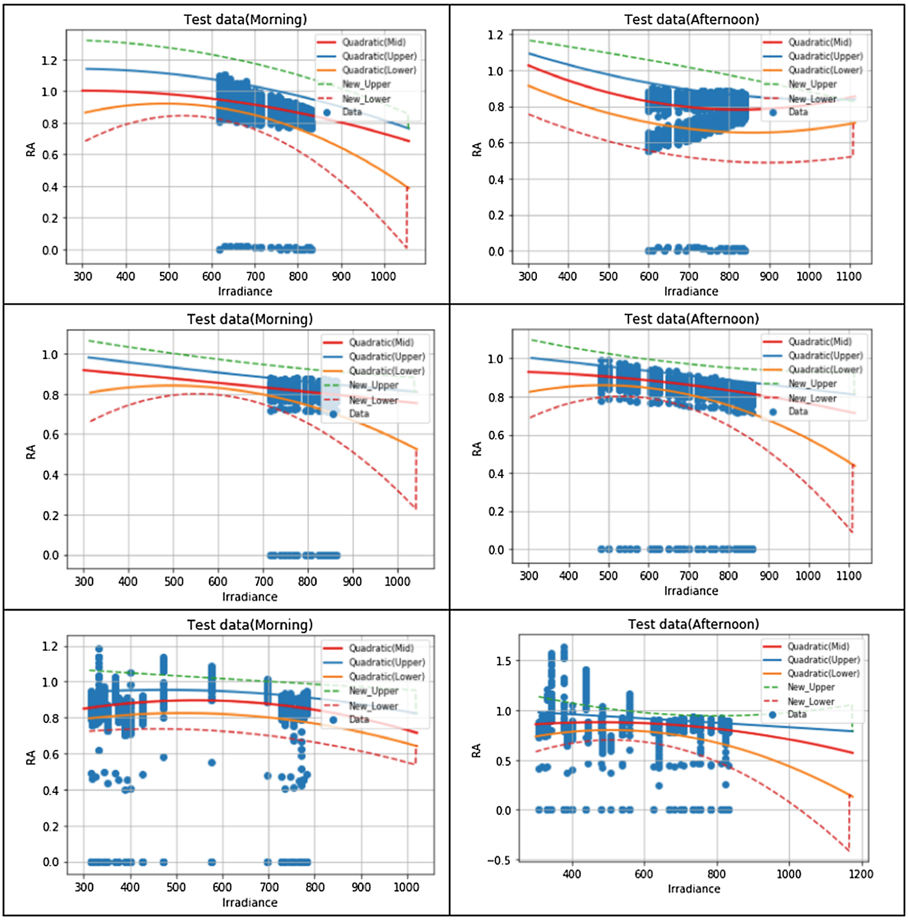

Fig. 6. RA distribution with respect to irradiance in the morning and afternoon, respectively, for Plant 1, 2, and 3.

Table 2. The numbers of the detected abnormal points and the points after merge

	No. of abnormal data points			No. of points after merge		
	Plant 1	Plant 2	Plant 3	Plant 1	Plant 2	Plant 3
KNN	54	0	167	6	0	10
NLR	131	140	438	2	3	30

The ground truth of the device status came from the field engineers which went to the plants to check and maintain, if necessary, the devices. Table 3 presents the diagnosis result and shows none of the methods reported false diagnosis. However, the NN method failed to report 13 faulty devices from the three plants in total. The reason

Table 3. The predicted results by the methods based on the *k*-nearest-neighbor technique (in the upper half) and based on the nonlinear regression technique (in the lower half)

		Predicted class					
		Plant 1		Plant 2		Plant 3	
		Healthy	Faulty	Healthy	Faulty	Healthy	Faulty
Real class	Healthy	5	0	0	0	7	0
	Faulty	0	1	0	0	0	3
Real class	Healthy	0	0	0	0	17	0
	Faulty	0	2	0	2	0	13

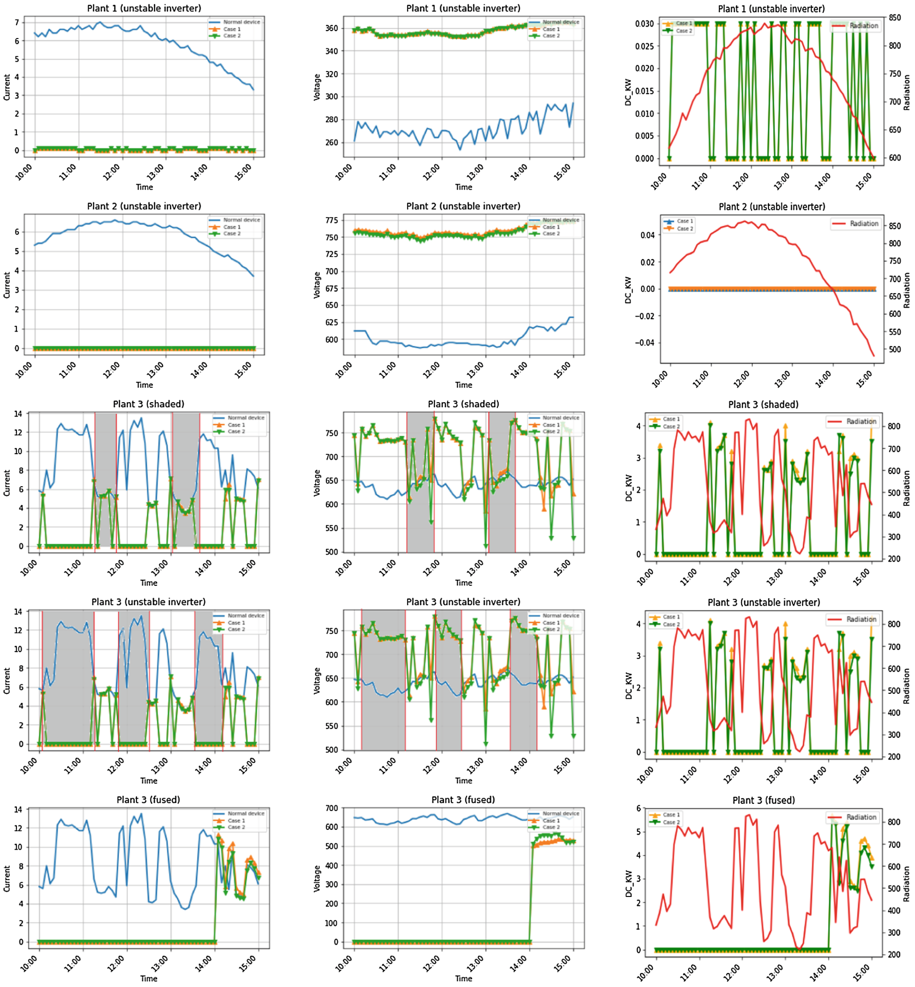

Fig. 7. Currents, voltages, and produced DC_KWs of the devices with various faults.

is that the NN is based on the historical data of the past seven days and if there are bad weather for more than seven days, the method cannot work as expected.

The first two rows in Fig. 7 present that an unstable inverter will have currents close to zero, voltages, higher than normal and unstable or zero DC powers. The last row shows that fused devices have current, voltage, and power all zero. Note that at 2 o'clock the faulty device was repaired.

5 Conclusion

In this paper, a detection method for faulty solar-power strings has been proposed. The RA range within which the values are considered normal is estimated by using nonlinear regression. To evaluate its performance, comparison is made with a method based on the KNN algorithm. The experimental results indicate that the proposed method can detect more faulty devices than the KNN approach, which can break under bad weather lasting for several days. Another advantage is that the proposed model is fast in detection since the model is constructed offline.

Acknowledgement. This work was partially supported by the grants from Ministry of Science and Technology, Reforecast Co., Ltd., and the Intelligent Recognition Industry Service Center.

References

1. Raza, M.Q., Nadarajah, M., Ekanayake, C.: On recent advances in PV output power forecast. Sol. Energy **136**, 125–144 (2016)
2. Europe SP: Global market outlook for solar power 2015–2019 Technical report Bruxelles: European Photovoltaic Industry Association (2015)
3. Oliver, M., Jackson, T.: The market for solar photovoltaics. Energy Policy **27**, 15 (1999)
4. Garoudja, E., Chouder, A., Kara, K., Silvestre, S.: An enhanced machine learning based approach for failures detection and diagnosis of PV systems. Energy Convers. Manage. **151**, 1246–1254 (2017)
5. Ding, H., et al.: Local outlier factor-based fault detection and evaluation of photovoltaic system. Sol. Energy **164**, 139–148 (2018)
6. Madeti, S.R., Singh, S.N.: Modeling of PV system based on experimental data for fault detection using kNN method. Sol. Energy **173**, 139–151 (2018)
7. Hsu, C.-C., Teng, C.-T., Cai, C.-J., Chang, A.: Real-time diagnosis of fault type for grid-connected photovoltaic plants. In: The 29th International Conference on Information Management, 3 June 2017, CSIM, Taichung (2018)
8. Khalid, A.M., Mitra, I., Warmuth, W., Schacht, V.: Performance ratio – crucial parameter for grid connected PV plants. Renew. Sustain. Energy Rev. **65**, 1139–1158 (2016)
9. Pillai, D.S., Rajasekar, N.: A comprehensive review on protection challenges and fault diagnosis in PV systems. Renew. Sustain. Energy Rev. **91**, 18–40 (2018)
10. Pillai, D.S., Rajasekar, N.: Metaheuristic algorithms for PV parameter identification: a comprehensive review with an application to threshold setting for fault detection in PV systems. Renew. Sustain. Energy Rev. **82**, 3503–3525 (2018)