Applied Energy Symposium 2019: Low carbon cities and urban energy systems
October 16-18, 2019, Xiamen, China
Paper ID: 0134

A NOVEL FAULT DIAGNOSIS METHOD FOR PV ARRAYS USING EXTREME GRADIENT BOOSTING CLASSIFIER

Yutao Gan[1], Zhicong Chen[1*], Lijun Wu[1], Chao Long[2], Shuying Cheng[1], Peijie Lin[1]

[1]College of Physics and Information Engineering, Fuzhou University, 2 XueYuan Road, 350116 Fuzhou, China

[2]Institute of Energy, School of Engineering, Cardiff University, Cardiff CF24 3AA, UK

ABSTRACT

A new online fault diagnostic method for photovoltaic array is proposed in this paper, which is based on the Extreme Gradient Boosting (XGBoost) classifier. Firstly, the string current, array voltage, temperature and irradiance are measured by a monitoring system, from which a seven-dimensional fault feature vector is extracted as the input of the fault diagnosis model. Secondly, based on the XGBoost classifier, a new fault diagnosis model is established. Lastly, the feasibility and superiority of the proposed XGBoost based fault diagnosis model are tested by both Simulink based simulation and real fault experiments on a laboratory PV system. The correct rate of fault diagnosis in Simulink simulation is 99.99%, while the correct rate of fault diagnosis in laboratory PV power plant simulation is over 99.90%. Extreme learning machines (ELM) and Random Forests (RF) are tested for comparison. Experimental results demonstrate the superiority of the proposed XGBoost based model.

Keywords: photovoltaic array, fault diagnosis, XGBoost, dynamic operating point

1. INTRODUCTION

With the depletion of traditional fossil energy, solar energy has received extensive attention as a clean energy source. As a core component of photovoltaic power generation systems, photovoltaic arrays are extremely vulnerable to damage due to their installation in harsh external environments. Photovoltaic array fault diagnosis is an important guarantee link to ensure good photovoltaic power generation. At present, many photovoltaic fault diagnosis methods have been proposed [1]. From the source of data information, it can be roughly divided into image-based processing methods and digital-based processing methods. Image-based processing methods include infrared thermal images of photovoltaic arrays [2], images based on voltage and current signal conversion; digital-based processing methods have current-voltage signals based on timing of photovoltaic array output, I-V curve, voltage and current at most power points [3] (MPP); The infrared thermal imaging-based photovoltaic fault diagnosis method can detect the abnormal hot spot of the PV array earlier and has the ability of early fault warning, but it also has the disadvantages of complicated operation and high equipment cost. The image processing method based on the conversion of the voltage and current output signals and the current-voltage signal based on the time change can detect the photovoltaic array in the transient state of the fault [4], but cannot judge the fault in the transient steady state and the steady state photovoltaic array. The I-V curve can completely preserve the information of the PV array. Therefore, the I-V curve based PV fault diagnosis method can diagnose a more comprehensive fault type [5], but the I-V curve needs to be measured offline, which will interrupt the normal operation of the PV power station and affect the power generation efficiency. Therefore, for online fault diagnosis of photovoltaic arrays, a more economical and practical method is based on the output current and voltage of the photovoltaic array, and machine learning to establish a fault diagnosis model online diagnosis. At present, many methods of machine learning have been applied to photovoltaic fault diagnosis. Support Vector Machine

* Corresponding author.
E-mail address: zhicong.chen@fzu.edu.cn (Zhicong Chen)

Selection and peer-review under responsibility of the scientific committee of CUE2019

(SVM) [6], Probabilistic neural network (PNN) [7], Decision Tree (DT) [8]. These machine learning-based methods have a single algorithm, which is easy to produce over-fitting and fall into local minimum. In order to overcome the shortcomings of only a single algorithm, integrated learning has been proposed in recent years. The integrated learning method refers to combining multiple learning models to achieve better results, and the combined model has stronger generalization ability. XGBoost is one of the boosting algorithms belong to integrated learning. The idea of the Boosting algorithm is to integrate many weak classifiers into a strong classifier. Because XGBoost is a lifting tree model, it integrates many tree models to form a strong classifier. Therefore, in this paper, we collect the original data extraction features such as the voltage and current at MPP of the PV array, and then use the XGBoost classifier for fault diagnosis.

2. FAULT DIAGNOSIS OF PV ARRAY USING XGBOOST

The main idea of integrated learning is to repeatedly train multiple models and combine them in a certain way to form a high-performance and powerful integrated model. In the boost algorithm system, a series of iterative serial forms are generally generated. Models, then linearly add these models to get the final integrated learner.

If the weak prediction model of each step of the boost algorithm is generated according to the gradient direction of the loss function, it is called gradient boosting, which can reduce the risk of over-fitting and realize the generation of weak learner. And the XGBoost algorithm does not use the search method but directly utilizes the first derivative value of the loss function. And through the techniques of pre-sorting, weighted quantile , the performance of the algorithm is greatly improved[9].

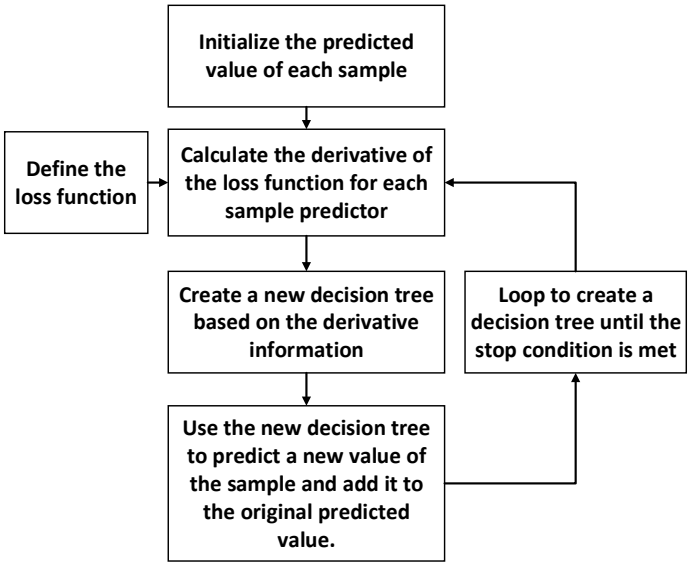

Fig 1 Process of creating a decision tree of XGBoost.

2.1 Data preprocessing and extraction of fault features

The idea of the data-based photovoltaic fault diagnosis method is that the output characteristics of the photovoltaic array are different from the normal state in the fault state, and the output power is often lower than the normal state when the fault occurs, which may cause the photovoltaic power station to have low power generation efficiency, which is also required for the photovoltaic array, and it is one of the important reasons for fault diagnosis research.

The photovoltaic array is composed of a plurality of solar cells arranged in a certain order. Solar cells work with the photovoltaic effect and require the presence of sunlight, while the output of the solar cells is also affected by temperature. For photovoltaic modules, the relationship of the short current, open voltage between the irradiance and temperature can be expressed by[10, 11]

$$I_{sc} = I_{sc\_stc}[1+\alpha(T_a-T_{stc})]\, G_a/G_{stc} \quad (1)$$

$$V_{oc} = V_{oc\_stc}[1+\beta(T_a-_{stc})]G_a/G_{stc}+nU_t ln(G_a/G_{stc}) \quad (2)$$

The current of maximum power point (I_{mpp}), voltage of maximum power point (V_{mpp}), and Temperature (T_a), irradiance (G_a), voltage, current, temperature and irradiance of standard test condition (V_{stc}, I_{stc}, T_{stc}, G_{stc}) of the PV array need to collect to extract features. The standard test conditions refer to an environment with a temperature of 25 ° C and an irradiance of 1000 W/m^2. And the fault occurs in different string reflect the different in the current between normal string and the fault sting of the PV array, so a feature needed to extract reflect the difference between each string currents. And the PV arrays of different manufacturers have different output characteristics. According to Eqs (1) and Eqs (2), need to collect the ideal factor n of the photovoltaic, the open circuit voltage temperature coefficient β, the short-circuit current temperature coefficient α, and the thermal voltage U_t, the number of solar cell strings in parallel in a solar cell array N_p, the number of solar cells in series in a solar cell string N_s, and the mean of the string current μ and the standard deviation of string current σ, According to these data and the Eqs (1) and Eqs (2), we define the Eqs (3) and Eqs (4) indicates the relationship between dynamic operating point voltage and current and irradiance temperature:

$$V_{op}=N_s V_{stc}[1+\beta(T_a-T_{stc})]G_a/G_{stc}+nU_t ln(G_a/G_{stc}) \quad (3)$$

$$I_{op} = N_p\, I_{stc}[1+\alpha(T_a-T_{stc})]\, G_a/G_{stc} \quad (4)$$

with these row data and formula, the following new seven-dimensional features vector is extracted and normalized eliminate the effects of changes in temperature irradiance with the Eqs (5) - (11):

$$V_n=V_{mpp}/V_{op} \quad (5)$$

$I_n = I_{mpp}/I_{op}$ (6)

$P_n = V_{mpp}I_{mpp}/V_{op}I_{op}$ (7)

$S_n = I_{mpp}V_{op}/I_{op}V_{mpp}$ (8)

$C_x = \mu/\sigma$ (9)

$G_n = G_a/G_{stc}$ (10)

$T_n = T_a/T_{stc}$ (11)

2.2 Extreme Gradient Boosting based fault detection and diagnosis model

According to the seven features selected in section 2.1 as the input of fault diagnosis model, and the fault diagnosis model based on XGBoost classifier is established for photovoltaic fault diagnosis. The model has seven inputs and one output, it can be directly used as the fault diagnosis model to predict the class of the unlabeled data. The process of establishing the fault diagnosis model is illustrated in follow step:

Step 1: Simulate a variety of different fault conditions and normal operating conditions of photovoltaic arrays in Simulink and laboratory photovoltaic power plants, and collect data such as PV array current, voltage, temperature and irradiance under various operating conditions as raw data;

Step 2: Extract the characteristics according to the raw data collected in step 1 and the specification data of the studied PV array according to section 2.1;

Step 3: The extracted feature samples are randomly divided into two parts, one of which accounts for 70% of the total sample as a training sample, and the other part accounts for 30% of the total sample as a test sample;

Step 4: The training samples are used to optimize and train the XGBoost classifier. After training the model, test samples are used to test the performance of the XGBoost classifier.

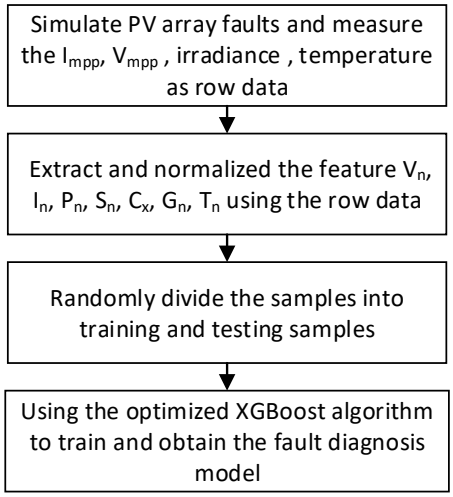

Fig 2 Brief flowchart of building the XGBoost based fault diagnostic model.

2.3 Training of the fault diagnosis model

In this study, the training samples were randomly divided into training set and verification set by 5-fold cross-validation, used for parameter optimization of XGBoost, selected different parameter values to be trained with training set, and then used verification set to verify the performance of using the parameter model.

The following are additional parameters used by XGBoost and their meanings：

The “max_depth” parameter defines the maximum depth of the decision tree.

The “min_child_weight” parameter defines the minimum leaf node weight sum, if in a split, the weight of all samples on the leaf node is less than min_child_weight, the splitting is stopped, which can effectively prevent over-fitting and prevent special samples from being learned.

The “subsample” parameter represents the proportion of samples randomly sampled per tree, reducing the value of this parameter, the algorithm will be more conservative, avoiding overfitting. However, if this value is set too small, it may cause an under-fitting.

The “colsample_bytree” parameter corresponds the percentage of columns used to control each random sample, each column is a feature.

The value of the above parameters will affect the quality of the model. Therefore, we need to adopt a more convenient and efficient parameter optimization method. The following is our parameter optimization method in model training in this study:

Step 1: The initialization parameter value is set to the default value. The author of XGBoost gives the default value of the parameter when creating the algorithm, and gives the adjustment interval of the parameter, which is convenient for the user to learn and use XGBoost. The default parameter value is defined as follows: “max_depth " = 6, "min_child_weight" = 1, "subsample" = 1, "colsample-bytree" = 1.

Step 2: Keep the other parameters unchanged, first adjust “max_depth”, adjust it near the default value, and adjust the direction with the correct rate until the parameter with the highest correct rate is found.

Step 3: Keep the parameter "max_depth" unchanged, then adjust "min_child_weight", adjust it near the default value, and adjust the direction with the correct rate until the parameter value of the highest correct rate is found.

Step 4: Keep the parameter "min_child_weight" unchanged, then adjust "subsample", adjust it near the default value, and adjust the direction with the correct

rate until the parameter value with the highest correct rate is found.

Step 5: Keep the parameter "subsample" unchanged, and finally adjust "colsample_bytree", adjust it near the default value, and adjust the direction as the correct rate increases until the parameter value with the highest correct rate is found.

Step 6: Use the above parameters as the optimal parameters and train the XGBoost model with training samples.

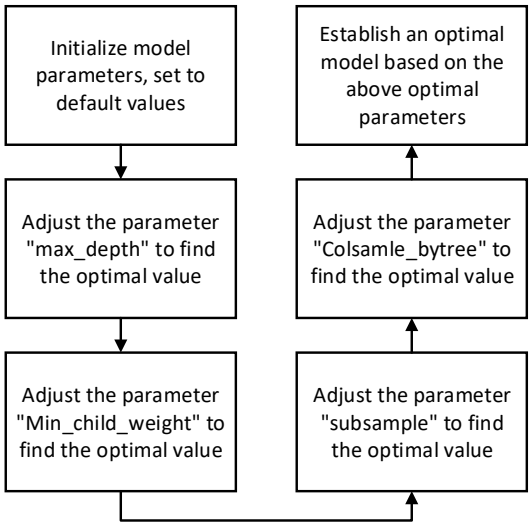

Fig 3 Flowchart of parameter optimization for the XGBoost model.

2.4 Simulation study

In this experiment, the data was collected by Simulink simulation at irradiance 100-975W/m^2 (take a data value every 25 W/m^2), with different combination temperature 25-70 °C (take a data value every 2.5 °C), simulate seven kinds under working conditions: normal condition(N), line-line fault of string level with one module difference in the same string(LL1) is simulated by connected a resistance between modules which is 0.001 ohms, line-line fault of array level with two modules difference(LL2) at two difference PV string is simulated by connected a resistance between modules which is 0.001 ohms, open-circuit fault on one string(OC) is simulated by connect a series resistor of 40000 ohms into the negative of the PV array, partial shading fault(PS) which is simulated through setting the irradiance gains K=0.5, degradation fault of PV array level(DA) is simulated by connect 4 ohm resistor to the negative output of the PV array, degradation fault of PV string level(DS) is simulated by connect 4 ohm resistors between a string of PV and the negative of the PV array. The data of each group of working conditions are 684 groups, a total of 4788 sets of data, according these data and parameters of photovoltaic models, features such as V_n, I_n, P_n, S_n, C_x, G_n, T_n extracted as the sample data of photovoltaic fault diagnosis model. The sample data is randomly divided into a training samples and a test samples, where in the training samples accounts for 70% of the total data, and the test samples accounts for 30% of the total data samples. The training samples are divided into training sets and verification sets by 5-fold crossover to search for optimal parameters of XGBoost. In order to avoid contingency, each parameter combination runs 50 times and the result is averaged, and then select the best combination of parameters as the optimal parameters for XGBoost. In this experiment, based on the data of Simulink, the results of finding the optimal parameters of XGBoost by K-fold are shown in Table 1 to Table 4.

Table 1

The correct rate of the model is when adjusting the "max_depth" parameter (50 times) for the simulation data sample

parameter	Max_depth				
Value	1	3	5	7	9
Train accuracy	99.63	99.89	99.91	99.80	99.78
Test accuracy	99.63	99.91	99.92	99.81	99.80

Table 2

The correct rate of the model is when adjusting the "min_child_weight" parameter (50 times) for the simulation data sample

parameter	Min_child_weight				
Value	0.5	1	3	5	7
Train accuracy	99.76	99.93	99.79	99.72	99.31
Test accuracy	99.72	99.92	99.80	99.74	99.32

Table 3

The correct rate of the model is when adjusting the "subsample" parameter (50 times) for the simulation data sample

parameter	Subsample			
Value	0.3	0.5	0.7	1
Train accuracy	98.63	99.84	99.95	99.93
Test accuracy	98.63	99.81	99.93	99.91

Table 4

The correct rate of the model is when adjusting the "colsample_bytree" parameter (50 times) for the simulation data sample

parameter	Colsample_bytree			
Value	0.3	0.5	0.7	1
Train accuracy	99.63	99.94	99.99	99.98
Test accuracy	99.63	99.91	99.99	99.97

According to the results of Table 1 to Table 4, we find that the result of parameter combination which is " max_depth = 5,min_child_weight = 1, subsample = 0.7,

colsample = 0.7" are the best combination parameter of fault diagnosis using XGBoost classifier, so we take this parameters combination as the optimal parameter of XGBoost classifier, and then divided all data samples are into 70% as the training set to training the fault diagnosis model based on the optimized XGBoost classifier, and 30% as the test set testing the accuracy of the trained model for fault diagnosis based on the optimized XGBoost classifier. In order to make the results more stable and reliable, the fault diagnosis model has trained and tested for 50 independent times, and to verify the excellent performance based on XGBoost, ELM and RF were used as comparisons in this experiment. ELM and RF used the same 50 independent times trained and tested as the same as the method of XGBoost, and the average results are shown in Table 5.

Table 5

Comparison of XGBoost, ELM and RF for the simulation data sample (50 times)

	XGBoost		ELM		RF	
Item	Training accuracy (%)	Testing accuracy (%)	Training accuracy (%)	Testing accuracy (%)	Training accuracy (%)	Testing accuracy (%)
O	100	100	100	100	100	100
LL1	100	100	100	100	100	100
LL2	99.99	99.99	99.90	99.99	99.98	99.98
PS	99.99	99,99	99.93	99.92	99.92	99.91
DS	99.99	99.99	99.92	99.96	99.94	99.95
DA	100	100	99.98	99.95	100	100
N	100	100	100	100	100	100

2.5 Experimental study

The 2kw small grid-connected photovoltaic system of the North Building of the College of Physics and Information Engineering of Fuzhou University was used to verify the photovoltaic fault diagnosis model proposed by this paper. Simulate the same fault as section 2.4 and collection data samples at on a clear day, data of each the fault condition is collected for 3 hours, and a data is collected every one second. The PV array's I_{mmp}, V_{mmp}, PV array temperature T_a and environmental irradiance G_a are collected. Each fault condition collects 10800 sets of data, and seven working conditions totals 75600 sets of data, and then extract fault features based on collected data. Training the PV fault diagnosis model based real PV array is the same as discussed and detailed in section 2.4. Because the experimental data samples are noisy and the simulation data samples are ideal so the trained accuracy and the tested accuracy of the experimental data is slightly less than the simulation data. However, the experimental accuracy is also relatively high so the PV fault diagnosis model is effective. Using the same method as section 2.4 and then come to result in Table 6 to Table 9 and Table 10. The results compared with ELM and RF show that the trained accuracy and the tested accuracy of the PV array fault diagnosis model based XGBoost are better than ELM and RF.

Table 6

The correct rate of the model is when adjusting the "max_depth" parameter (50 times) the experimental data sample

parameter	Max_depth				
Value	1	3	5	7	9
Train accuracy	98.85	99.88	99.90	99.80	99.75
Test accuracy	98.84	99.89	99.92	99.81	99.73

Table 7

The correct rate of the model is when adjusting the "min_child_weight" parameter (50 times) the experimental data sample

parameter	Min_child_weight				
Value	0.5	1	3	5	7
Train accuracy	98.78	99.83	99.81	99.79	99.78
Test accuracy	98.75	99.82	99.80	99.80	99.79

Table 8

The correct rate of the model is when adjusting the "subsample" parameter (50 times) the experimental data sample

parameter	Subsample			
Value	0.3	0.5	0.7	1
Train accuracy	98.63	99.84	99.95	99.93
Test accuracy	98.63	99.81	99.91	99.91

Table 9

The correct rate of the model is when adjusting the "colsample_bytree" parameter (50 times) the experimental data sample

parameter	Colsample_bytree			
Value	0.3	0.5	0.7	1
Train accuracy	99.90	99.91	99.93	99.92
Test accuracy	99.89	99.91	99.94	99.91

Table 10

Comparison of XGBoost, ELM and RF the experimental data sample (50 times)

	XGBoost		ELM		RF	
Item	Training accuracy (%)	Testing accuracy (%)	Training accuracy (%)	Testing accuracy (%)	Training accuracy (%)	Testing accuracy (%)
O	100	100	99.96	99.98	100	100
LL1	99.94	99.96	99.89	99.78	99.93	99.95
LL2	99.92	99.98	97.53	98.80	99.90	99.91
PS	99.94	99.92	99.93	99.89	99.92	99.91
DS	99.96	99.92	97.54	96.90	99.94	99.90
DA	99.98	99.98	99.99	99.98	99.98	99.98
N	99.97	99.98	99.36	99.42	99.87	99.85

*C*ONCLUSIONS

In this paper, a new online fault diagnosis method based on XGBoost classifier is proposed for PV arrays. From the monitored voltage, current, temperature and irradiance of the PV array, a seven dimensional fault features is proposed as the XGBoost classifier based fault diagnosis model. The proposed model can Identify some faults commonly occurring in PV arrays, including open circuit faults, line faults, local shadows, and degradation faults. In order to obtain a stable and optimal model, K-folding is used to optimize the parameters of the PV fault diagnosis model. The faults are simulated both in Simulink and laboratory PV power plants, from which a large number of data samples are collected to train the model and test the performance. The ELM and RF algorithm are also tested for comparison with the XGBoost based model. The comparison results demonstrate that the accuracy of the ensemble learning algorithms such as XGBoost and RF is significantly higher than that of the ELM single algorithm, and XGBoost is better than RF due to the addition of regular terms in the objective function. The accuracy of fault diagnosis model of Simulink simulation is as high as 99.99%, while that of real experiment is as high as 99.90%. In the subsequent work, feature reduction will be further introduced to reduce the computational complexity of the model to obtain a better PV fault diagnosis model.

ACKNOWLEDGEMENT

The authors would like to acknowledge the financial supports by the National Natural Science Foundation of China (Grant Nos. 61601127, 51508105 and 61574038), the Fujian Provincial Department of Science and Technology of China (Grant Nos. 2016H6012 and 2018J01774), and the Fujian Provincial Department of Education of China (Grant No. JAT160073).

REFERENCE

[1] M. K. Alam, F. H. Khan, J. Johnson, and J. Flicker, "PV faults: Overview, modeling, prevention and detection techniques," in *Control and Modeling for Power Electronics (COMPEL), 2013 IEEE 14th Workshop on*, 2013, pp. 1-7: IEEE.

[2] Z. A. Jaffery, A. K. Dubey, Irshad, and A. Haque, "Scheme for predictive fault diagnosis in photo-voltaic modules using thermal imaging," *Infrared Physics & Technology,* vol. 83, pp. 182-187, 2017.

[3] Z. Chen *et al.*, "Random forest based intelligent fault diagnosis for PV arrays using array voltage and string currents," *Energy Conversion and Management,* vol. 178, pp. 250-264, 2018.

[4] Z. Yi and A. Etemadi, "Fault Detection for Photovoltaic Systems Based on Multi-resolution Signal Decomposition and Fuzzy Inference Systems," *IEEE Transactions on Smart Grid,* vol. PP, no. 99, pp. 1-1, 2017.

[5] Z. Chen, L. Wu, S. Cheng, P. Lin, Y. Wu, and W. Lin, "Intelligent fault diagnosis of photovoltaic arrays based on optimized kernel extreme learning machine and IV characteristics," *Applied Energy,* 2017.

[6] Z. Yi and A. H. Etemadi, "A novel detection algorithm for Line-to-Line faults in Photovoltaic (PV) arrays based on support vector machine (SVM)," in *Power and Energy Society General Meeting (PESGM), 2016*, 2016, pp. 1-4: IEEE.

[7] H. Zhu, L. Lu, J. Yao, S. Dai, and Y. Hu, "Fault diagnosis approach for photovoltaic arrays based on unsupervised sample clustering and probabilistic neural network model," *Solar Energy,* vol. 176, pp. 395-405, 2018.

[8] R. Benkercha and S. Moulahoum, "Fault detection and diagnosis based on C4.5 decision tree algorithm for grid connected PV system," *Solar Energy,* vol. 173, pp. 610-634, 2018.

[9] T. Chen and C. Guestrin, "XGBoost," pp. 785-794, 2016.

[10] M. Barukčić, V. Ćorluka, and K. Miklošević, "The irradiance and temperature dependent mathematical model for estimation of photovoltaic panel performances," *Energy Conversion & Management,* vol. 101, no. C, pp. 229-238, 2015.

[11] D. L. King, J. A. Kratochvil, and W. E. Boyson, "Temperature coefficients for PV modules and arrays: measurement methods, difficulties, and results," in *Photovoltaic Specialists Conference, 1997., Conference Record of the Twenty-Sixth IEEE*, 1997, pp. 1183-1186: IEEE.