Fault Detection for Photovoltaic Systems Using Multivariate Analysis With Electrical and Environmental Variables

Gyu Gwang Kim, Wonbin Lee, Byeong Gwan Bhang, Jin Ho Choi, and Hyung-Keun Ahn, *Member, IEEE*

***Abstract*—Fault detection and repair of the components of photovoltaic (PV) systems are essential to avoid economic losses and facility accidents, thereby ensuring reliable and safe systems. This article presents a method to detect faults in a PV system based on power ratio (PR), voltage ratio (VR), and current ratio (IR). The lower control limit (LCL) and upper control limit (UCL) of each ratio were defined using the data of a test site system under normal operating conditions. If PR exceeded the set range, the algorithm considered a fault. Subsequently, PR and IR were examined via the algorithm to diagnose faults in the system as series, parallel, or total faults. The results showed that PR exceeded the designated range between LCL (0.93) and UCL (1.02) by dropping to 0.91–0.68, 0.88–0.62, and 0.66–0.33 for series, total, and parallel faults, respectively. Moreover, VR exceeded the LCL (0.99) and UCL (1.01) by 0.95–0.69 and 0.91–0.62 for series and total faults, respectively, but not under parallel faults condition. IR did not change in series and total faults but exceeded the range of LCL (0.93) and UCL (1.05) by dropping to 0.66–0.33. Thus, faults in PV systems can be detected and diagnosed by analyzing quantitative output values.**

***Index Terms*—Correlation coefficient, current ratio, electrical variables, fault detection algorithm, multivariate analysis, power ratio, regression analysis, voltage ratio, weather variables.**

I. Introduction

THE significant developments in photovoltaic (PV) systems in China during the past three years have contributed to a global increase in the number of PV systems. In 2017, the top ten countries had an installed PV capacity of 344.5 GW [1]. However, owing to this increasing use of PV systems, the importance of system reliability and safety is also expanding [2]. Moreover, additional time and costs are incurred when failures or faults in the system are failed to be detected in time. Therefore, to maintain a high-quality system for extended periods, it is essential to calculate and analyze quantitative values to promptly identify the locations and times of faults and failures [3].

Furthermore, PV generation is influenced by various environmental factors, i.e., irradiance, ambient temperature (T_a), relative humidity (RH), and wind speed (WS) [4]. The relationship between these factors and the output of the PV system can be expressed using a correlation coefficient. When the correlation coefficient is close to the absolute value of 1, it indicates a strong relationship between the variables, whereas a coefficient value close to 0 indicates a weak relationship between the two variables.

The maximum power output of the PV system can be achieved when incident irradiance enters the surface at a right angle [5]. Changes in surrounding environmental conditions, such as shading, are also among the main causes of power drops in PV systems [6]. When a PV system uses a bifacial PV module, the shading conditions of the rear side also need to be considered to avoid undesired power drops [7], [8]. When the global horizontal irradiance (GHI) and the plane of array (POA) irradiance were compared, POA showed a stronger correlation with the output power of the PV module than GHI, as shown in our previous work [9].

A PV system is greatly influenced by changes in the temperature. As the temperature of the PV module rises, the output power decreases. Generally, when the temperature of a crystalline silicon PV module increases by 1, the output power of the module decreases by 0.35%–0.4% [10]–[12]. In a floating PV system, the operating temperature of the PV module is lower than that of a land-based system; therefore, a floating PV system can generate 10% more energy [13]. The correlation coefficient between the output power of the PV module and temperature is 0.71 for the module temperature and 0.13 for the ambient temperature. According to theoretical analyses, the coefficient between the temperature and output power should be negative; however, according to a correlation analysis, the temperature rises as irradiance rises, resulting in positive correlation coefficients, also shown in our previous work [9].

Similarly, the RH affects PV systems in a similar fashion as dust accumulation affects the output power of the PV system. Moreover, water vapor particles in the air reduce the amount of irradiance, and light rays hitting these water droplets are scattered via refraction, reflection, or diffraction [14]. The correlation coefficient between the output power and RH is −0.46, which indicates that when irradiance at the installation site

Manuscript received July 13, 2020; revised August 7, 2020 and September 10, 2020; accepted October 16, 2020. Date of publication November 4, 2020; date of current version December 21, 2020. This work was supported in part by the New and Renewable Energy Technology Program of the Korea Institute of Energy, Technology, Evaluation, and Planning, and in part by the Ministry of Trade, Industry, and Energy, Republic of Korea, under Grant 20183010014260. *(Gyu Gwang Kim and Wonbin Lee are co-first authors.) (Corresponding author: Hyung-Keun Ahn.)*

The authors are with the Next Generation Photovoltaic Module and Power System Research Center, Konkuk University, Seoul 05029, South Korea (e-mail: rbrhkd00@konkuk.ac.kr; dnjsqls6766@konkuk.ac.kr; shorev@konkuk.ac.kr; bbk0627@konkuk.ac.kr; hkahn@konkuk.ac.kr).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JPHOTOV.2020.3032974

increases, the RH also decreases, resulting in increased output power of the system [15].

Finally, the correlation coefficient between the WS and the output power of the PV system is 0.19, which indicates a weak relationship. Nevertheless, it is still a positive value because wind cools the surface of the PV module, leading to an increase in the output power of the PV system [16].

In this study, a half-cut 36-cell PV module that produces one-sixth of the value of a 72-cell crystalline silicon PV module was used to design a PV system. The data of irradiance, module temperature, ambient temperature, RH, and WS from a test site were collected to model and estimate the output power, voltage, and current in the system. Calculations were performed to detect and diagnose faults in the PV system using the proposed algorithm based on output data under normal operating conditions. This algorithm can be particularly essential for ground, floating, and/or marine-based PV systems, where installation conditions are sensitive to sudden climatic changes.

II. Fault Detection and Diagnosis of PV System

A. Performance Ratio (PR)

The PR is a general evaluation method for a PV generation system and calculates the ratio between the theoretical output and actual output as [17]

$$\mathrm{PR} = Y_p/Y_r. \tag{1}$$

The numerator and denominator of the right-hand-side term of (1) each represent the power rate of the PV system and the theoretical yield when incident irradiation hits the PV system, respectively. These are presented as follows:

$$Y_p = E_{P,d}/P_{A,s} \tag{2}$$

$$Y_r = H_{A,d}/G_{\mathrm{STC}}. \tag{3}$$

In (1), Y_p represents the yield of the PV system, which is the rate of generated power with units in $[\mathrm{kWh}\cdot d^{-1}\cdot(\mathrm{kW})^{-1}]$ or $[\mathrm{h}\cdot d^{-1}]$. $E_{p,d}$ is the total cumulative energy generated by the PV generation system, and its unit is [kWh/d]. $P_{A,s}$ is the rated output of the PV system, and its unit is kilowatt [kW]. All variables capable of affecting the output of the PV system, such as the power loss, from the increasing resistance owing to multiple connections of modules in PV arrays or changes in output owing to changes in operating temperature are considered in the output value that is processed through the inverter, which is the final stage, and are included in Y_p.

The reference yield is expressed as Y_r, and its unit is $[\mathrm{kWh}\cdot d^{-1}\cdot(\mathrm{kW})^{-1}]$ or $[\mathrm{h}\cdot d^{-1}]$. $H_{A,d}$ is the total cumulative solar irradiation for a day, and its unit is [kWh/m^2]. G_{STC} is a day's worth of solar irradiation according to standard testing conditions (STC), and its unit is [d $\times$ 1 kW/m^2]. Thus, Y_r is the ratio of the actual cumulative solar irradiance to the STC solar irradiance. Since a PV system is installed outdoors, it is affected by various factors, and the temperature is one of the factors with the greatest impact. When the temperature rises by 1 °C, the power decreases by a certain percentage, which is known as the temperature coefficient. If the power ratio is calculated without considering the temperature coefficient, errors may arise [18].

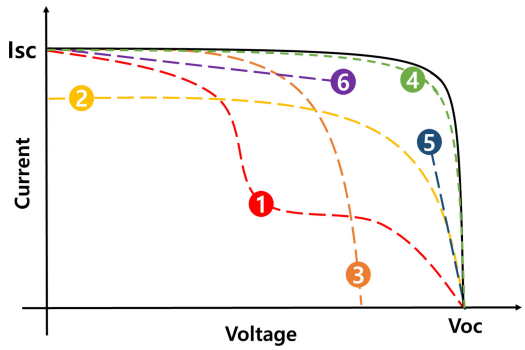

Fig. 1. *I–V* curves under different PV module fault conditions.

B. I–V Curve

The *I–V* curve of a PV module is represented by a graph of current and voltage. When the voltage of the PV module is 0, a short-circuit current (I_{sc}) is generated, owing to the generation and collection of photogenerated carriers. In an ideal case where the power loss from the resistance is same as the ideal calculation, I_{sc} and the light-induced current are identical. I_{sc} is the maximum current available from the PV module, and the open-circuit voltage (V_{oc}) is the maximum voltage available from the PV module. The fill factor (FF) is an element that determines the maximum power. The *I–V* curve of a PV system can be categorized into six types according to each fault condition. The *I–V* curve of PV module under the normal condition is calculated and visualized via a one-diode equation. The detection and diagnosis of faults in the PV module could be done by detecting different changes in *I–V* curve according to different types of faults. One method of using *I–V* curve to detect fault was studied in [19] by applying neurofuzzy fault detection method. Another method of using *I–V* curve to detect faults in the PV module was studied by Tingting and Xiaohong [20] where they used MATLAB and Simulink to make a hybrid model with variables, such as the number of PV modules, V_{oc}, and I_{sc} at maximum power point condition.

1) Mismatch: PV modules in a PV array are connected in either series or parallel. When a string, which refers to several PV modules connected in series, is connected to other strings in series and parallel, multiple dysfunctions can take place. One possible dysfunction is that the current produced from one of the strings could be lower than that of the other strings. This situation can occur when PV modules connected in strings are not capable of generating electricity in the normal way because the PV module has been somehow polluted or damaged. When such mismatching has occurred, the *I–V* curve of the system has the shape of curve ① in Fig. 1.

2) Uniform Soiling: The mismatching of modules in strings or arrays usually occurs because of differences in generated electricity. Curve ② of Fig. 1 is a case where the front surface of a PV system has been generally contaminated or plenty of time has passed since installation, resulting in decreasing generation.

3) Bypass Diode: When the surface of the PV module is soiled or shaded, PV cells in that particular region are unable to generate the same amount of electricity as normal cells. In such a fault condition, the voltage generated from other normal

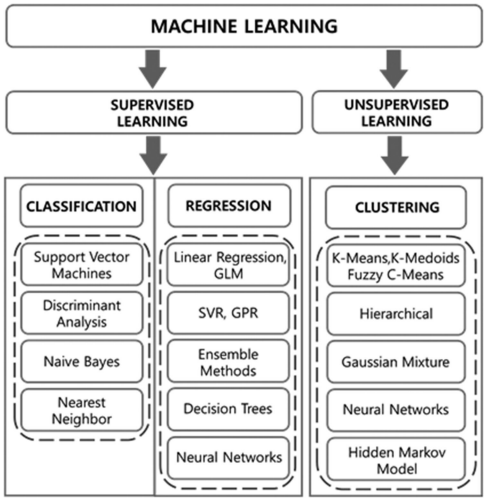

Fig. 2. General category of machine learning method [24].

cells would be cast to the defected cells, resulting in hotspots or other malfunctions. A bypass diode is installed in the PV module to minimize the damage caused in such a situation. When the bypass diode is functioning, currents generated from normal cells flow through the bypass diode instead of the defective cells. When the bypass diode is in operation, the open-circuit voltage decreases, resulting in an *I–V* curve in the form of curves ① and ③ in Fig. 1.

4) Unknown Knee: Curve ④ of Fig. 1 is an *I–V* curve of an unknown knee, which is the case where degradation occurs with no clear reason. Such a case requires significant time and general inspections throughout the entire PV system.

5) Series Resistance: There are various sources of series resistance in PV cells and modules. The contact resistance between the metal contact and silicon of a PV cell or mismatch between the PV modules can increase the series resistance. A major influence on the series resistance of a PV system is the reduction of the FF. However, if the series resistance is high enough, it could also result in a reduction of the short-circuit current. The series resistance does not affect the PV cell at the open-circuit voltage because the overall current flows through the PV cell, resulting in the series resistance of 0. By contrast, around the open-circuit voltage, the series resistance greatly affects the *I–V* curve. The *I–V* curve of a PV module as the series resistance increases could be drawn in a similar form as curve ⑤ in Fig. 1. The series resistance can be approximated as the slope of the *I–V* curve at the open-circuit voltage point.

6) Shunt Resistance: A decline in power generation caused by a reduction in the shunt resistance generally occurs during manufacturing. If the shunt resistance is too low, the light-induced current could flow through undesignated paths, resulting in a power reduction. When such an unwanted bypass is formed, the current generated from the PV module and open-circuit voltage decreases. A fault caused by the shunt resistance is even more drastic at low-irradiation conditions because the light-induced current is lesser in this situation. The *I–V* curve of the PV module as the shunt resistance decreases could be drawn in a similar form as curve ⑥ in Fig. 1. The parallel resistance can be approximated as the slope of the *I–V* curve at the short-circuit current point.

7) Design Factor: The design factor is an analysis method that uses the operation time from the time of incident irradiation at the surface of the PV system to the time of the ac output generation of the inverter.

This method is similar to the PR method except that the design factor is capable of detecting faults in each part of the PV system through the yield. This is represented by quantified values derived from analyzing each section of the PV system. The reference yield of solar irradiation (Y_r), maximum PV array yield (Y_{am}), temperature-corrected PV array yield (Y_{at}), optimized PV array yield (Y_{ao}), PV array yield (Y_a), and PV system yield (Y_f) are defined as follows:

$$Y_r G_{a,\mathrm{meas.}}/G_{a,\mathrm{ref}} \tag{4}$$

$$Y_{am} a_m \cdot Y_r \tag{5}$$

$$Y_{ao} Y_{am} - l_a/P_{as} \tag{6}$$

$$Y_{at} Y_a/\left[1 + a_t \times (T_m - 25)\right] \tag{7}$$

$$Y_a P_{a,\mathrm{meas.}}/P_{as} \tag{8}$$

$$Y_f P_{f,\mathrm{meas.}}/P_{as} \tag{9}$$

where $G_{a,\mathrm{meas}}$ irradiation at a degree of POA ;
$G_{a,\mathrm{ref}}$ 1 [kW/m^2];
a_m maximum efficiency coefficient of PV array;
l_a dc cable loss of PV array;
P_{as} total capacity of PV array;
a_t temperature coefficient of PV array;
T_m module temperature;
$P_{a,\mathrm{meas}}$ output power of PV array [kW];
$P_{f,\mathrm{meas}}$ output power of power conditioning system.

The general loss rate of a PV array is calculated using the equations presented above. The loss rate caused by a transition of irradiation, contamination of the surface, aging of PV modules, and other factors is calculated by comparing Y_{ao} with Y_{at} [21], [22].

C. Exponentially Weighted Moving Average (EWMA)

EWMA is an effective method to detect small changes in the process mean with no concern for the size of the measured value. The moving average refers to a method of attaining more weight to new data, whereas the past data are given less weight exponentially, and the total sum of the weight is 1. The EWMA is defined as

$$S_i r Z_i + (1-r)\, S_{i-1} = \sum_{j=1}^{i-1} r(1-r)^j Z_{i-j} + (1-r)^i S_0, i = 1, 2, \ldots. \tag{10}$$

In (10), S_0 is the initial value, which is set to be the mean of controlled data μ_{Z0}, and r is the weight with a range between 0 and 1. The present and past statistical values were used to express (10), where r is exponentially weighted. When $\theta = \theta_0$,

the mean of S_i is derived as follows:

$$E_{\theta_0}\sum_{j=1}^{i-1} r(1-r)^j\mu_{Z_0} + (1-r)^i\mu_{Z_0} = \mu_{Z_0} \tag{11}$$

$$\begin{aligned}\mathrm{Var}_{\theta_0}(S_i) &= \sum_{j=1}^{i-1} r^2(1-r)^{2j}\,\mathrm{Var}(Z_{i-j}) \\ &= \sigma_{Z_0}^2\left(\frac{r}{2-r}\right)\left[1-(1-r)^{2i}\right].\end{aligned} \tag{12}$$

$[1-(1-r)^{2i}]$ in (12) sharply converges to 1 as *i* increases. The variance of S_i is derived as

$$\lim_{i\to\infty}\mathrm{Var}_{\theta_0}(S_i) = \sigma_{Z_0}^2\left(\frac{r}{2-r}\right). \tag{13}$$

As a result, the upper control limit (UCL) and lower control limit (LCL) are defined as follows:

$$\begin{pmatrix}\mathrm{UCL}\\ \mathrm{LCL}\end{pmatrix} = \begin{pmatrix}\mu_{Z_0} + k\sigma_{Z_0}\left(\frac{r}{2-r}\right)\\ \mu_{Z_0} - k\sigma_{Z_0}\left(\frac{r}{2-r}\right)\end{pmatrix}. \tag{14}$$

By determining the upper and lower limits of the control range, abnormalities can easily be detected. However, since the control range greatly depends on the weight, determining the proper weight is crucial [23].

D. Machine Learning Methods

The machine learning method can be classified mainly into two categories: supervised learning and unsupervised learning. Supervised learning trains model with a dataset where input and output are already known so that the future output could be predicted, whereas the unsupervised learning method focuses on finding a hidden pattern or intrinsic structure from input data.

The supervised learning method could further be divided into two groups: classification and regression. Classification predicts a discrete response, whereas regression predicts a continuous response. Chen *et al*. [25] applied a random forest ensemble learning algorithm to detect and diagnose PV array faults.

Clustering, which in the most common method of unsupervised learning technique, is used to find hidden patterns in a dataset by exploratory data analysis. Chen *et al*. [26] merged one-diode model and MATLAB Simulink to analyze and detect faults in PV array. Park and Ahn [27] proposed a method to detect and diagnose faults in the PV system with an RNN-based deep-learning algorithm.

As shown in Fig. 3, various causes of faults in the PV system ultimately cause the output power of the PV system to decrease. The term "fault" in this study refers to a condition where PV generation does not meet the user's (owner or administrator, etc., of PV system) expectation or does not generate electricity at all. In this study, change in current, voltage, and power of the PV system under a fault condition is studied to detect and categorize the faults [28].

Generally, when designing the output model of PV, the one-diode model is the most used method. Since this method is the most general one that corresponds to most cases at a certain degree, this method cannot be the most optimized model for each PV system's installation conditions. The method proposed in this study merges the output data with environmental data of the targeted PV system. The user or operator of the system would be able to derive the best fit model for a particular system to analyze and detect possible faults. A simple but effective fault detection algorithm using the proposed ratio of current, voltage, and power is presented in Sections III–VI.

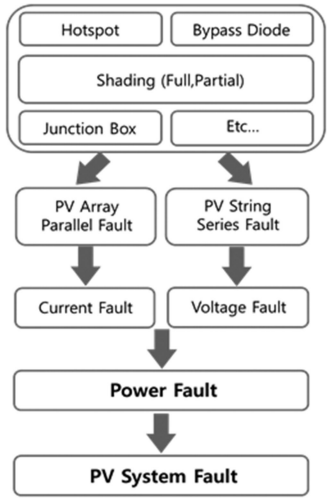

Fig. 3. Flow of power decrease in the PV system from various faults.

III. Mathematical Models of PV System

A. Correlation Analysis Between the PV System and Environmental Variables

Correlation analysis is a method of elucidating the interrelationship between two variables. Generally, correlation analysis can be divided into two main groups: Spearman and Pearson. In a Spearman analysis, the order of strength in the interrelationship between the two variables is analyzed [29]. However, in a Pearson analysis, the correlation between the variables is analyzed based on the linear relationship between the two variables and expressed as a number between -1 and $+1$. When the two variables are in a proportional relationship, the coefficients are positive, where a number closer to 1 indicates a stronger relationship. By contrast, when two variables are in an inverse proportional relationship, the coefficients are negative [30].

In this study, a Pearson correlation analysis was chosen to analyze the interrelationship between the output of the PV system (voltage, current, and power at the maximum power point) and the environmental variables (irradiance, module temperature T_m, T_a, WS, and RH) [31].

The Pearson correlation coefficient *r* between two variables x_i and y_i is derived as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x}_i)(y_i-\bar{y}_i)}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x}_i)^2\sum_{i=1}^{n}(y_i-\bar{y}_i)^2}} \tag{15}$$

$$\bar{x}_i = \frac{1}{N}\sum_{i=1}^{N}x_i,\quad \bar{y}_i = \frac{1}{N}\sum_{i=1}^{N}y_i \tag{16}$$

where $\bar{x}_i$ and $\bar{y}_i$ are the average values of x_i and y_i, respectively, and n is the size of the samples of each variable. In Figs. 1, 3, and 5, the numbers in the upper half refer to the correlation coefficients between the variables, and the red lines in the lower half indicate the linearity between the variables [32]. The data used for the Pearson correlation analysis were collected from the test site, which is discussed in detail in Section IV.

B. PV Output Prediction Model

In this study, environmental variables to be used to design PV prediction models were prioritized by evaluating the correlation between the PV generation and environmental variables. Prediction models for voltage, current, and power were derived through a correlation analysis.

Coefficients in this study are estimated by the least-square regression method, which is described in (17), along with the multiregressive model

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \beta_3 x_{31} + \ldots + e_1 \\ y_2 &= \beta_0 + \beta_1 x_{12} + \beta_2 x_{22} + \beta_3 x_{32} + \ldots + e_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \beta_3 x_{3n} + \ldots + e_n. \end{aligned} \tag{17}$$

The relationship among the dependent variables (y_i), independent variables (x_i), and the regression coefficient (β_n) is expressed by

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + e_i \tag{18}$$

where e_i is the error term, and its mean value must be minimized to enhance the accuracy of the correlation analysis. β that makes the mean value of e_i to be 0 is derived by the least-square regression method and is described as follows:

$$\begin{aligned} L &= \min\left[\sum_{i=1}^{n} {e_i}^2\right] = \min\left[\sum_{i=1}^{n} (y_i - \widehat{y_i})^2\right] \\ &= \left(y - X\hat{\beta}\right)' \left(y - X\hat{\beta}\right) \\ &= Y'\,Y - Y'X\hat{\beta} - \beta' X'Y + \hat{\beta}' X'X\hat{\beta} \end{aligned} \tag{19}$$

$$\frac{\partial L}{\partial \hat{\beta}} = -2X'Y' + 2X'X\,\hat{\beta} = 0 \tag{20}$$

$$\hat{\beta} = (X'X)^{-1}\,X'Y. \tag{21}$$

L is defined as the total sum of squared e_i in (19). Then, L is partially differentiated by β in (20), and its value is set to 0. β that fits the term is calculated in (21) [33].

The mean absolute percentage error (MAPE) and the root-mean-square error (RMSE) were calculated for each model to evaluate their accuracy. The calculation methods of MAPE and RMSE are presented in (22) and (23). MAPE is generally expressed as a percentage. A demerit of this method is a lack of an upper limit for the error rate when the outcome of the prediction is too great. The result of RMSE corresponds to the square of the error; therefore, it is greatly affected by outliers.

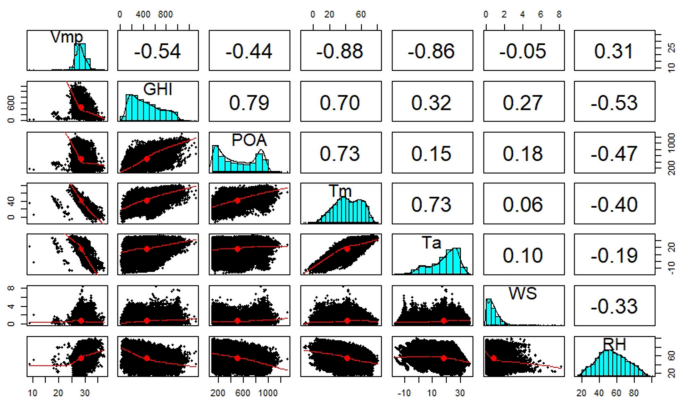

Fig. 4. Pairwise scatter plots and correlation coefficients between V_{mp} and environmental variables.

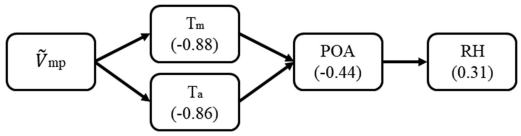

Fig. 5. Process flow of selecting variables for maximum voltage prediction model through correlation analysis and scatter plot.

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|{(y_i - \widehat{y_i})}/{y_i}\right| \times 100\% \tag{22}$$

$$\text{RMSE} = 100\% \times \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}\Big/\frac{1}{n}\sum_{i=1}^{n} y_i. \tag{23}$$

1) Maximum Power Point Voltage: Fig. 4 depicts a scatter plot of the interrelationship between the maximum power point voltage V_{mp} and other environmental variables. It is evident that V_{mp} decreases as the temperature rises in the actual environment, which is in agreement with the theoretical analysis. Moreover, V_{mp} exhibits a tendency to decrease as the extent of solar irradiance increases, which contradicts the theoretical analysis of the PV cell. However, this outcome can be attributed to the high correlation between the temperature and irradiance.

High irradiance leads to high temperatures, resulting in a decrease in V_{mp}. There is a positive correlation between V_{mp} and humidity. This is because when RH is high, it indicates a cloudy day, which limits irradiation and causes low temperatures; hence, V_{mp} increases. This analysis was done with the data collected from a test site in South Korea. If the same analysis is carried out with data from the PV system in different environmental conditions, slightly different correlation coefficients are attained. Fig. 5 depicts the process flow of selecting the variables used in the V_{mp} prediction model.

The prediction models of V_{mp} were designed following the process flow diagram, as shown in Fig. 4. The models and the error rate of each model are listed in Table I. The models of the current and power of the PV system are presented as variables multiplied by POA since the current and power are proportional

TABLE I
PREDICTION MODELS OF V_{mp}, I_{mp}, AND P_{mp}

Model	Equations	MAPE (%)	RMSE (%)
$\tilde{V}$mp	$C_0 + \ln(POA)(C_1 + C_2 T_m)$	0.81	2.26
	$C_0 + \ln(POA)(C_1 + C_2 T_a)$	1.64	4.04
	$C_0 + \ln(POA)(C_1 + C_2 T_m + C_3 \ln(POA)RH)$	0.81	2.26
	$C_0 + \ln(POA)(C_1 + C_2 T_a + C_3 \ln(POA)RH)$	1.64	4.04
$\tilde{I}$mp	$POA(C_0 + C_1 T_m)$	2.36	0.91
	$POA[C_0 + POA(C_1 T_a + C_2 WS)]$	2.30	0.09
	$POA[C_0 + C_1 T_m + POA(C_2 WS + C_3 RH)]$	2.22	0.09
	$POA[C_0 + POA(C_1 T_a + C_2 WS + C_3 RH)]$	2.17	0.08
$\tilde{P}$mp	$POA(C_0 + C_1 T_m)$	2.72	20.69
	$POA[C_0 + POA(C_1 T_a + C_2 WS)]$	2.74	22.81
	$POA[C_0 + C_1 T_m + POA(C_2 WS + C_3 RH)]$	2.75	22.67
	$POA[C_0 + POA(C_1 T_a + C_2 WS + C_3 RH)]$	2.65	20.01

to POA. On the other hand, voltage is proportional to POA up to a certain point where it saturates. Hence, the prediction models of voltage take the form of a natural logarithm function [9].

Other environmental variables except for POA were deployed in the model either as a product of POA or a natural logarithm of POA when the correlation coefficient is below 0.5.

It could be seen that the error in the model utilizing only T_a and POA data is greater than that in the model utilizing only T_m and POA data. This is because a rise in module temperature leads to a voltage drop in the PV module. In this study, the model used to predict V_{mp} employs POA, T_a, and RH (the fourth model in Table I). The error rate of the prediction model using T_m is smaller than the model using T_a; however, from our previous study [9] that was done with a larger data group proved that the model using T_a, WS, RH, and POA shows smaller error compared with the model using T_m, WS, RH, and POA. From the model that does not utilize the RH and the model that exhibits an identical error rate, the latter was chosen in this study because previous studies reported that this model yields a smaller error rate on cloudy days [9], [34].

2) Maximum Power Point Current: Fig. 6 depicts a scatter plot of the interrelationship between the maximum power point current (I_{mp}) and other environmental variables. It could be seen that I_{mp} increases as the amount of solar irradiation increases, and it shows a high correlation coefficient POA, which is the irradiation at an incident angle perpendicular to the plane of the module. I_{mp} increases according to the temperature. This

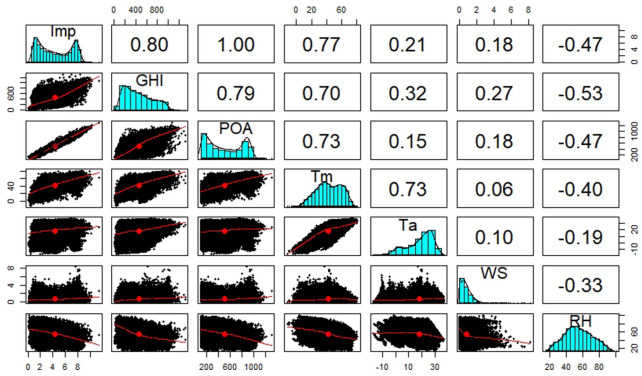

Fig. 6. Pairwise scatter plots and correlation coefficients between I_{mp} and environmental variables.

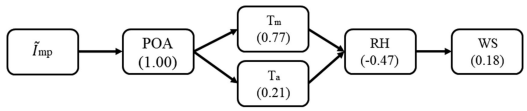

Fig. 7. Process flow of selecting variables for I_{mp} prediction model through correlation analysis and scatter plot.

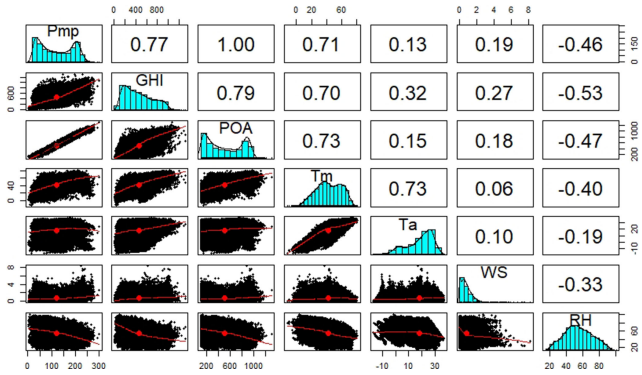

Fig. 8. Pairwise scatter plots and correlation coefficients between P_{mp} and environmental variables [33].

outcome can be attributed to the high correlation between the temperature and irradiance.

I_{mp} indicates a negative correlation with humidity, and it could be seen that I_{mp} decreases owing to a decrease in the amount of solar irradiation, which is caused by increasing humidity. Fig. 7 presents the flow of selecting the variables to be used for the I_{mp} prediction model. Using this flow diagram, I_{mp} is expressed as a mathematical model in terms of the correlation coefficient, as listed in Table I. When predicting I_{mp}, it was evident that the model utilizing T_a, WS, and RH yields fewer errors as compared with the model that utilizes module temperature data. Therefore, the fourth I_{mp} prediction equation in Table I was used to predict I_{mp} in this study.

3) Maximum Power Point Power: Fig. 8 depicts a scatter plot of the interrelationship between the maximum power point power (P_{mp}) and other variables. It is evident that P_{mp} increases as the amount of solar irradiation increases. Similar to I_{mp}, P_{mp} exhibits a high correlation to POA. P_{mp} appears to increase as the temperature increases; however, this is a result of the positive

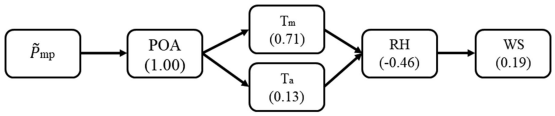

Fig. 9. Process flow of selecting variables for maximum power prediction model through correlation analysis and scatter plot.

correlation between the POA and the temperature. There is a positive correlation between P_{mp} and the temperature, which does not agree with the theoretical and practical analyses. However, this discordance is because there is a positive correlation between the POA and the temperature and between I_{mp} and P_{mp}.

Fig. 9 illustrates the flow of selecting the variables to be used in the P_{mp} prediction model. Using this flow diagram, the mathematical expression for the correlation coefficient for P_{mp} is listed in Table I.

Two methods to predict P_{mp} were considered: multiplying the prediction of V_{mp} and I_{mp} obtained from the models, as listed in Table I, and designing a new prediction model for P_{mp} using the cases, as presented in Table I. With an error rate of 2.29 MAPE (%) and 18.67 RMSE (%), the errors in the prediction obtained by multiplying the predicted voltage and current were lower than those of the cases, as listed in Table I. Therefore, in this study, the former prediction method was used to predict P_{mp}.

C. Voltage, Current, and Power Ratios

1) Voltage Ratio: To validate the fault detection and diagnosis algorithm, changes in voltage, current, and power ratio according to different configurations were analyzed.

To evaluate the operating conditions of the PV system, the measured value of V_{mp} was compared with the predicted value that was derived from the prediction model, as listed in Table I. The following equation defines the voltage ratio (V_R), where the numerator is the actual measured value of $V_{\text{meas.}}$ of the PV system, and the denominator is the predicted value of $\tilde{V}_{mp}$:

$$V_R = V_{\text{meas.}}/\tilde{V}mp \tag{24}$$

Using (24), the V_R for various cases of defective modules was calculated and is illustrated in Fig. 10. A comparison was made for four configurations: 6 (series) by 6 (parallel), 9 by 4, 12 by 3, and 18 by 2.

The faults were categorized as either faults that take place at PV modules that are connected in series to form string in a PV system or faults that take place at connecting part between the strings which form PV array, and they are further referred to as series faults and parallel faults, respectively. When the series fault occurs, the faulty module in the string gets short circuited and the bypass diode is then activated, resulting drop in the voltage of the array. In a series fault, the number of faulty modules in the entire system increased one at a time. When the parallel fault occurs, the faulty string gets open circuited, resulting decrease in the current of the array. In a parallel fault, the number of faulty modules increased in the unit of a string. For instance, in the case of a 6-by-6 configuration, the number of faulty modules increased by 6.

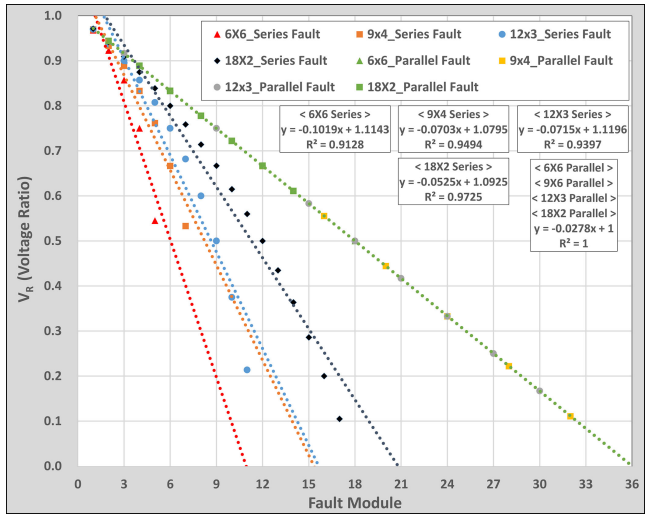

Fig. 10. Calculation results of V_R with different numbers of faulty PV modules in series and parallel connections.

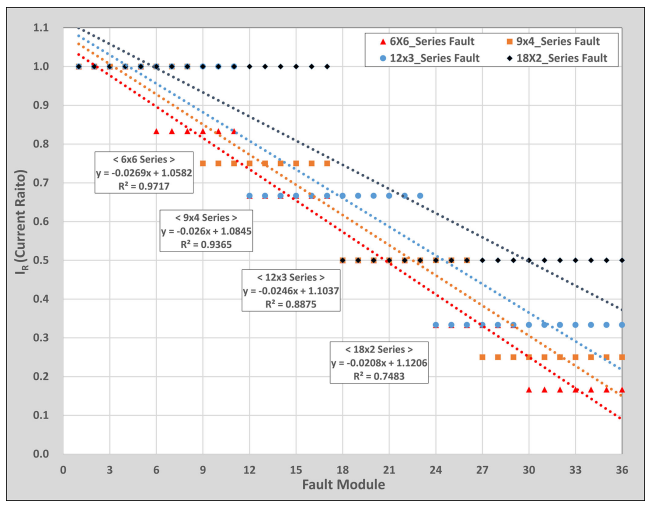

Fig. 11. Calculation results of I_R with different numbers of faulty PV modules.

The calculation results show that V_R dropped more sharply in a series fault than in a parallel fault. In series–parallel connections of PV modules, a series fault causes the total voltage of the entire system to drop considerably.

2) Current Ratio (I_R): The equation representing the current ratio is expressed in (25), and the prediction of I_R obtained by the equation is illustrated in Fig. 11

$$I_R = I_{\text{meas.}}/\tilde{I}mp. \tag{25}$$

The series and parallel structures of the PV system have the same characteristics of a regular parallel circuit: the voltage is a constant value and the current increases proportionally to the number of parallel connections. In a series connection of PV modules, the current is insensitive to the number of faulty modules in a string when the bypass diode is activated. However, when the entire string becomes faulty, the current of the PV array decreases, which is the definition of parallel fault. This is shown

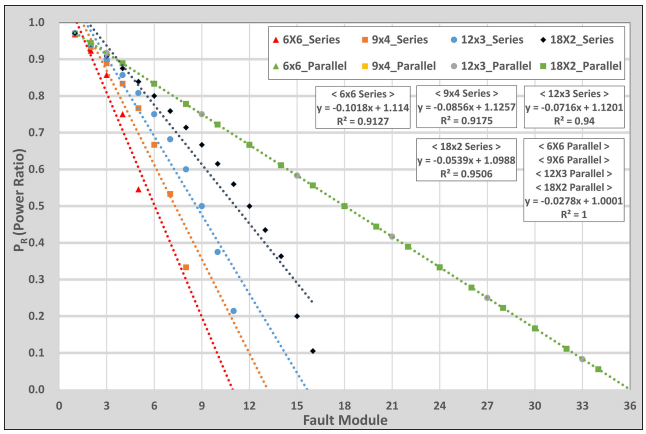

Fig. 12. Calculation results of P_R with numbers of faulty PV modules in series and parallel.

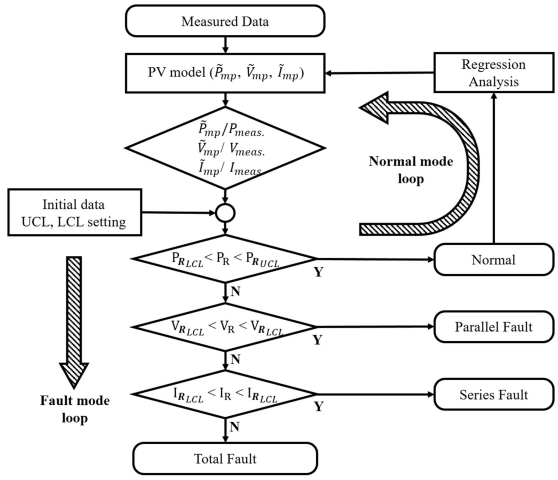

Fig. 13. Fault detection/diagnosis algorithm for the PV system.

in Fig. 11, as the current of each configuration decreases in a step-like manner.

3) Power Ratio (P_R): The power ratio measured at maximum power point is calculated by dividing the measured value of P_{mp} by the multiplied value of the voltage and current, which were predicted using the model, as presented in Table I. P_R is calculated using

$$P_R = P_{\text{meas.}} / \left(\tilde{V}mp \times \tilde{I}mp \right). \tag{26}$$

Fig. 12 illustrates the change in the calculated P_R as the number of faulty modules changes in the PV system, which consists of the same number of PV modules but in a different configuration of series and parallel connections. In a parallel fault, all three configurations showed a similar reduction rate of PR. Notably, the reduction rates of P_R and V_R under the parallel fault condition were the same. In a series fault, the reduction rate of P_R differed according to the number of PV modules connected in series.

IV. PV Fault Detection/Diagnosis Algorithm

A fault detection and diagnosis algorithm for PV systems was designed, as shown in Fig. 13. In the first stage of the algorithm, the measured values of the output data and environmental variables are fed to the PV system. Subsequently, P_R, V_R, and I_R are calculated using the equations presented in Section II. These values of P_R, V_R, and I_R are the standard values used to determine if faults have occurred in the PV system. Thereafter, the UCL and LCL can be defined using the initial data from the PV system under the normal operation condition.

In the fault detection stage, the P_R of the input data is evaluated. When the value exceeds the range between the LCL and UCL, it is identified as a fault, and it proceeds to the fault diagnosis stage. Otherwise, the input data are identified as a normal operation condition, and it proceeds to the normal-mode loop and is fed back to regression analysis for a better prediction of the PV output.

The next stage is the fault diagnosis stage. If the V_R of the input data lies within the range between V_{LCL} and V_{UCL}, it is identified as a parallel fault; otherwise, it proceeds to the next stage. The I_R of the input data is examined to check whether it lies within the range of I_{LCL} and I_{UCL}. If I_R is within this designated range, then it is identified as a series fault. Otherwise, the input data are identified as a total fault, which refers to a case where complex series and parallel faults have occurred simultaneously.

In this study, faults in PV system are defined as a condition where PV generation does not meet the user's expectation or does not generate electricity at all. Therefore, UCL and LCL would be chosen by the administrator, such as the owner or user of the PV system, as their expectations.

Fig. 14. Outdoor testing site.

V. Testing Site

An outdoor testing site was designed to verify the fault detection/diagnosis algorithm, as presented in Section III, and is shown in Fig. 14. The PV system was composed of 36 PV modules with the characteristics of 50 W (18 V and 2.77 A). A total of 12 modules are connected in series forming a string and three strings are connected in parallel to form an array. The specifications of the PV modules used in the test are listed in Table II. The geographical coordinates of the test site location were 36° 54′ 08.3″ N and 127° 32′ 26.4″ E. It was installed facing south, and the installation orientation was 30°. The data were collected for a week to determine the coefficients for the PV output prediction model. To enhance the accuracy of the model, data with less than 100 W/m^2 of irradiation were excluded from the analysis.

TABLE II
SPECIFICATIONS FOR PV MODULE

Cell (ea)	V_{oc} (V)	V_{mp} (V)	I_{sc} (A)	I_{mp} (A)	P_{mp} (W)
36	21.6	18	3.62	2.77	50

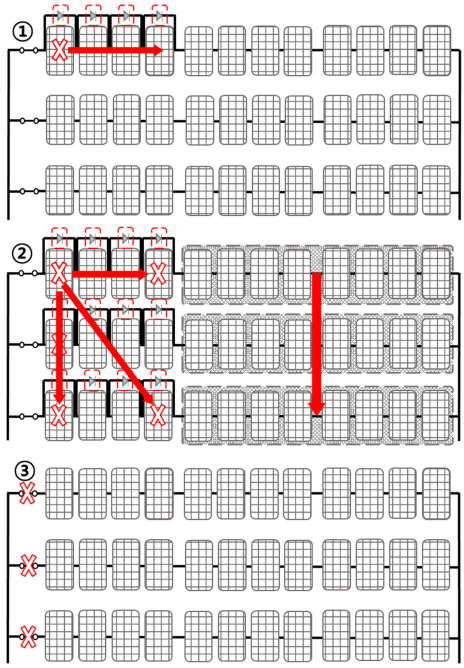

Fig. 15. Types of fault under test conditions (1: Series fault, 2: total fault, and 3: parallel fault).

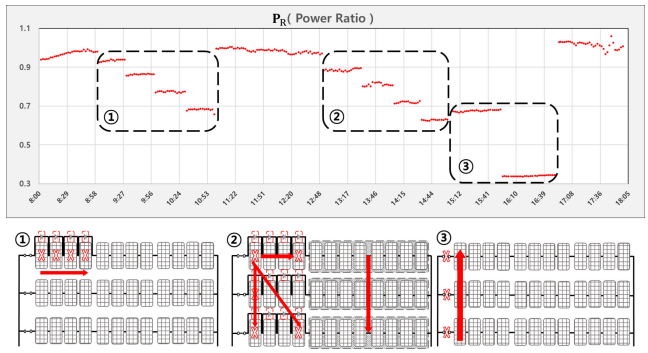

Fig. 16. Variation of P_R according to fault conditions (1: Series fault, 2: total fault, and 3: parallel fault).

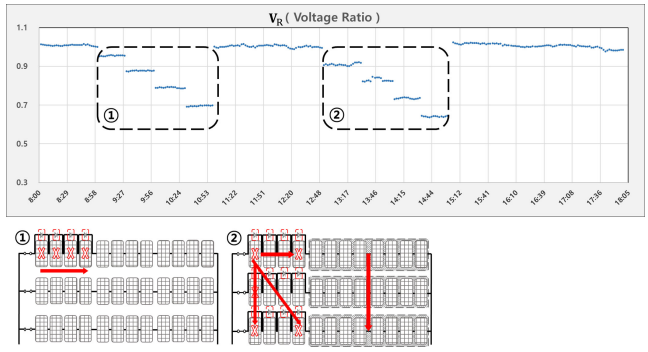

Fig. 17. Variation of V_R according to fault conditions (1: Series fault and 2: total fault).

The UCL and LCL of P_R, V_R, and I_R were defined using the initial output data of the PV system, which were considered during the normal operation. Therefore, UCL and LCL would be defined as different values for different testing sites.

Three types of fault conditions in the PV system were designed: series, parallel, and total. These are shown in Fig. 15. The series and parallel faults were defined in Section III-C. The total fault is a case where both voltage and current of PV array have dropped due to complex reasons, such as open-/short-circuited problem and shading. In this study, the total fault condition has been emulated by decreasing the current and voltage of the array. Current is decreased by covering the surface of the PV module with mesh, and voltage is decreased by increasing the number of short-circuited PV modules. The series fault was emulated by short circuiting a faulty module and then activating a bypass diode. The parallel fault was emulated by increasing the number of the open-circuited sting.

VI. TEST RESULT AND VALIDATION OF FAULT DETECTION/DIAGNOSIS ALGORITHM

The LCL and UCL of P_{mp}, V_{mp}, and I_{mp} were defined as 0.93–1.02, 0.99–1.01, and 0.93–1.05, respectively. This was derived from the output data of the testing site under normal operation. The validation process for the fault detection algorithm was created by artificially applying a short-circuited and open-circuited connection between the PV modules.

The change in P_R value for each fault situation is shown in Fig. 16. In the case of a series fault, P_R decreased in phase to 0.91, 0.85, 0.76, and 0.68 as the number of faulty modules increased. This result is outside the control range of LCL and UCL, which is defined as 0.93–1.02. In the case of a total fault, P_R shows a similar tendency as a series fault by decreasing to 0.88, 0.80, 0.71, and 0.62 as the number of faulty modules increased. Finally, for the case of a parallel fault, the P_R decreased sharply to 0.66 and 0.33 as connections between the array were short circuited. The result showed that, in general, P_R decreases in phase for all three fault types.

Fig. 17 shows changes in V_R of the PV system under each fault condition. In the case of a series fault, V_R dropped to 0.95, 0.87, 0.79, and 0.69 on average, which is below the control range of LCL and UCL, defined as 0.99–1.01. In the case of a total fault, the decrease in V_R is similar to that of a series fault with the values of 0.91, 0.82, 0.73, and 0.62. Finally, for a parallel fault, P_R showed no change as the connection between the strings was open circuited. The voltage of the PV array was not affected by a change in the number of parallel connections. In conclusion, V_R changes in phase under series and total fault conditions but not under a parallel fault condition. Therefore, V_R can be used as a criterion to diagnose parallel faults and the other two fault types: If V_R is within a designated range, and yet still there is a

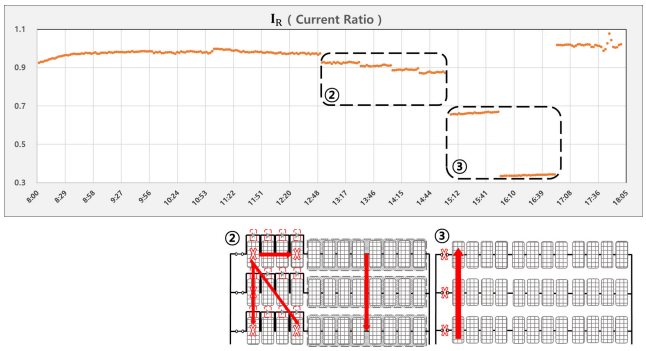

Fig. 18. Variation of I_R according to the fault condition (2: total fault and 3: parallel fault).

drop in P_R, it means that there is a parallel fault in the system; otherwise, there is either series or total fault.

Fig. 18 shows the change in I_R under each fault condition. In a series connection, the series fault does not affect the current of the PV array; therefore, there are no changes in I_R under series fault conditions.

The parallel fault and total fault conditions, however, cause a change in the current flowing through the circuit, which results in an abnormality of I_R. Under a parallel fault condition, for each open-circuited connection between three strings, I_R dropped to 0.66 and then to 0.33, indicating that I_R decreases proportionally to the number of short-circuited strings. Under a total fault condition, I_R dropped in phase to 0.92, 0.91, 0.89, and 0.87. From the test results, it is seen that each type of fault condition shows diversity in changes of P_R, V_R, and I_R; therefore, faults in the PV system can be diagnosed using the proposed method.

VII. Conclusion

This study presented a method to predict the output of a PV system using environmental variables and subsequently detect and diagnose faults in the system. The normal running condition of a PV system was defined from a regression model by utilizing adapted environmental variables.

There are various causes of faults in PV systems; these causes can be categorized as series, parallel, and total faults that eventually lead to a decrease in the electrical output of the system.

An output model was generated using the data of environmental variables collected from a test site. The accuracy of the V_{mp} prediction model was 0.81 MAPE (%) and 2.26 RMSE (%). The accuracy of the I_{mp} prediction model was 2.17 MAPE (%) and 0.083 RMSE (%). A prediction model of P_{mp} was designed by multiplying the predicted value of V_{mp} and I_{mp} with the smallest error rate, thus achieving the accuracies of 2.29 MAPE (%) and 18.67 RMSE (%). The modeled output data of the PV system were compared with the actual output data of a PV system collected from the same test site. P_R was used as the criterion to detect faults in the PV systems. When a fault is caused in a parallel connection, the current in the circuit is affected; however, if the fault is caused in a series connection, the voltage in the circuit is affected. Based on these characteristics of the PV system, faults were diagnosed using V_R and I_R.

Using this research, it is possible to realize a secure power generation through the integrated management of PV systems. The proposed system can enable the owners of PV systems and companies in operations and maintenance to monitor the operating conditions of a system through quantitative data. Furthermore, faults in the PV system can be detected and diagnosed through basic output data of the PV system without requiring the installation of additional monitoring systems. The innovation of this work is that the proposed method is capable of detecting a fault and diagnosing the adapt maintenance method of the PV system, especially in cases of floating and marine PV system, which have a vast deviation in output power because of their sensitiveness to changes in environmental conditions. This algorithm can be particularly useful to monitor and maintain ground, floating, and marine-based PVs, wherein the installation conditions are sensitive to sudden climatic changes.

For future works, the presented method could be used for the diagnosis of PV power plants using general PV modules. In particular, this analysis method is planned to be applied to develop and diagnose faults in a testbed of PV power plant composed with a high-density power module, such as bifacial or shingled PV module, which is capable of generating higher power under low-temperature condition.

References

[1] "Snapshot of global photovoltaic markets," Int. Energy Agency, Paris, France, Rep. IEA PVPS T1-33:2018, 2018.
[2] Y. Mahmoud and E. F. El-Saadany, "Enhanced reconfiguration method for reducing mismatch losses in PV systems," *IEEE J. Photovolt.*, vol. 7, no. 6, pp. 1746–1754, Nov. 2017.
[3] A. Walker, "PV O&M cost model and cost reduction," presented at Photovolt. Module Rel. Workshop, 2017. [Online]. Available: https://www.nrel.gov/docs/fy17osti/68023.pdf
[4] F. Touati, A. Massoud, J. A. Hamad, and S. A. Saeed, "Effects of environmental and climatic conditions on PV efficiency in Qatar," in *Proc. Int. Conf. Renewable Energies Power Qual.*, Mar. 2013, pp. 262–267.
[5] H. A. Kazem, T. T. N. Khatib, and A. Alwaeli, "Optimization of photovoltaic modules title angle for Oman," in *Proc. 7th IEEE Int. Power Eng. Optim. Conf.*, Jun. 2013, pp. 700–704.
[6] M. Catelani *et al.*, "Characterization of photovoltaic panels: The effect of dust," in *Proc. IEEE Int. Energy Conf. Exhib.*, 2012, pp. 45–50.
[7] B. G. Bhang *et al.*, "Power performance of bifacial c-Si PV module with different shading ratios," *IEEE J. Photovolt.*, vol. 9, no. 5, pp. 1413–1420, Sep. 2019.
[8] H. Cha, B. G. Bhang, S. Park, J. Choi, and H. Ahn, "Power prediction of bifacial Si PV module with different reflection conditions on rooftop," *Appl. Sci.*, vol. 8, no. 10, pp. 1752–1771, Sep. 2018.
[9] G. G. Kim *et al.*, "Prediction model for PV performance with correlation analysis of environmental variables," *IEEE J. Photovolt.*, vol. 9, no. 3, pp. 832–841, May 2019.
[10] S. Dubey, J. N. Sarvaiya, and B. Seshadri, "Temperature dependent photovoltaic (PV) efficiency and its effect on PV production in the world—A review," *Energy Procedia*, vol. 33, pp. 311–321, 2013.
[11] C. M. Whitaker *et al.*, "Effects of irradiance and other factors on PV temperature coefficients," in *Proc. Conf. Rec. 22nd IEEE Photovolt. Spec. Conf.*, 1991, pp. 608–613.
[12] B. Rupnik and O. Westbrook, "Ambient temperature correction of photovoltaic system performance data," in *Proc. IEEE 40th Photovolt. Spec. Conf.*, 2014, pp. 1973–1977.
[13] W. C. L. Kamuyu, J. R. Lim, C. S. Won, and H. K. Ahn, "Prediction model of PV module temperature for power performance of floating PVs," *Energies*, vol. 11, no. 2, Feb. 2018, Art. no. 447.
[14] S. Mekhilef, R. Saidur, and M. Kamalisarvestani, "Effect of dust, humidity and air velocity on efficiency of photovoltaic cells," *Renewable Sustain. Energy Rev.*, vol. 16, pp. 2920–2925, 2012.

[15] H. A. Kazem and M. T. Chaichan, "Effect of humidity on photovoltaic performance based on experimental study," *Int. J. Appl. Eng. Res.*, vol. 10, no. 5, pp. 3979–3982, 2016.
[16] J. K. Kaldellis, M. Kapsali, and K. A. Kavadias, "Temperature and wind speed impact on the efficiency of PV installations: Experience obtained from outdoor measurements in Greece," *Renewable Energy*, vol. 66, pp. 612–624, Jun. 2014.
[17] "Photovoltaic system performance monitoring guidelines for measurement, data exchange and analysis," IEC 61724, Edition 1.0, 1998.
[18] N. H. Reich *et al.*, "Performance ratio revisited: Is PR >90% realistic?" *Prog. Photovolt., Res. Appl.*, vol. 20, pp. 717–726, 2012.
[19] L. Bonsignore, M. Davarifar, A. Rabhi, G. M. Tina, and A. Elhajjaji, "Neuro-fuzzy fault detection method for photovoltaic systems," *Energy Procedia*, vol. 62, pp. 431–441, Dec. 2014.
[20] P. Tingting and H. Xiaohong, "A fault detection method for photovoltaic systems based on voltage and current observation and evaluation," *Energies*, vol. 12, 2019, Art. no. 1712, doi: 10.3390/en12091712.
[21] H. Haeberlin and C. Beutler, "Normalized representation of energy and power for analysis of performance and on-line error detection in PV-systems," in *Proc. 13th Eur. Photovolt. Sol. Energy Conf.*, Nice, France, 1995.
[22] J.-H. So *et al.*, "Design factor calculation and analysis of grid-connected photovoltaic system," *J. Korean Sol. Energy Soc.*, vol. 33, no. 5, pp. 89–94, 2013.
[23] F. Harrou, Y. Sun, B. Taghezouit, A. Saidi, and M.-E. Hamlati, "Reliable fault detection and diagnosis of photovoltaic systems based on statistical monitoring approaches," *Renewable Energy*, vol. 116, pp. 22–37, 2018.
[24] Introducing machine learning, Section 1, Mathworks. 2020. [Online]. Available: https://kr.mathworks.com/campaigns/offers/machine-learning-with-matlab.html?elqCampaignId = 10588
[25] Z.-C. Chen *et al.*, "Random forest based intelligent fault diagnosis for PV arrays using array voltage and string currents energy," *Energy Convers. Manage.*, vol. 178, pp. 250–264, 2018.
[26] Z. Chen, Y. Chen, L. Wu, S. Cheng, and P. Lin, "Deep residual network based fault detection and diagnosis of photovoltaic arrays using current-voltage curves and ambient conditions," *Energy Convers. Manage.*, vol. 198, 2019, Art. no. 111793.
[27] N. Park and H. K. Ahn, "Multi-layer RNN-based short-term photovoltaic power forecasting using IoT dataset," in *Proc. AEIT Int. Ann. Conf.*, Florence, Italy, 2019, pp. 1–5, doi: 10.23919/AEIT.2019.8893348.
[28] A. Mellit, G. M. Tina, and S. A. Kalogirou, "Fault detection and diagnosis methods for photovoltaic systems: A review," *Renewable Sustain. Energy Rev.*, vol. 91, pp. 1–17, 2018.
[29] C. Spearman, "The proof and measurement of association between two things," *Amer. J. Psychol.*, vol. 100, no. 3/4, pp. 441–471, 1904.
[30] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise Reduction in Speech Processing*, vol. 2. New York, NY, USA: Springer, Mar. 2009. [Online]. Available: https://link.springer.com/chapter/10.1007%2F978-3-642-00296-0_5
[31] C. Thirumalai, S. A. Chandhini, and M. Vaishnavi, "Analysing the concrete compressive strength using Pearson and Spearman," in *Proc. Int. Conf. Electron., Commun. Aerosp. Technol.*, Dec. 2017, pp. 215–218.
[32] W. Revelle, Psych v1.8.4, 2018. [Online]. Available: https://www.rdocumentation.org/packages/psych/versions/1.8.4/topics/pairs.panels
[33] W. C. Lawrence, J. R. Lim, C. S. Won, and H. K. Ahn, "Prediction model of photovoltaic module temperature for power performance of floating PVs," *Energies*, vol. 11, 2018, Art. no. 447.
[34] D. L. King, W. E. Boyson, and J. A. Kratochvill, "Photovoltaic array performance model," Sandia Rep. SAND2004-3535, Dec. 2004.