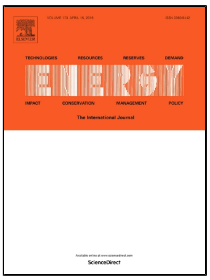
Accepted Manuscript



Development of Photovoltaic Abnormal Condition Detection System Using Combined Regression and Support Vector Machine

Fauzan Hanif Jufri, Seongmun Oh, Jaesung Jung

PII:	S0360-5442(19)30635-8
DOI:	10.1016/j.energy.2019.04.016
Reference:	EGY 15045
To appear in:	*Energy*
Received Date:	27 June 2018
Accepted Date:	03 April 2019

Please cite this article as: Fauzan Hanif Jufri, Seongmun Oh, Jaesung Jung, Development of Photovoltaic Abnormal Condition Detection System Using Combined Regression and Support Vector Machine, *Energy* (2019), doi: 10.1016/j.energy.2019.04.016

Development of Photovoltaic Abnormal Condition Detection System Using Combined Regression and Support Vector Machine

Fauzan Hanif Jufri[1], Seongmun Oh[2], Jaesung Jung[*2]

[1]*Electric Power and Energy Studies (EPES), Department of Electrical Engineering, Universitas Indonesia, Depok, Indonesia*

[2]*Department of Energy Systems Research, Ajou University, Suwon, South Korea*

*Corresponding author. Tel.: +82-31-219-2695; e-mail: jjung@ajou.ac.kr (J. Jung)

Postal address: Power System Laboratory, Energy Center 210, Ajou University, 206 Worldcup-ro, Yeongtong-gu, Suwon, South Korea 16499

Abstract

It is essential to monitor and detect the abnormal conditions in Photovoltaic (PV) system as early as possible to maintain its productivity. This paper presents the development of a PV abnormal condition detection system by combining regression and Support Vector Machine (SVM) models. The regression model is used to estimate the expected power generation under the respective solar irradiance, which is used as the input for the SVM model. The SVM model is then used to identify the abnormal condition of a PV system. The proposed model does not require installing additional measurement devices and can be developed at low cost, because the data that is used as the input variable for the model is retrieved from the Power Conversion System (PCS). Furthermore, the accuracy of the detection system is improved by taking into consideration the daylight time and the interactions between the independent variables, as well as the implementation of the multi-stage k-fold cross-validation technique. The proposed detection system is validated by using actual data retrieved from a PV site, and the results show that it can successfully distinguish the normal condition, as well as identify the abnormal condition of a PV system by using the basic measurements.

Keywords: Photovoltaic, PV Abnormal Detection, PV Fault Detection, Support Vector Machine (SVM)

1. Introduction

Photovoltaic (PV) is considered as a promising alternative energy generation to replace the fossil-based energy generation which is less environmental-friendly due to its CO_2 emission. However, the implementation of PV system involves certain challenges such as the uncertainty, efficiency, and reliability. In addition, PV system is susceptible to various environmental conditions such as dust, humidity, debris intrusion, animal interference, and human error, which may lead to the failure of the PV system and affect its productivity. On the other hand, the assurance of consistent productivity necessitates expanding the implementation of PV system as an alternative energy resource. Therefore, the occurrence of failures must be minimized so that the productivity of the PV system can be maintained at its expected level in order to obtain optimum technical and economic benefits.

Triki-Lahiani et al. classified the failures of PV system into three categories, viz., PV module failures, inverter failures, and other components failures [1]. Amongst these, PV module failure is closely related to the

environmental disturbances, and includes module overheating, blocked surfaces, broken module or connections, moisture or insect penetration, module stealing, etc. Whereas the inverter and other components failures include failures due to overvoltage, short-circuit faults, overheating, mechanical vibration, etc. that damages the components. These type of disturbances cause PV system to generate less power even when the solar irradiance incident on the surface of the PV panel is high [2]–[4]. In another words, PV system may not operate normally and generate the expected amount of power because of these failures. Therefore, it is important to provide an early detection system that can recognize the abnormal conditions in the PV systems, and immediate measures can be taken to prevent further damages and losses.

Methods to identify the abnormal conditions of PV systems have been formulated by some studies. Additional sensors were used to detect the abnormal condition of PV system by using infrared technology or voltages and current sensors. Infrared cameras were used to observe the abnormal condition of PV modules in [5]–[8], wherein the cameras collected images of the PV module, which were then processed to verify the status of the modules. Whereas in [9] and [10], voltage and current sensors were added to individual PV modules, following which the obtained data were processed to determine the abnormal condition of PV system. These methods can provide a real-time output and a precise location of the failure at the module level. However, they require the installation of a considerable number of additional devices, which becomes costly, and hence, can only be feasible for a large-scale PV plant.

Another methodology to monitor the abnormal condition is by analyzing the power loss using the I-V characteristics of PV systems [11]–[15]. The power loss is determined by calculating the difference between the ideal and the actual power. The ideal power is evaluated by retrieving the parameters of an ideal one-diode PV model, such as the photo-generated current, dark saturation current, series and shunt resistance of PV, and the diode's ideality factor. However, the determination of these parameters uses the assumption that all modules in a PV array have the same characteristics, whereas under actual conditions each PV module will have different characteristics. In addition, this method applies an iterative technique which is time-consuming, may have convergence issues, and the results may vary when different initialization is applied [20].

Statistical methods can also be used to identify the condition of PV system. In [16], the approach used the least square method to obtain the residual between the ideal and actual parameters of PV systems such as power and voltage. It also included the input of shading and solar irradiance information into the fuzzy inference system in order to determine the condition of the PV module. In [17], the measured variables of the PV system were checked for significance by using a *t*-test, and the ratio between the ideal and actual power and voltage were calculated. The ratio was then compared with thresholds obtained from a training algorithm to determine the abnormal conditions. In [18], the Exponentially Weighted Moving Average (EWMA) control chart was used to detect PV system condition. This method used the normal historical dataset to obtain the time-weighted statistical information in order to recognize deviations in the new dataset that are considered as an abnormality. In [19], the statistical distribution of the variables of the PV system was analyzed to define the boundary of the normal operating limits, wherein the PV system was classified as abnormal when a new variable of the system occurred beyond the defined boundary. However, these methods only consider the failures related to the PV modules, and do not include the abnormal conditions caused by failures of the inverter and other components.

In this paper, a different approach has been used to calculate the ideal power generation and takes into consideration all three categories of failures of PV systems, i.e., failures of PV module, inverter, and other

components, for detection of abnormal conditions in the PV system. Moreover, the proposed method uses variables that are even available in a small-scale PV system and installation of additional costly sensory devices are not needed. These variables include power, voltage, and current, which can be retrieved from the Power Conversion System (PCS), solar irradiance on the surface of the PV panels, which is measured using the pyranometer, and ambient and PV cell temperatures that are measured using the thermometer. Therefore, the proposed PV abnormal condition detection system can be effectively used in a small-scale PV system, or as an early warning system to the PV operator/owner for conducting further investigations on the system.

The PV abnormal condition detection system has been developed using regression and Support Vector Machine (SVM) models. The regression analysis is used to obtain a new variable, which is the expected power generation. Whereas, the SVM model is used to identify the abnormal condition of PV by using various variables, including the expected power generation obtained from the regression model. The detection model has been improved by considering the daylight time, incorporating additional variables that can be obtained from the interaction between independent variables, and implementing the multi-stage *k*-fold cross-validation technique. Actual data retrieved from a PV site is used to validate the proposed detection system.

The rest of this paper is organized as follows: Section 2 presents the development of the models used in the PV abnormal condition detection system, which are regression and SVM algorithms. Section 3 describes the strategies for improving the accuracy of the detection system. Section 4 details the simulation results, and finally Section 5 provides the conclusion.

2. PV Abnormal Condition Detection System

The proposed detection model aims to distinguish between the normal and abnormal operating conditions of PV system. A normal operating condition is defined as a state when the PV system generates as much power as is proportionally possible to the amount of solar irradiance incident on the surface of the PV panel. Whereas, an abnormal operating condition is described as a state when the PV system generates less power than the expected in accordance with the solar irradiance. However, classifying a state as an abnormal condition, when the actual power generated is much lower than the expected power generation, is incorrect because the solar irradiance and the generated power do not follow a linear relationship. The influence of temperature, daylight time, and maximum power point characteristics must also be considered in addition to the comparison between the expected and actual power generation. Therefore, further analysis and a more appropriate algorithm are required to distinguish this phenomenon.

A PV system generally consists of PV modules and PCS, including charge controller, inverter, and protection and metering equipment. In addition, it is equipped with a pyranometer to measure solar irradiance, and thermometers to measure ambient and PV cell temperatures. The data retrieved from these equipments are processed to determine the condition of the PV system. The proposed detection system does not require installing any additional devices. Therefore, this approach can be considered more economical compared to the application of expensive sensory devices, while being utilized for the same purpose.

The configuration of the PV abnormal condition detection system is presented in **Fig. 1**. The PCS measures the data such as power generation (P), DC voltage (V), and DC current (I), as well as collects the data of solar irradiance (S) from the pyranometer, and ambient and PV cell temperatures (T_A and T_{PV}) from the thermometer.

The measurement data is stored in the data logger and then processed through two different steps. Firstly, it is recorded in the database to be used for model development, which can be performed frequently or by request. Secondly, it is used to determine the condition (H) of PV system.

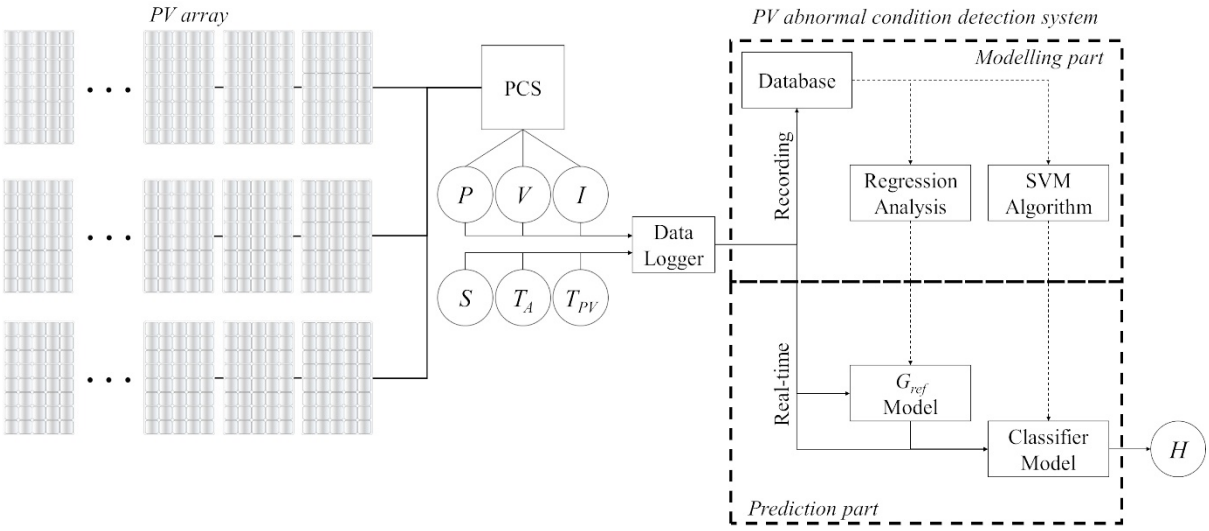


Fig. 1. PV system configuration and the measurement point

Two models are developed in the PV abnormal condition detection system. The first model is to estimate the expected PV generation (G_{ref}) by using the regression model and the second one is to detect PV abnormal condition by using the SVM model.

2.1. Regression model to estimate the expected PV generation

PV generation ideally has a proportional correlation with the solar irradiance measured on the panel surface, given that the PV system operates under normal circumstances. Therefore, calculation of the expected power generation using solar irradiance is an important input in determining the condition of the PV system. Other than the ideal one-diode PV model, the power generated from the neighboring PV system can also be used as a substitute to estimate the expected power generation, as long as the amount of solar irradiance incident on the surface of the PV panel and weather conditions between the two systems are typical. However, this method can only be applied when the neighboring PV system exists in the same area.

To solve issues relating to extraction of ideal PV parameters or in the absence of information from the neighboring PV system, this paper proposes a model using regression analysis. The expected PV generation is determined by the following general equation:

$$G_{ref} = \beta X + \varepsilon \quad (1)$$

$$\beta = (X^T X)^{-1} X^T G_{ref} \quad (2)$$

where G_{ref} is the expected power generation (Watt), β is the vector of the regression parameters, and X is the matrix of the set of input variables. The set of input variables is a combination of variables including solar irradiance, ambient temperature, and PV cell temperature. Ordinary Least Square (OLS) method is used to estimate the regression parameters (β) by using the set of data under normal operating conditions as shown in (2).

2.2. SVM model to detect PV abnormal condition

The difference between the expected and measured power generation alone is unable to identify the abnormal condition of PV system with high accuracy because of their non-linear relationship and factors such as temperatures, daylight time, and the characteristic of maximum power point. It requires variables such as power, voltage, current, solar irradiance, ambient and PV cell temperatures to provide additional information in determining the condition of the PV system. Hence, the SVM algorithm is used since it is widely known and proven for solving classification problems with non-linear data.

The SVM is a supervised learning algorithm widely used to classify either binary or multi-classes objects. Moreover, SVM with kernels can be used for more complex and non-linear classifications even in the case of limited amount of data. SVM detection is based on the concept of margin detection by building a hyperplane (or hyperplanes) that equally separate two or more classes of objects. A proper separator is the one that has the longest distances to the nearest objects. Therefore, SVM is also often called as Large Margin Classifier. The nearest objects are called the support vectors, and the hyperplanes are usually defined by the kernel function. The hyperplane that classifies the objects is defined by the following equation:

$$H_\theta(x) = \theta^T f = \theta_1 f_1 + \theta_2 f_2 + \theta_3 f_3 + \ldots + \theta_n f_n \quad (3)$$

where $H_\theta(x)$ is the detection system equation, θ represents SVM parameters, and f is the new feature that is defined by the kernel function. One of the commonly used non-linear kernels in SVM classification is the Radial Basis Function (RBF), also called the Gaussian kernel, which is defined by the following equation:

$$f_i = exp\left(-\gamma \|x - x^{(i)}\|^2\right) \quad (4)$$

where f_i is the new feature and defined as the similarity value of two objects, x and $x^{(i)}$. x is the object under investigation, $x^{(i)}$ represents the other objects in the space, and γ is the parameter of the Gaussian kernel.

The parameter θ in (3) can be estimated through the optimization of the penalty function based on the following equation:

$$\min_\theta C\sum_{i=1}^{m}\left(y^{(i)}\text{cost}_1(\theta^T f^{(i)}) + (1-y^{(i)})\text{cost}_0(\theta^T f^{(i)})\right) + \frac{1}{2}\sum_{j=1}^{n}\theta_j{}^2 \quad (5)$$

where C is the SVM regularization parameter, $y^{(i)}$ is the i-th actual output, cost_1 and cost_0 are the penalty functions for classifying when $y^{(i)} = 1$ and when $y^{(i)} = 0$, respectively. They can be defined as follows:

$$\text{cost}_1(\theta^T f^{(i)}) = \max\left(0, (1 + \theta^T f^{(i)})\right) \quad (6)$$

$$\text{cost}_0(\theta^T f^{(i)}) = \max\left(0, (1 - \theta^T f^{(i)})\right) \quad (7)$$

The SVM regularization parameter (C) and the Gaussian kernel parameter (γ) can be any number. These two parameters are tightly coupled with each other [21]. For instance, a larger value of γ means a narrow kernel and thus, lower variance. Therefore, it requires a smaller value of C. However, a small value of C may cause the model to fail to generalize the classification. Therefore, the SVM parameters must be carefully selected to obtain a robust and adaptive classification model. In this paper, C and γ are determined through cross-validation technique.

3. Improvement Strategies for PV Abnormal Condition Detection System

3.1. Consideration of the daylight time

The measurement data recorded in the database does not consider the daylight time, which means that the recording takes place for a certain interval within 24 hours. However, a PV system generates power during the

period between the sunrise and the sunset. Therefore, the proposed PV abnormal condition detection system has been developed to operate within this daylight time window. The sunrise and sunset time of a day at a particular place can be estimated by using the latitude and solar declination on that day at the examined site [22]. It can be mathematically calculated by the following equations:

$$SR = 12{:}00 - \left(H_{SR}/15\right) - 4(LTM - LG) - E \quad (8)$$

$$H_{SR} = \cos^{-1}\left(-\tan LT \tan \delta\right) \quad (9)$$

$$E = 9.87\sin 2\omega - 7.53\cos \omega - 1.5\sin \omega \quad (10)$$

$$\omega = \left(360/364\right) \times (n - 81) \quad (11)$$

$$\delta = 23.45\sin\left(\left(360/365\right) \times (n - 81)\right) \quad (12)$$

where *SR* is the sunrise/sunset time, H_{SR} is sunrise hour angle, *LTM* is Local Time Meridian, *LG* is the site longitude, *E* is the equation of time, *LT* is the site latitude, δ is the solar declination, and *n* is the day number in a year. The value of H_{SR} may be positive or negative due to the inverse cosine, such that the positive value is used to find the sunrise time and the negative value is for sunset time.

Additionally, it is observed from the dataset that the power generation is generally lower during the time from the sunrise to 1 hour after sunrise, and from 1 hour before the sunset until the sunset. Therefore, the exclusion of this period in the model development can also improve the accuracy of the detection system [23].

3.2. The interactions between independent variables

The dependent variable, i.e., the output of the detection system represents the condition of the PV system, whether it is normal or abnormal. This variable is dependent on power generation, voltage, current, solar irradiance, ambient and PV cell temperatures, which are considered as independent variables from the point of view of the output. However, these independent variables or so-called the input variables are dependent each other which means that if the value of one variable changes, then the value of the other variable that depends on it will also change. Therefore, the dependency between input variables is an important factor for improving the accuracy and sensitivity of the detection model. The interaction between the independent variables can be evaluated by the following equations [22]:

$$P = P_{rated}\left(1 - \varphi\left(T_{PV} - 25℃\right)\right) \quad (13)$$

$$T_{PV} = T_A + \left(\frac{NOCT - 20°}{0.8}\right) \cdot S \quad (14)$$

where *P* is the power generation considering the PV cell temperature, P_{rated} is the rated power of the PV system dependent on its design and assuming a full solar irradiance condition, T_{PV} is the PV cell temperature, T_A is the ambient temperature, *NOCT* is the Nominal Operating Cell Temperature, *S* is the solar irradiance, and φ is the temperature coefficient of power.

It can be observed that power generation is dependent upon the PV cell temperature, as well as proportional to the solar irradiance. Alternately, the PV cell temperature is dependent upon the ambient temperature and the solar irradiance. Higher solar irradiance can generate more power, but it also increases the PV cell temperature, which may ultimately degrade the power generation. Therefore, it requires additional variables as inputs that can provide more information about the interactions between the main variables. This interaction is commonly represented by the product of the respective variables [24]–[26]. Equation (13) and (14) define the correlations between the input

variables and their interactions that produce new variables (termed as interactions variables) can be summarized as follows:

- Power generation (P) is influenced by solar irradiance (S), ambient temperature (T_A), and PV cell temperature (T_{PV}). Thus, new variables inherited from these interactions are indicated as PS, PT_A, and PT_{PV}.
- The PV cell temperature (T_{PV}) is influenced by the solar irradiance (S) and ambient temperature (T_A). Thus, new variables inherited from these interactions are indicated as $T_{PV}S$ and $T_{PV}T_A$.

3.3. Multi-stage k-fold cross-validation technique

The expansion of the input variables results in multiple possible regression and SVM models for detecting the abnormal condition in a PV system. All possible models are presented by the configuration of the independent and interactions variables along with their respective parameters. The regression and SVM models that are finally selected for the PV abnormal condition detection system are determined based on the lowest Mean Absolute Deviation (MAD). The optimization function for the regression model is expressed by the following equations:

$$\min_{X,\beta}\left\{MAD_X^{REG}\right\} = \min_{X,\beta}\left\{\frac{\sum_{i=1}^{n}\left|P^{(i)} - G_{ref,X}^{(i)}\right|}{n}\right\} \tag{15}$$

where MAD_X^{REG} is the MAD of the regression model with the variable configuration X, β is the regression parameter for the respective variable configuration, $P^{(i)}$ is the i-th measurement data of the power generation, $G_{ref,X}^{(i)}$ is the i-th expected power generation calculated by the regression model with the configuration X, and n is the number of data.

On the other hand, the optimization function for selecting the SVM model is expressed by the following equation:

$$\min_{X,C,\gamma,\theta}\left\{MAD_{X,C,\gamma}^{SVM}\right\} = \min_{X,C,\gamma,\theta}\left\{\frac{\sum_{i=1}^{n}\left|H_{actual}^{(i)} - H_{predict,X,C,\gamma}^{(i)}\right|}{n}\right\} \tag{16}$$

where $MAD_{X,C,\gamma}^{SVM}$ is the MAD of the SVM model with the variable configuration X, regularization parameter C, and Gaussian kernel parameter γ. θ represents the SVM parameters of the respective variable configuration, $H_{actual}^{(i)}$ is the i-th actual condition of the PV system, $H_{predict,X,C,\gamma}^{(i)}$ is the i-th predicted condition of the PV system, which is calculated by the SVM model with configuration X, C, and γ, and n is the number of data.

The models are selected by solving (15)–(16) through multi-stage k-fold cross-validation technique. The first stage of the k-fold cross-validation is to find the input variable configuration (X) for the regression model and their related parameters (β). Whereas, the second stage of the k-fold cross-validation is to find the input variables configuration (X) for the SVM model and their related SVM parameters (θ), as well as regularization parameter (C) and the kernel parameter (γ). The process flow of the multi-stage k-fold cross-validation technique to develop the PV abnormal condition detection system is shown in **Fig. 2**.

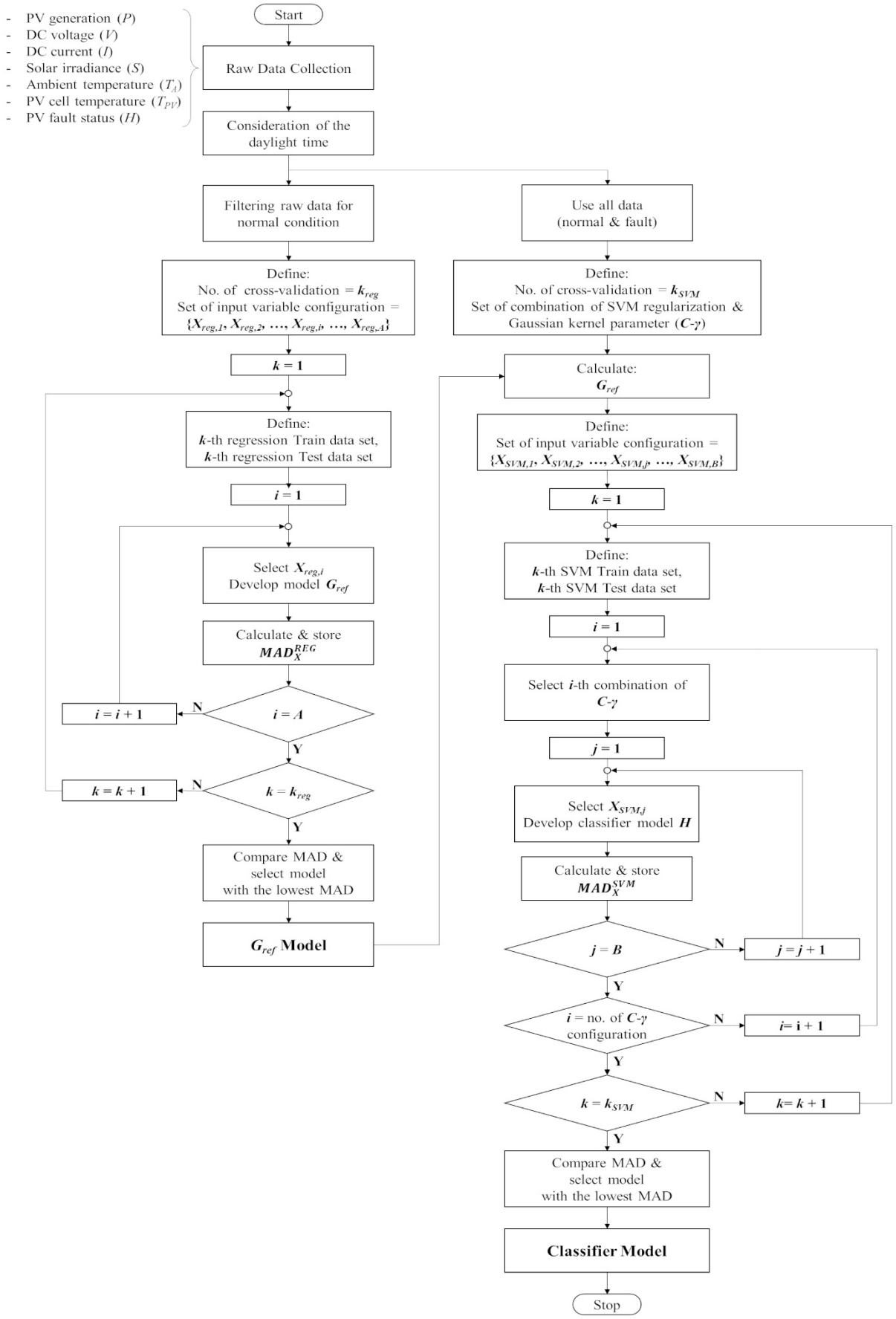


Fig. 2. The multi-stage k-fold cross-validation technique process flow

First, the raw data retrieved from the PCS, pyranometer, and thermometer is refined by taking into account the daylight time. The first stage of the *k*-fold cross-validation is performed to obtain the most suitable regression model to estimate the expected PV generation. The steps involved are explained as follows: (1) the data is filtered based on the normal conditions of the PV system, (2) various configurations of the input variables are constructed by including the independent and interactions variables, and by implementing the rules of hierarchically well-formulated methodology [27]–[29], (3) the data is divided into *k* number of groups, the training dataset is selected from *k*-1 group, and the remaining dataset is selected as the cross-validation dataset, (4) the input variable configuration is established for each pair of the dataset and the regression parameters are calculated, and finally, (5) the MAD_X^{REG} for each combination is recorded. This process is repeated for all input variables configuration until all the pairs of cross-validation data have been assessed. All the recorded MADs are then compared, and the model with the lowest MAD is selected as the final model to be used for estimating the expected PV generation.

The second stage of the *k*-fold cross-validation is performed to determine the SVM model to detect the condition of the PV system. All the refined data post consideration of the daylight time is used to develop this model. The steps involved in the process can be explained as: (1) the set of the SVM regularization and Gaussian kernel parameters is defined based on the exponentially growing values such as $C \in \{e^{-7},e^{-6},...,e^{5}\}$, and $\gamma \in \{e^{-7},e^{-6},...,2^{5}\}$ [30], (2) the data is divided into *k* number of groups, the train dataset is selected from the *k*-1 group and the remaining dataset is selected as the cross-validation dataset, (3) for each pair of the dataset, the expected PV generation is calculated, (4) the set of input variables configuration is defined, (5) for each combination of *C, γ*, and input variables configuration, the SVM model parameters are calculated, and finally, (6) the $MAD_{X,C,\gamma}^{SVM}$ for each combination is recorded. Eventually, the SVM model with the lowest MAD value is selected as the final model.

4. Simulations and Results

The simulation has been carried out by utilizing the data retrieved from a PV system located in Beonyeong-ro, Jeju City, South Korea, having a capacity of 1,600 Wp. Simulations of four case studies are presented in this paper to evaluate the proposed PV abnormal condition detection system, as well as the strategies that can be implemented to improve the system performance. The first and the second case study represent the reference models, and the third and the fourth case study represent the combination of both methods as well as the improvement strategies proposed in this paper. The first case study is performed by comparing the actual PV generation with the expected PV generation resulted by a regression analysis. The second case study develops the detection system by using the traditional SVM model by using standard variables available in PCS such as power generation, DC voltage, DC current, solar irradiance, ambient temperature, and PV cell temperature. In the third case study, the detection system is developed by utilizing the expected PV generation obtained from the regression analysis into the SVM model, as explained earlier in Section 2. The fourth case study considers the expected PV generation from the regression analysis and implements the improvement strategies described in Section 3. The summary of each simulation is presented in **Table 1**.

Table 1. The Simulation Case Study

Case	Detection System Models	Input Variables
1	Regression analysis	G_{ref}, P
2	Traditional SVM	P, V, I, S, T_A, T_{PV}
3	SVM and regression	G_{ref}, P, V, I, S, T_A, T_{PV}
4	SVM, regression, and improvement strategies	G_{ref}, P, V, I, S, T_A, T_{PV}, PS, PT_A, PT_{PV}, $T_{PV}S$, $T_{PV}T_A$

The model is developed using actual dataset obtained from an established PV generation site. Each model in each of the three case studies is tested on the other acquired data that is separated from the data used to develop the model. Two days (Day-1 and Day-2) represent scenarios when the abnormal condition occurs randomly, one day (Day-3) represents a scenario when there is no abnormal condition during the entire day, and one day (Day-4) represents a scenario when the abnormal condition persists during the whole day.

The accuracy of the detection system is evaluated as a representation of the True Positive Rate (*TPR*), the True Negative Rate (*TNR*), and the Total Accuracy (*TA*). The *TPR* is used to evaluate the accuracy of predicting the abnormal condition, while *TNR* is used to evaluate the accuracy of distinguishing the normal condition to prevent sending false alarm to the operator. Accordingly, *TPR* is defined as the ratio of the number of abnormal conditions detected correctly (*CF*) to the total number of actual abnormal conditions (*TF*), and *TNR* is defined as the ratio of the number of the normal conditions detected correctly (*CN*) to the total number of the actual normal conditions (*TN*). Finally, *TA* is defined as the ratio of the number of correct results (*CR*) to the total samples (*TS*). They can be mathematically expressed by the following equations:

$$TPR = (CF/TF) \times 100\% \tag{17}$$

$$TNR = (CN/TN) \times 100\% \tag{18}$$

$$TA = (CR/TS) \times 100\% \tag{19}$$

4.1. Case 1: Regression analysis

In the first case study, the PV abnormal condition detection system is developed by evaluation the actual PV generation and the expected PV generation which is developed by using a regression analysis. The difference between these variables is compared with an optimal threshold to classify the condition of PV systems as normal or abnormal. In this case study, the optimal threshold is calculated as 20%, which means that if the difference between the actual and expected PV generation is less than 20% then, it is considered as normal, otherwise, it is considered as abnormal. The results obtained for all four days are presented in **Fig. 3** and **Table 2**.

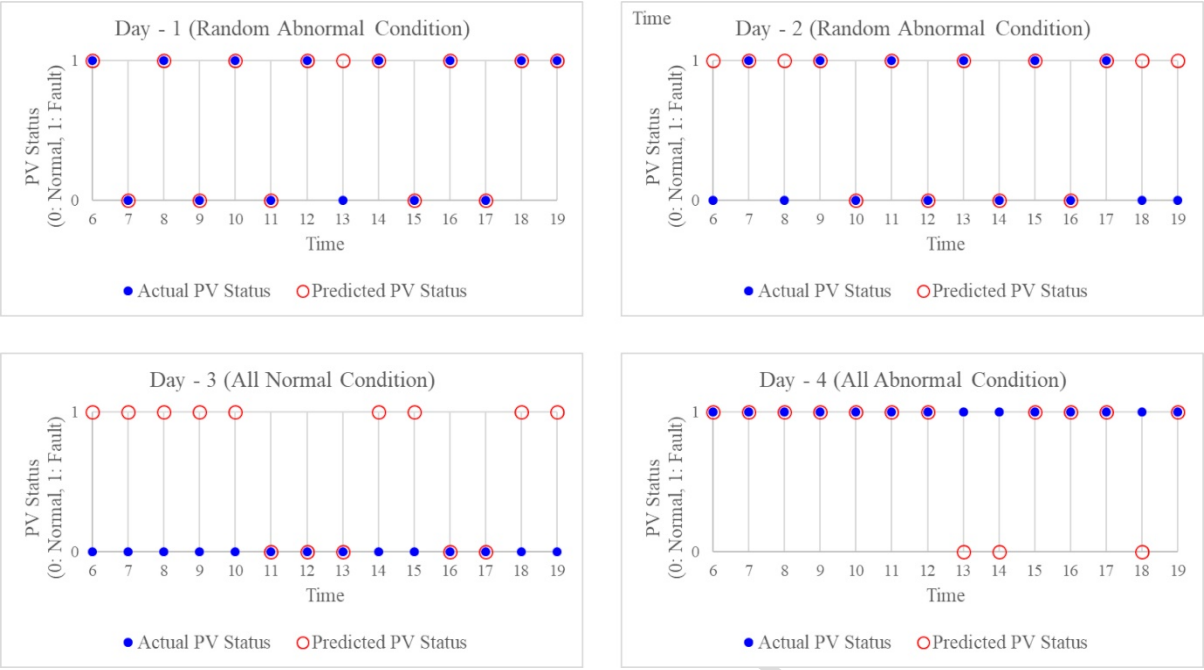


Fig. 3. Comparison between the actual and the predicted PV abnormal condition (Case 1)

Table 2. TPR, TNR, and TA Results for Each Day in Case 1

Day	TPR (%)	TNR (%)	TA (%)
Day-1 (Random Abnormal Condition)	100	83.33	92.86
Day-2 (Random Abnormal Condition)	100	50	71.43
Day-3 (All Normal Condition)	-	35.71	35.71
Day-4 (All Abnormal Condition)	78.57	-	78.57
Average	**89.29**	**50**	**69.64**

*The *TPR* for all normal condition and *TNR* for all abnormal condition cannot be defined since they have zero denominators (*TF* = 0 and *TN* = 0 respectively in TPR and TNR).

The detection system can adequately detect the abnormal condition, as indicated by 100% of *TPR* in Day-1 and Day-2, and 78.57% of *TPR* in Day-4, thereby having an average *TPR* of 89.29%. However, it fails to distinguish the normal condition appearing in Day-2 and Day-3 with 50% and 35.71% of *TNR*, respectively, even though the *TNR* of Day-1 is relatively high (83.33%). The average *TNR* is observed as 50%, which means that the detection system can only distinguish half of the normal conditions. In addition, the average *TA* of the prediction system in this case is 69.64%, which indicates that the comparison method by using regression analysis is unable to distinguish normal conditions in a PV system with high accuracy.

4.2. Case 2: Traditional SVM

In this second case study, the PV abnormal condition detection system is developed by using the traditional SVM model. All input variables used to develop the SVM model are acquired from the available measurement equipment. These include the measured power generation, DC voltage, and DC current, solar irradiance, ambient, and PV cell temperatures. The results obtained for all four days are presented in **Fig. 4** and **Table 3**.

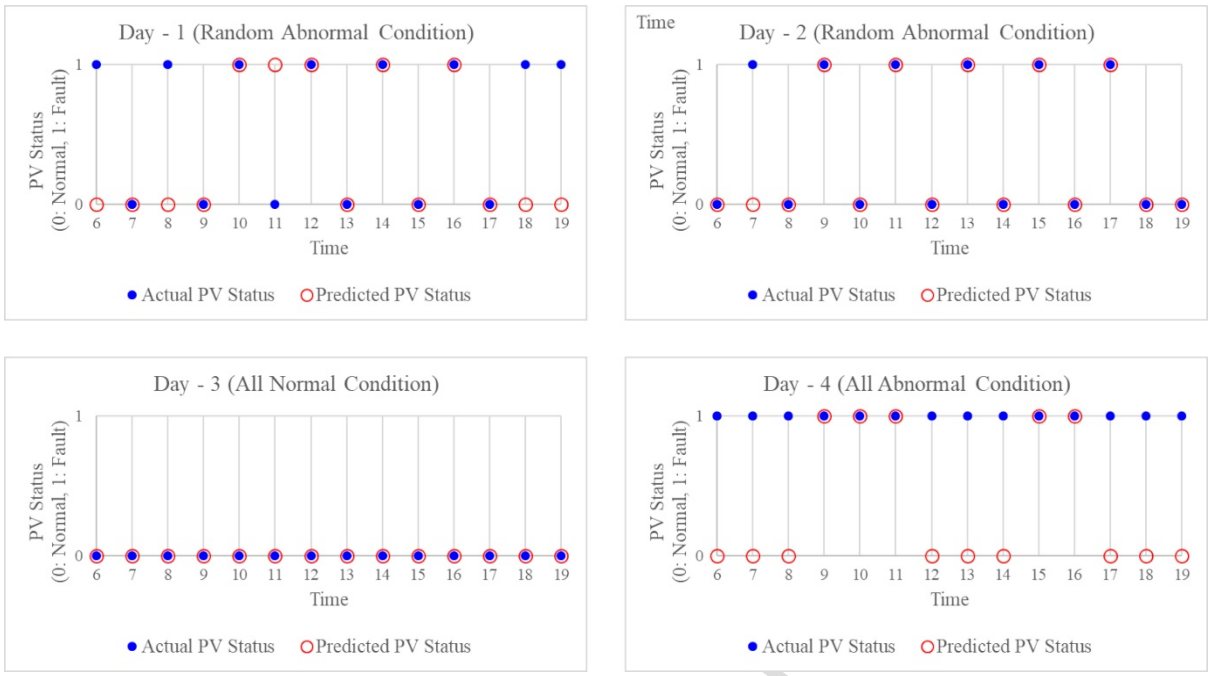


Fig. 4. Comparison between the actual and the predicted PV abnormal condition (Case 2)

Table 3. TPR, TNR, and TA Results for Each Day in Case 2

Day	TPR (%)	TNR (%)	TA (%)
Day-1 (Random Abnormal Condition)	50	83.33	64.29
Day-2 (Random Abnormal Condition)	83.33	100	92.86
Day-3 (All Normal Condition)	-	100	100
Day-4 (All Abnormal Condition)	35.71	-	35.71
Average	**50**	**96.43**	**73.21**

The detection system fails to detect the abnormal condition appearing in Day-1 and Day-4, with 50% and 35.71% of *TPR* respectively, even though the *TPR* of Day-2 is relatively high (83.33%). The average *TPR* is observed as 50%, which means that the detection system can only identify half of the abnormal condition events. However, it can adequately distinguish the normal condition, as indicated by 83.33% of *TNR* in Day-1, 100% of *TNR* in Day-2 and Day-3, thereby having an average *TNR* of 96.43%. In addition, the average *TA* of the prediction system in this case is 73.21%. Therefore, it can be summarized that the traditional SVM model is unable to detect abnormal conditions in a PV system with high accuracy, and thus, further developments are required to improve the detection accuracy.

4.3. Case 3: SVM and Regression

A fundamental idea of PV abnormal condition detection system is to detect the abnormal condition by comparing the measured power generation with the expected power generation. The detection system in this case study is developed by considering the expected PV generation obtained from the regression analysis algorithm, which includes the variables of solar irradiance, ambient temperature, and PV cell temperature. It is mathematically expressed by the following equation:

$$G_{ref} = 10.3427 + 1.7503S - 0.7496T_A + 0.3734T_{PV} \quad (20)$$

The accuracy of this model is presented as 57.40 of MAD. **Fig. 5** shows the comparison between the expected PV generation obtained from this model and the actual measured power.

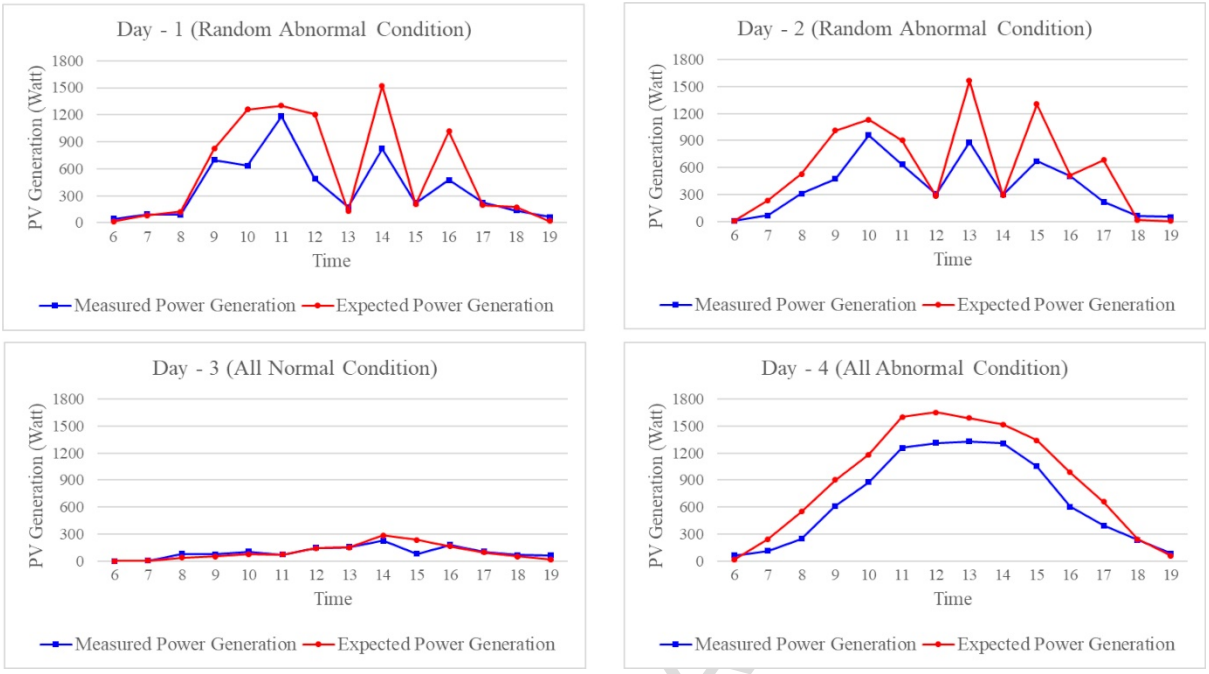


Fig. 5. Comparison between the measured and expected PV generation (Case 3)

It can be observed that the actual PV generation corresponds with the expected PV generation, although there are some periods where the actual PV generation is below the expected one. This condition shows that there exists a possibility of an abnormal condition occurring in the PV system. Therefore, the difference between the measured and the expected PV generation can provide additional information required to detect the abnormal condition.

The results of Days 1-4 based on the regression and SVM models in the detection system are presented in **Fig. 6** and **Table 4**. In this case, *TPR* of Day-4 is improved to 50%, but *TPR* of Day-1 and Day-2 remains the same (50% and 83.33% respectively), in comparison to Case 1. At 57.14% of average *TPR*, this model is only moderately better than the traditional SVM model of Case 1 On the other hand, this model shows an accurate prediction of the normal condition where the average *TNR* is 100%. Hence, it can be observed that the overall accuracy of the prediction system can be improved by including the expected PV generation information as shown by 78.57% of the average *TA*. To summarize, even though this detection system can accurately distinguish the normal condition, its capability to detect the abnormal condition is still relatively low.

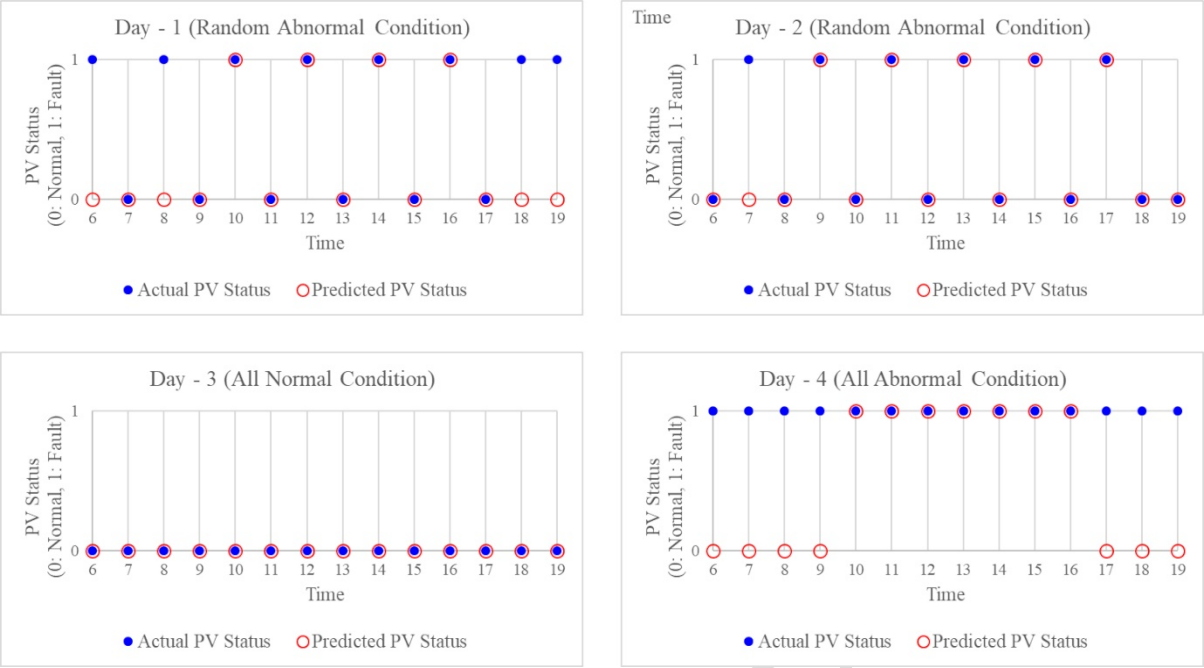


Fig. 6. Comparison between the actual and the predicted PV abnormal condition (Case 3)

Table 4. TPR, TNR, and TA Results for Each Day in Case 3

Day	TPR (%)	TNR (%)	TA (%)
Day-1 (Random Abnormal Condition)	50	100	71.43
Day-2 (Random Abnormal Condition)	83.33	100	92.86
Day-3 (All Normal Condition)	-	100	100
Day-4 (All Abnormal Condition)	50	-	50
Average	**57.14**	**100**	**78.57**

4.4. Case 4: SVM & Regression with improvement strategies

The fourth case study uses SVM and regression models along with the improvement strategies to develop the PV abnormal condition detection system. First, the model is trained by using the refined data based on consideration of the daylight time. Hence, the dataset is refined by considering the data between one hour after sunrise and one hour before sunset. The sunrise and sunset at the PV generation site in the period of this study are estimated by using (8)-(12) and the data shown in **Table 5**. The results show that the sunrise is between 5:15 and 5:36, and the sunset is between 19:40 and 19:56. Therefore, the dataset used to develop the model is the measurement data between 6:00 and 19:00.

Table 5. Location Information of PV System

Site Latitude (*LT*)	33.5°
Site Longitude (*LG*)	126.5°
Local Time Meridian (*LTM*)	135°

Secondly, the additional variables are developed by applying the interactions between the independent variables. The considered variables in this case study include G_{ref}, P, V, I, S, T_A, T_{PV}, PS, PT_A, PT_{PV}, $T_{PV}S$, and $T_{PV}T_A$. Lastly, the configuration of the input variables of the detection system are optimized and selected for both

the regression and SVM models through multi-stage *k*-fold cross-validation algorithm.

The configuration of input variables is constructed by following the hierarchically well-formulated algorithm for the second-order polynomial regression model. All possible predictors (X) are included as the first model, and then the size of the model is sequentially reduced by deleting the highest order predictors until only one variable remains. Amongst all possible models, the best model is selected through *k*-fold cross-validation. The top three regression models with the lowest MAD value obtained from the cross-validation dataset are presented in **Table 6**.

Table 6. Top Three Input Variables Configuration for the Regression Model

Model configuration	MAD
$X = \{S, T_A, S^2, ST_A\}$	46.51
$X = \{S, T_{PV}, ST_{PV}\}$	48.34
$X = \{S, T_{PV}, S^2, ST_{PV}\}$	48.67

The final regression model to estimate the expected PV generation (G_{ref}) under normal operating conditions can be determined from the following equation:

$$G_{ref} = -225.6390 + 2.2635S + 10.0758T_A + 0.0007S^2 - 0.0354ST_A \quad (21)$$

The results of the comparison between the expected PV generation obtained from the model in (21) and the actual measured power is shown in **Fig. 7**. It is observed that the error of the model decreases from 57.40 MAD in Case 2 to 46.51 in this case.

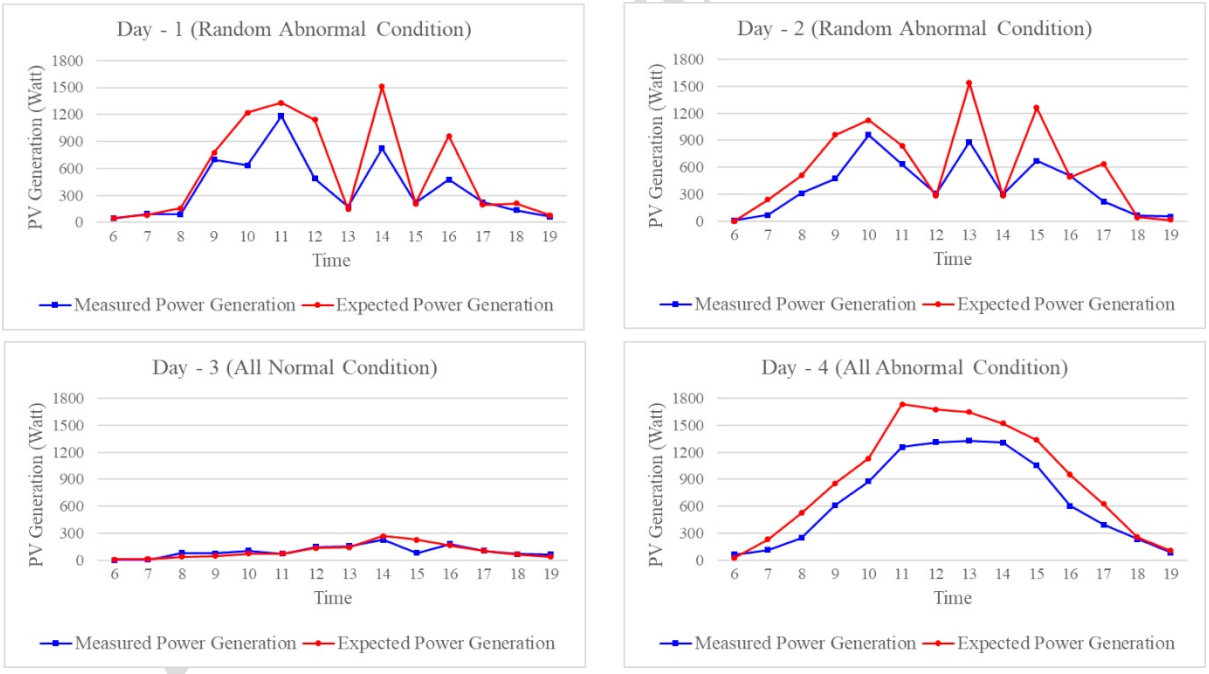


Fig. 7. Comparison between the measured and expected PV generation (Case 4)

As for SVM model, the top three input variables configuration with the lowest MAD on cross-validation dataset are presented in **Table 7**.

Table 7. Top Three Input Variables Configuration for the SVM Model

Model configuration	MAD
$X = \{G_{ref}, P, V, I, S, T_A, T_{PV}, PS, PT_A, PT_{PV}, T_{PV}S, T_{PV}T_A\}$	7.41
$X = \{P, V, I, S, T_A, T_{PV}, PS, PT_A, PT_{PV}, T_{PV}S, T_{PV}T_A\}$	10.71
$X = \{P, V, I, S, T_A, T_{PV}\}$	14.29

The results of Days 1-4 using the regression and SVM models with the improvement strategies in the detection system are presented in **Fig. 8** and **Table 8**. It is observed that this model can accurately predict the abnormal condition as shown by 92.86% of TPR, as well as distinguish the normal condition as shown by 96.43% of TNR. In Day-4, the identification of the abnormal condition is more accurate than the previous models as presented by 85.71% of TPR in this case (Case 1 is 35.71% and Case 2 is 50%). As a result, the overall TA value is also improved from 73.21% (Case-1) and 78.57% (Case-2) to 94.64%.

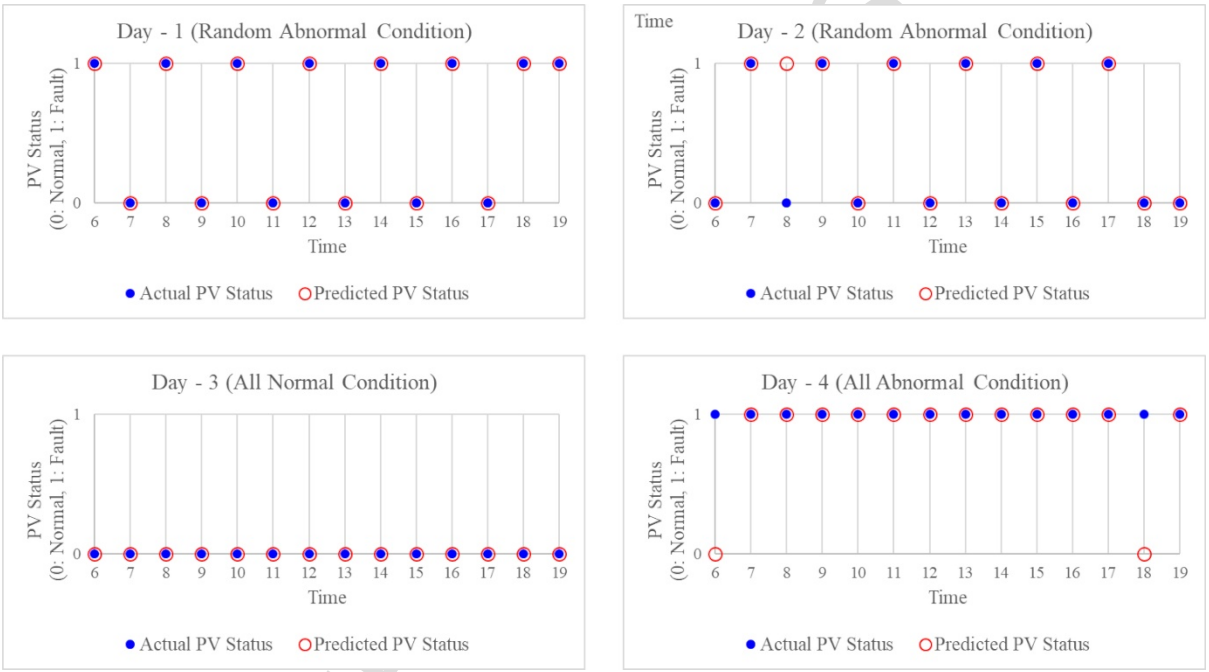


Fig. 8. Comparison between the actual and the predicted PV abnormal condition (Case 4)

Table 8. TPR, TNR, and TA Results for Each Day in Case 4

Day	TPR (%)	TNR (%)	TA (%)
Day-1 (Random Abnormal Condition)	100	100	100
Day-2 (Random Abnormal Condition)	100	87.5	92.86
Day-3 (All Normal Condition)	-	100	100
Day-4 (All Abnormal Condition)	85.71	-	85.71
Average	**92.86**	**96.43**	**94.64**

5. Conclusion

This paper proposes the PV abnormal condition detection system, which can be developed by utilizing minimum amount of data available in the PCS of a small-scale PV generation system. These data include

generated power, voltage, current, solar irradiance, ambient and PV cell temperatures. Therefore, the primary advantage of the proposed detection system is that it does not require installation of any additional devices specifically intended for detecting the abnormal condition. The PV abnormal condition detection system has been developed through a combination of regression and SVM models, wherein the regression model is used to calculate the expected power generation, while the SVM model is used to detect the PV abnormal condition.

This paper presents the limitations of other methodologies such as regression analysis and conventional SVM. It is shown that even though the regression analysis and conventional SVM can detect the abnormal conditions in some cases, but they are not able to accurately classify the conditions of PV system in other cases, especially when the difference between the forecasted and actual PV generation is not very small. Hence, the conventional regression and SVM result in a detection system which tends to detect normal conditions or abnormal conditions only. This paper also shows that the combination of regression and SVM result in a better detection since it does not only take the difference between the forecasted and actual PV generation, but also other parameters which are available in the PCS.

The accuracy of the proposed detection system is improved by considering the daylight time, interactions variables, and the implementation of multi-stage *k*-fold cross-validation optimization technique. The daylight time is used to eliminate the outliers in the data arising from low solar irradiance. The interactions variables are used to include the correlation between the independent variables, while the multi-stage *k*-fold cross-validation optimization technique is applied to select the most suitable model amongst various possible models. The results show that the proposed model can successfully detect the abnormal condition of a PV system, as well as adequately distinguish the normal condition, regardless of the randomness or the uniformity of the PV condition within a certain period of time. Hence, the proposed detection system can be effectively applied to a small-scale PV generation system for determining abnormal conditions with high accuracy at minimum cost.

Acknowledgements

This research was supported by the Ministry of Trade, Industry & Energy (MOTIE), Korea Institute for Advancement of Technology (KIAT) through the Encouragement Program for The Industries of Economic Cooperation Region (No. P0006091).

This work was supported by the Ajou University research fund.

References

[1] A. Triki-Lahiani, A. B.-B. Abdelghani, and I. Slama-Belkhodja, "Fault detection and monitoring systems for photovoltaic installations: A review," *Renewable and Sustainable Energy Reviews*, no. July, p. , 2017.

[2] S. Mekhilef, R. Saidur, and M. Kamalisarvestani, "Effect of dust, humidity and air velocity on efficiency of photovoltaic cells," *Renewable and Sustainable Energy Reviews*, vol. 16, no. 5, pp. 2920–2925, 2012.

[3] M. J. Adinoyi and S. A. M. Said, "Effect of dust accumulation on the power outputs of solar photovoltaic modules," *Renewable Energy*, vol. 60, pp. 633–636, 2013.

[4] S. A. M. Said and H. M. Walwil, "Fundamental studies on dust fouling effects on PV module performance," *Solar Energy*, vol. 107, pp. 328–337, 2014.

[5] B. Nian, Z. Fu, L. Wang, and X. Cao, "Automatic Detection of Defects in Solar Modules: Image Processing in Detecting," *2010 International Conference on Computational Intelligence and Software Engineering*, pp. 1–4, 2010.

[6] Z. Fu *et al.*, "Solar cell crack inspection by image processing," in *Proceedings of 2004 International Conference on the Business of Electronic Product Reliability and Liability (IEEE Cat. No.04EX809)*, 2004, vol. 200030, pp. 77–80.

[7] L. Jiang, J. Su, and X. Li, "Hot Spots Detection of Operating PV Arrays through IR Thermal Image Using Method Based on Curve Fitting of Gray Histogram," *MATEC Web of Conferences*, vol. 61, p. 06017, Jun. 2016.

[8] J. A. Tsanakas, D. Chrysostomou, P. N. Botsaris, and A. Gasteratos, "Fault diagnosis of photovoltaic modules through image processing and Canny edge detection on field thermographic measurements," *International Journal of Sustainable Energy*, vol. 34, no. 6, pp. 351–372, 2015.

[9] Y. Liu, B. Li, and Z. Cheng, "Research on PV module structure based on fault detection," *2010 Chinese Control and Decision Conference*, pp. 3891–3895, 2010.

[10] S. R. Madeti and S. N. Singh, "Online fault detection and the economic analysis of grid-connected photovoltaic systems," *Energy*, vol. 134, pp. 121–135, 2017.

[11] B. K. Kang, S. T. Kim, S. H. Bae, and J. W. Park, "Diagnosis of output power lowering in a PV array by using the kalman-filter algorithm," *IEEE Transactions on Energy Conversion*, vol. 27, no. 4, pp. 885–894, 2012.

[12] Z. M. Omer, A. A. Fardoun, and A. Hussain, "Large scale photovoltaic array fault diagnosis for optimized solar cell parameters extracted by heuristic evolutionary algorithm," *IEEE Power and Energy Society General Meeting*, vol. 2016–Novem, 2016.

[13] S. Spataru, D. Sera, T. Kerekes, and R. Teodorescu, "Diagnostic method for photovoltaic systems based on light I-V measurements," *Solar Energy*, vol. 119, pp. 29–44, 2015.

[14] S. Silvestre, M. A. Da Silva, A. Chouder, D. Guasch, and E. Karatepe, "New procedure for fault detection in grid connected PV systems based on the evaluation of current and voltage indicators," *Energy Conversion and Management*, vol. 86, pp. 241–249, 2014.

[15] N. Gokmen, E. Karatepe, S. Silvestre, B. Celik, and P. Ortega, "An efficient fault diagnosis method for PV systems based on operating voltage-window," *Energy Conversion and Management*, vol. 73, pp. 350–360, 2013.

[16] T. Andrianajaina, E. J. R. Sambatra, C. B. Andrianirina, T. D. Razafimahefa, and N. Heraud, "PV Fault Detection Using the Least Squares Method," no. Epe, pp. 20–22, 2016.

[17] M. Dhimish and V. Holmes, "Fault detection algorithm for grid-connected photovoltaic plants," *Solar Energy*, vol. 137, pp. 236–245, 2016.

[18] E. Garoudja, F. Harrou, Y. Sun, K. Kara, A. Chouder, and S. Silvestre, "Statistical fault detection in photovoltaic systems," *Solar Energy*, vol. 150, pp. 485–499, 2017.

[19] R. Platon, J. Martel, N. Woodruff, and T. Y. Chau, "Online Fault Detection in PV Systems," *IEEE Transactions on Sustainable Energy*, vol. 6, no. 4, pp. 1200–1207, Oct. 2015.

[20] A. Chatterjee, A. Keyhani, and D. Kapoor, "Identification of photovoltaic source models," *IEEE Transactions on Energy Conversion*, vol. 26, no. 3, pp. 883–889, 2011.

[21] K. P. Murphy, *Machine learning : a probabilistic perspective*. MIT Press, 2012.

[22] G. M. Masters, *Renewable and Efficient Electric Power Systems*, 2nd Editio. Wiley-IEEE Press, 2013.

[23] M. Fan, V. Vittal, G. T. Heydt, and R. Ayyanar, "Preprocessing Uncertain Photovoltaic Data," *IEEE Transactions on Sustainable Energy*, vol. 5, no. 1, pp. 351–352, Jan. 2014.

[24] D. R. Cox, "Interaction," *International Statistical Review / Revue Internationale de Statistique*, vol. 52, no. 1, p. 1, Apr. 1984.

[25] L. Gunter, J. Zhu, and S. A. Murphy, "Variable selection for qualitative interactions," *Statistical Methodology*, vol. 8, no. 1, pp. 42–55, 2011.

[26] D. Hsu, "Identifying key variables and interactions in statistical models of building energy consumption using regularization," *Energy*, vol. 83, pp. 144–155, 2015.

[27] J. L. Peixoto, "Hierarchical Variable Selection in Polynomial Regression Models," *The American Statistician*, vol. 41, no. 4, p. 311, Nov. 1987.

[28] J. L. Peixoto, "A Property of Well-Formulated Polynomial Regression Models," *The American Statistician*, vol. 44, no. 1, p. 26, Feb. 1990.

[29] J. A. Nelder, "The Selection of Terms in Response-Surface Models-How Strong is the Weak-Heredity Principle?," *The American Statistician*, vol. 52, no. 4, p. 315, Nov. 1998.

[30] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A Practical Guide to Support Vector Classification," *BJU international*, vol. 101, no. 1, pp. 1396–400, 2008.

- PV abnormal condition detection system is developed
- The model does not require to install any additional measurement devices
- Regression analysis is employed to estimate the ideal PV generation
- Support Vector Machine (SVM) algorithm is used to identify PV abnormal condition
- The proposed detection system is validated by using the actual data