Fault detection and classification in solar photovoltaic system using graph base semi-supervised learning and support vector machine

[1]Vitthal S. Sagde, [2]Nitin J. Phadkule
[1]Student, [2]Assistant Professor
Government college of Engineering, Amravati

Abstract - The behaviour of solar photovoltaic systems is different than conventional power sources as regard to fault detection and classification(FDC). In solar photovoltaic (PV) systems FDC is an ultimate need for increasing safety and reliability in PV systems. A various type of faults may become difficult to detect by conventional protection devices because nonlinear characteristics of PV system, causes safety issues and fire risk in PV system. The machine learning methods such as supervised, unsupervised, and semi-supervised are widely used in various applications. This paper focus on machine learning methods such as graph based semi-supervised learning (GBSSL) and support vector machine (SVM) to mitigate the protection issues. The GBSSL algorithm have been proposed for FDC using measurements, such as PV system voltage, current, irradiance, and temperature and the result obtained is compare with SVM (supervised learning models), which are trained by large amount of labelled data and therefore, have drawbacks: the labelled PV data are difficult or expensive to obtain, the trained model is not easy to update and the model is difficult to visualize. To mitigate these issues, this paper proposes a GBSSL algorithm only using a few labelled data that are normalized by reference values for better visualization. The feature of GBSSL model is to not only detects the fault, but also identifies the possible type of fault in order to get easier system recovery. The model can learn the all the status of the PV systems under various changing weather conditions. The simulation results and their analysis show the effectiveness of fault detection and classification of the proposed GBSSL method and drawbacks of SVM.

Index Terms - Fault detection, GBSSL, SVM, PV arrays, Semi-supervised learning, etc.

I. Introduction

The solar PV arrays do not have any moving parts and therefore, requires low maintenance but still they suffer various faults along the PV systems, batteries, conditioning unit, utility grid connections, and wiring [2] and [3]. The PV array is always energized by sunlight in daytime therefore, it is difficult to shut down PV modules completely during faults. The conventional PV configurations(series–parallel) increase ratings of voltage and current, leading to more risk of dc arcs. This may cause to reduced PV system efficiency, accelerated aging, and system fire hazards [5].

The over-current protection devices (OCPD) and ground-fault detection interrupters (GFDI) conventionally used within PV system for fault protection and detection [6]. However, some faults type in PV system may be non-cleared by OCPD or GFDI due to high value of fault impedances, non-linear output characteristics of PV system, low irradiance, or maximum power point tracker (MPPT) of PV system [3] and [7]. This non-cleared difficulty my cause blind spots in the protection schemes, leading to similar high current or dc arcing and fire hazards reported previously in [5] and [7].

To understand the behaviour of PV system, the operation of PV arrays under normal conditions as well as fault conditions have been well modelled and simulated in [4] and [8]. The fault analysis, reliability, and protection issues have been studied for PV systems in [3]. Use of dc–dc converters on module level to increase energy yield under partial shading conditions and to mitigate the mismatch effects on solar modules, [12] and [13].

II. Literature Review

Several methods of fault detection have been proposed to fill this PV protection gap. For fault detection in solar PV systems, the power loss analysis is proposed in [14] and [15]. In Satellite observed irradiance information has been analysed for fault detection in photovoltaic systems [16]. In [17], based on PV module temperature, current, and voltage using the Kalman filter a fault detection model is proposed. Furthermore, machine learning techniques, such as local outlier factor in [19], artificial neural networks in [20] and statistic outlier detection rules in [18], are analysed to find the health conditions of PV arrays. The soiling effects and their evaluation in PV systems using polynomial regression methods and Bayesian neural network are studied in [21]. The machine learning methods not only detects the faults but also used for fault classification. This can indicate the type of fault and further help maintenance people to expedite the system's recovery from the fault. To detect and classify the types of fault in PV systems, a decision tree model has been proposed in [22].

A clustering-based machine learning method is used for quantifying PV array's effects on utility power grids [23]. However, the drawbacks of these classification models are as follows.

1) A large amount of expensive training data may have required, which could hinder their effectiveness. The training data are difficult or expensive to obtain because they are collected in typical faults under real working conditions.

2) The another challenge lies in the nonlinear operating point of the PV array, which varies widely as the environment condition changes or reduced solar cells life (degradation of solar cell) [22]. This means that a trained model in winter (low irradiance, low temperature) might mistakenly classify fault PV operation as a normal during summer (high irradiance, high temperature). Therefore, the model needs updating the labels continuously over time.

2) The another drawback is lies in the non-linear operating point of the PV system, which varies widely as regards to weather condition changes or reduced solar cells life (degradation of solar cell) [22]. Therefore, a trained model in winter (low temperature, low irradiance) might mistakenly classify fault PV operation as a normal during summer (high irradiance, high temperature). Therefore, the model needs updating the labels continuously over time.

3) The trained models are designed only based on specific operating data conditions on certain PV array installations. This means that the trained model built on a different-case basis that may not be possible to have a general fault detection model. Furthermore, the models are difficult to visualize by PV maintenance people. To overcome the above mentioned issues, this paper proposes a GBSSL for fault detection and classification (FDC) in PV system. The proposed method can detect faults those are non-cleared by conventional OCPD or GFDI, as well as identify the fault classifications.

The GBSSL method only uses easily available measurement such as MPPT voltage and currents, PV irradiation and operating temperature and do not required additional sensing devices or circuits [18]. Therefore, the GBSSL method avoids cost required for labour and hardware installation to conventional PV arrays.

This paper focuses on two types of general faults that may be difficult to detect or clear by conventional OCPD: 1) line to line (LL) fault is an accidental short circuiting between two points in the PV array at different potentials. It may be caused by a short-circuit fault between two different points or a double ground fault in PV fields, and 2) open circuit fault defined as an accidental disconnection at a normal current-carrying conductor, which can be caused by cracking PV cells/modules, loose string connections or blown PV fuses. Note that ground-fault is not discussed in this paper, since essentially it is a special case LL fault involving a ground point. Also, it is relatively easy to detect and clear by GFDI [2].

III. MODELLING OF PV CELL

The single diode PV cell model represented by equivalent circuit which consist of diode, current source, shunt and series resistor as shown in Fig. 1.

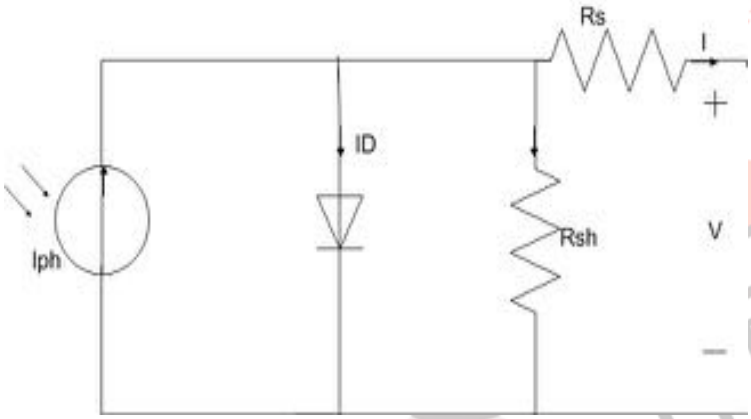


Fig. 1. Single diode PV cell model.

From equivalent circuit we can find the load current deliver by the cell using following equation.

$$I = I_{ph} - I_D - I_{sh} \quad (1)$$

Where I_D is diode current given by shackle diode equation.

$$I_D = I_S \left\{ e^{\left(\frac{V_{PV} + I_{PV} R_S}{A \cdot K \cdot T \cdot N_S} q \right)} - 1 \right\} \quad (2)$$

The photovoltaic generated current is

$$I_{ph} = \frac{G}{G_0} \cdot \{ I_{L0} + C_T \cdot (T - T_0) \} \quad (3)$$

And the current through shunt resistance

$$I_{sh} = \frac{V_{PV} + I_{PV} \cdot R_S}{R_{sh}} \quad (4)$$

The numerical solution of equation (1) is implemented to simulate the solar PV model for types of fault analysis [1].

IV. BASICS OF MACHINE LEARNING METHODS IN SOLAR PV ARRAYS

A. Basics of Machine Learning Methods

A machine learning (ML) is a part of artificial intelligence which includes statistical algorithm that automatically learn and extracts the knowledge from given data set. The data set have of number of data instances and each instance has parts as attribute and label. The attribute is a feature of the data instance which determines its classification and the class label, identifying which class this instance is belongs to. For example, a data set may consist of n number of instances x_i (i = 1 to n). Each data instance x_i is assign as of data set of $\{x_{1,i}, x_{2,i}, \ldots x_{d,i}, y_i\}$, including d number of attributes (d-dimensional data) and a class label y_i . After fault classification implementation, class labels y_i will be identified as "normal condition" or any specific "fault type." The data set is called "labelled data" if its class labels y_i are known. Otherwise, it is unlabelled data. Generally, machine learning methods can be divided into three categories: supervised machine learning, unsupervised machine learning, and semi-supervised learning (SSL)as shown in Fig. 1 [24]. From Fig. 2(a), supervised machine learning uses completely labelled training data with known class labels. However, labelling of data are costly, requiring more human effort (manual verification) and expertise for classification. F Fig. 1(b), unsupervised learning uses only large amount of unlabelled training data with clustering based algorithm. The combination of supervised machine learning and unsupervised machine learning is called semi-supervised learning as shown in Fig. 1(c). The SSL uses both labelled data and unlabelled data for training. This paper focuses on semi-supervised learning since it only requires a few of labelled data and can improve learning accuracy considerably using inexpensive unlabelled data.

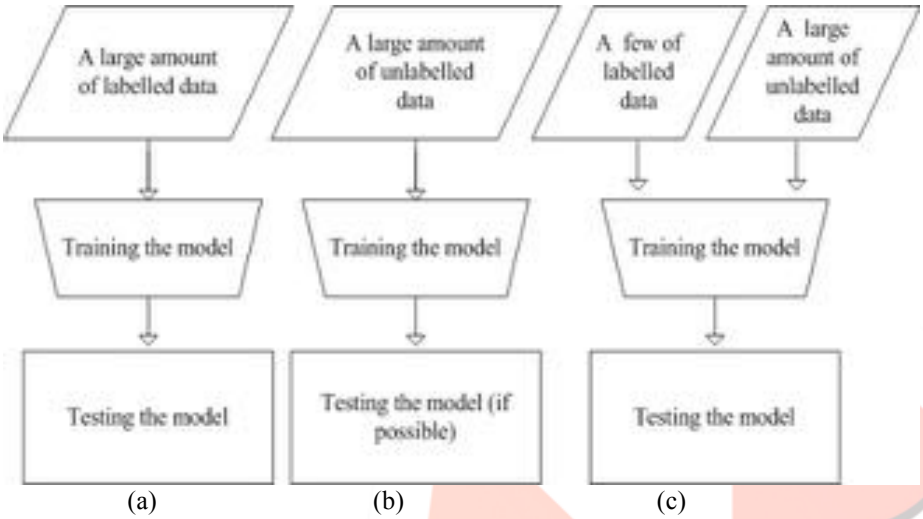


Fig. 2. (a) Supervised learning (b) Unsupervised learning (c) Semi-supervised machine learning.

B. Implementation of ML methods in PV arrays

The schematic of the proposed PV system is shown in Fig. 3, including a typical grid-connected photovoltaic system and the proposed machine learning GBSSL model for FDC.

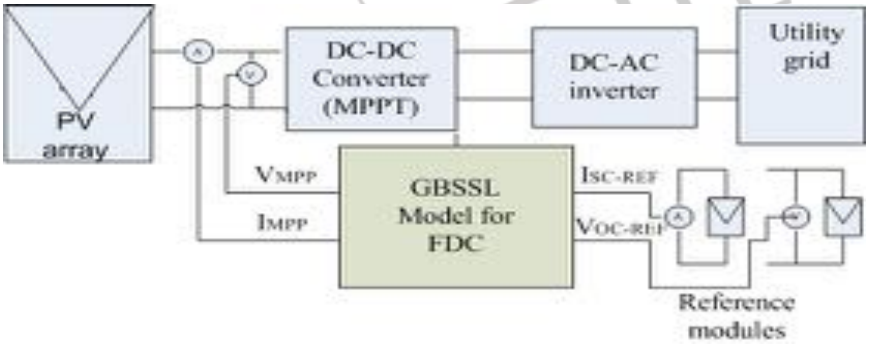


Fig. 3. GBSSL model for Fault Detection and Classification.

1) *Typical PV system:* consists of a PV modules (array), a PV inverter, utility grid, and conventional fault protection devices. The MPPT (DC-DC boost converter) algorithm is utilised to harvest the maximum output power from the PV array. The output power generated from the solar array is feeds into the utility-grid using DC-AC inverter. The photovoltaic inverter includes DC-DC converter along with MPPT and DC-AC inverter. The output curves of PV array exhibit non-linear current versus voltage (I–V) characteristics, for both normal or fault conditions. When the photovoltaic module is faulted, it changes its configuration as a result of change in I–V curves and reduced maximum power points (MPP). And, if the fault is hidden or not cleared properly, it is likely that the photovoltaic inverter will still work, as long as the photovoltaic array can achieve the minimum operating voltage of the PV inverter [3]. Consequently, it is expected that PV system still work with hazardous MPP, and faulted I–V curves. This changing PV MPPs is useful to provide information to the proposed GBSSL model.

2) *Proposed GBSSL model can be integrated into PV inverters.*

The GBSSL model for FDC measures the instantaneous open circuit voltage reference ($V_{OC\text{-}REF}$) and short circuit current ($I_{SC\text{-}REF}$) reference of the small reference module, and receives the PV-array MPP voltage (V_{MPP}) and current(I_{MPP}) at PV's inverter. The reference module is a small power PV module. Alternatively, instead of using reference modules, it is also possible to obtain $I_{SC\text{-}REF}$ and $V_{OC\text{-}REF}$ from simulation models using instantaneous solar cell temperature and solar irradiance monitored by weather stations that are commonly installed in PV fields. As shown in Fig. 3, the GBSSL model only monitors the photovoltaic array (dc side) at MPP under steady state, so it does not rely on any particular power conversion unit. Thus, one advantage of the GBSSL method is the easy integration within a photovoltaic inverter of any circuit topology.

C. Normalized Parameters for better data visualization

The operating MPP points of PV arrays are widely depends on environmental conditions such as solar temperature and irradiance. This causes fluctuations in the PV voltage(V_{MPP}) and PV current(I_{MPP}) widely over a year. Fig. 4, shows the V_{MPP} versus I_{MPP} characteristics for different solar temperature and irradiance. The specific PV array under normal condition and fault conditions are simulated in the MATLAB to obtain this two dimensional plot. In this we consider two types of faults L-L and Open circuit fault. In Fig. 4, even though LL faults tend to have lower V_{MPP} than Normal and Open, it demonstrates the common occurrence of overlapping on MPPs. This means, the faulted photovoltaic array may have similar V_{MPP} and I_{MPP} to a normal array, which creates challenges in identifying faults from normal conditions.

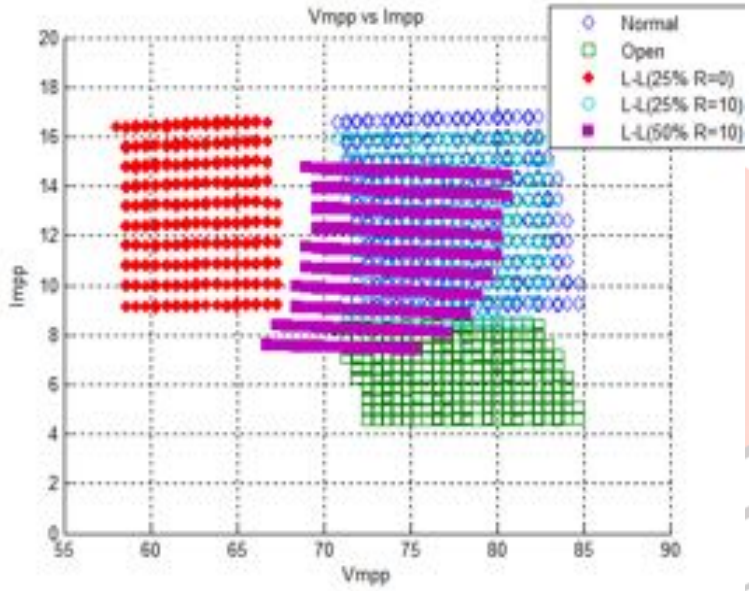


Fig. 4. Overlapping of MPPs of PV system at wide range of solar temperature and irradiance (difficult for FDC).

To mitigate above challenges and better visualize /identify the photovoltaic faults, we have introduced the new parameters (normalised voltages and currents) in Fig. 9 as the data attributes, which shows better data clustering and easier data visualization than the original attributes (V_{MPP} versus I_{MPP} in Fig. 4). Detailed discussion about Fig. 9 will be given in Section VI. Two new normalized parameters are as follows:

1) The normalized PV voltage V_{NORM}. Defined as $V_{MPP}/(N_{MOD} * V_{OC\text{-}REF})$, where N_{MOD} is the number of modules in series per PV string, and $V_{OC\text{-}REF}$ is the open circuit voltage reference.
2) The normalized PV current I_{NORM}. Defined as $I_{MPP}/(N_{STR} * I_{SC\text{-}REF})$, where N_{STR} is the number of strings in parallel in the array, and $I_{SC\text{-}REF}$ is the short circuit-current reference.

The helpfulness of the visualization in Fig. 9 depends on the accuracy and life-cycle degradation of the reference module measurements. It is assumed that the PV module and the reference modules has same life and degrade similarly or at insignificant rates. However, since the PV module the and reference module have same working environments, such as solar irradiance, thermal stress (temperature), or humidity, the reference modules and PV modules are expected to have similar front surface soling or optical degradation over time [25] and [26]. In addition to that, the effect on $V_{OC\text{-}REF}$ and $I_{SC\text{-}REF}$ caused by difference of cell temperature associated with short circuit and open circuit conditions is not much significant [27]. Therefore, it is advantageous to have normalized parameters V_{NORM} and I_{NORM}, which would remain constant, even as the photovoltaic modules degrade uniformly. In this paper, reference modules are chosen to have the identical PV model parameters as other operating modules in the photovoltaic system. However, any partial shadings or mismatch on the reference modules or over the PV modules may cause bad data in V_{NORM} and I_{NORM}, leading to incorrect FDC. Because of this reason, the partial shading on PV modules is not considered in this paper.

V. GBSSL ALGORITHM IN SOLAR PHOTOVOLTAIC ARRAYS

A. Introduction of Graph Data in PV System

The GBSSL consist of PV system data represented in unidirectional graph ($G = (X, E)$) as given from Fig. 5, which has two elements namely vertices X and edges E. The graph (G) is connected only if there is always a path for every vertex to any other vertex. Suppose there are no internal loops or multiple edges in the graph. The vertices X (node) represent the data points which include labelled data and unlabelled data and edges E represent the weight or degree of closeness (similarity) between the points [28].

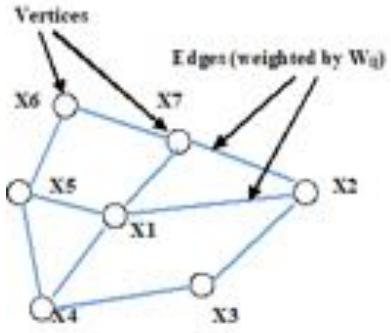
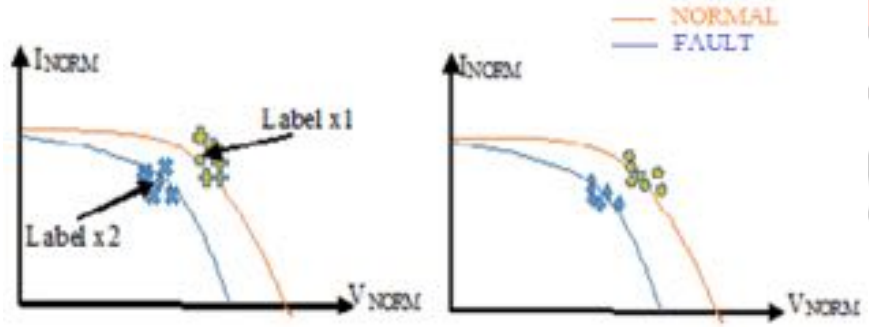


Fig. 5. Graph data consisting of edges and vertices.

The GBSSL algorithm proposed in [28] and [29] will be extended to be applied to solar PV systems. For illustration purpose, in Fig. 6, PV systems normal condition as NORMAL and the fault condition as FAULT are considered in a normalized current versus voltage curve in 2-D, which has possible normalized values of MPPs of a PV system. Note that normalized MPPs vary slightly during the temperature and irradiance changes, since they tend to cluster in their own category (NORMAL or FAULT). Also, the GBSSL assumes that nearby points are falls into the same class label. The aim of the GBSSL algorithm is to let the labelled data, for example, x_1 and x_2 in Fig. 6(a), spread their class label information to unlabelled neighbouring points in order to achieve a global stable state. As a result of GBSSL in Fig. 6(b), unlabelled data will be classified by the initial labels accordingly. This demonstrates the growing label size of the GBSSL algorithm and its self-learning ability in real-time operation. As more data are collected and labelled, the GBSSL for FDC on this specific PV installation is improved. Thus, a few initial labelled data allow a large amount of unlabelled data to be classified.



(a) (b)

Fig. 6. Example of GBSSL for PV array.

B. GBSSL Algorithm for FDC:

A data set as $X = \{x_1, \ldots x_l, x_{l+1}, \ldots, x_n\}$ is represented by a $d \times n$ matrix as (5) shown below, in which each column vector $xi = [x_{1,i}, x_{2,i} \cdots x_{d,i}]^T{}_{1 \times d}$ is data instance in data set of X. A label set $L = \{1, \ldots, c\}$ has the c number of class labels. A label column vector is $Y = [y_1 \cdots y_i \cdots y_n]^T{}_{1 \times n}$, in which each entry $yi \in L$ represents the class label of the instance xi. The first l instances xi ($i \le l$) in X are labelled data with labels yi. The remaining instances x_u ($l + 1 \le u \le n$) in X are unlabelled data which have $y_u = 0$ ($l + 1 \le u \le n$).

$$X = \begin{bmatrix} x_{1,1} & \cdots & x_{1,l} & x_{1,l+1} & \cdots & x_{1,n} \\ x_{2,1} & \cdots & x_{2,l} & x_{2,l+1} & \cdots & x_{2,n} \\ \vdots & & \vdots & \vdots & & \vdots \\ x_{d,1} & \cdots & x_{d,l} & x_{d,l+1} & \cdots & x_{d,n} \end{bmatrix} \quad (5)$$

Specifically, by using the proposed normalized attributes shown in Fig. 9, the matrix X has two rows ($d = 2$, two dimensions for VNORM and INORM) and n number of columns. The two entries for each column vector xi are defined in (6) as an instance of PV measurement at time ti

$$x_i = [Vnorm_i \quad Inorm_i]^T \tag{6}$$

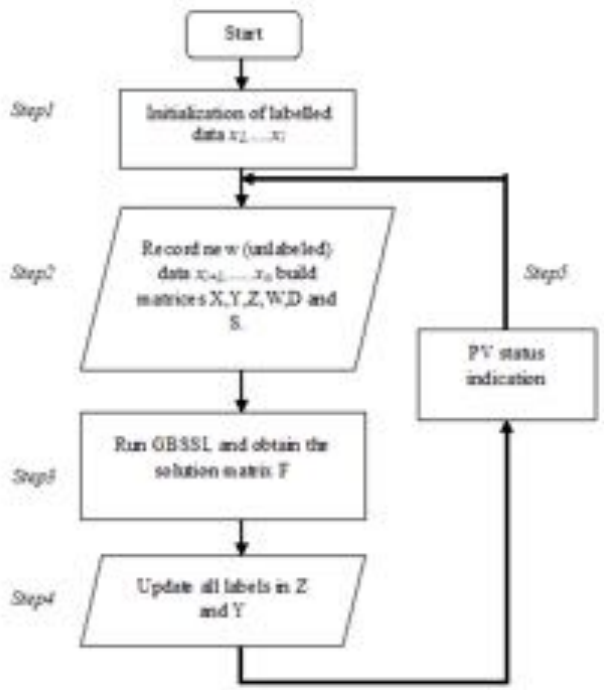


Fig.7. Proposed GBSSL algorithm for FDC in PV system.

The GBSSL algorithm for FDC is demonstrated in five steps as shown in Fig.7, the detailed explanation about GBSSL algorithm are as given as follows:

1) *Step 1:* First, the prior knowledge and information about the PV system is given, including the initial labelled data $\{x_1, \ldots, x_l\}$ and their class labels $\{y_1, \ldots, y_l\}$. For example, the number of class labels is $c = 3$ in our PV application. Thus $y_i = 1$, 2 or 3 represent NORMAL, LL or OPEN conditions.

2) *Step 2:* New instances $\{x_{l+1}, \ldots, x_n\}$ are recorded with unknown labels ($y_u = 0$, $l + 1 \leq u \leq n$). Then, the data set matrix X and the label column Y are constructed accordingly. Using the label column Y, a label matrix Z is established of size $n \times c$ ($c = 3$ in our PV application), whose elements z_{ij} at the i^{th} row and the j^{th} column are defined in

$$Z_{ij} = 1 \; if \; y_i = j \epsilon 1,2,..c \; otherwise \; 0 \tag{7}$$

After that, a weight matrix W (size $n \times n$) is introduced to represent the closeness of its entries W_{ij} as defined in (8). Thus, the matrix W is a pairwise relationship matrix on the data set X with zero diagonal elements

$$W_{ij} = \exp\left(\frac{-(x_i - x_j)^2}{2\sigma^2}\right) \quad if \; i \neq j$$
$$= 0 \qquad\qquad if \; i = j \tag{8}$$

where σ = 1as the bandwidth parameter. For example, in Fig. 5 the edges between the points x_i and x_j in the graph are weighted by the element W_{ij}. Equation (8) implies that if $i \neq j$, W_{ij} is large when x_i and x_j are close to each other, and small when they are far way.

Next, a matrix $s = \frac{W_{ij}}{\sqrt{D_{ii}D_{jj}}}$ (size $n \times n$) is established, where D is a diagonal matrix having D_{ii} as diagonal element equal to the sum of the i^{th} row of the matrix W: $D_{ii} = \sum W_{ij}$. Note that W is the weight matrix is normalized symmetrically.

3) *Step 3:* Once the matrix Z and the matrix S are built, the GBSSL algorithm can be implemented to find the classification matrix F (solution matrix) according to equation (9), where $0 < \alpha < 1$ and $(I - \alpha S)^{-1}$ can be viewed as a graph or diffusion kernel [29]:

$$F = (I - \alpha S)^{-1} \cdot Z \tag{9}$$

The solution matrix is of size $n \times c$(the same size as the matrix Z) with nonnegative entries f_{ij}, where the row vector is $Fi = [f_{i,1}, f_{i,2} \cdots f_{i,c}]$

$$F = \begin{bmatrix} f_{1,1} & f_{1,2} & f_{1,3} & f_{1,4} & \cdots & f_{1,c} \\ f_{2,1} & f_{2,2} & f_{2,3} & f_{2,4} & \cdots & f_{2,c} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ f_{n,1} & f_{n,2} & f_{n,3} & f_{n,4} & \cdots & f_{n,c} \end{bmatrix} \quad (10)$$

4) *Step 4:* The label matrix Z and label column vector Y can be updated based on the classification result the matrix F found previously. The rule of classification is straightforward: the matrix F corresponds to classification criteria on the point set X by labelling each instance xi as a label $y_i = arg\ max\ f_{ij}$. In other words, the class label for each point x_i will be chosen as the class j $(1 \leq j \leq c)$ if $f_{i,j}$ is the maximum entry in the corresponding row vector F_i. For example, there are three classes in the PV application $(c = 3)$, in relation to class labels Class1–NORMAL, Class2–LL, and Class3–OPEN. For instance, x_i, if the corresponding row vector F_i equals to [0.4, 0.6, 1.4], then x_i will be classified into Class3–OPEN, so that $y_i = 3$, since the third entry in F_i has the maximum value.

5) Step 5: FDC results will be indicated continuously. If any faults are identified, fault alarms will be sent out.

VI SUPPORT VECTOR MACHINE (SVM)

SVM is a powerful and universal machine learning method which is widely used in classification and regression. For the nonlinear classification, the main idea is to map the input vector to a high dimensional feature space non-linearly. In this feature space, a linear decision surface is constructed. The high generalization ability of SVM is ensured by the special properties of the decision surface [34]. After the classification has finished in high-dimensional space, the result will be mapped into the original space. Since PCA can only process and improve the classification of dataset. In order to further classify the different working conditions of PV array, a classification model is needed. SVM adopts the Structure Risk Minimization to achieve the optimal generalization ability and avoid over-fitting by balancing the error of training set and maximizing the classification interval. It can solve practical problems such as small samples and non-linearity well. Therefore, SVM is applied in this paper to achieve this purpose. In this paper, the classification model will be trained by the toolbox LIBSVM which compiled in MATLAB [35].

VI. SIMULATION RESULTS

A. Simulated PV Systems

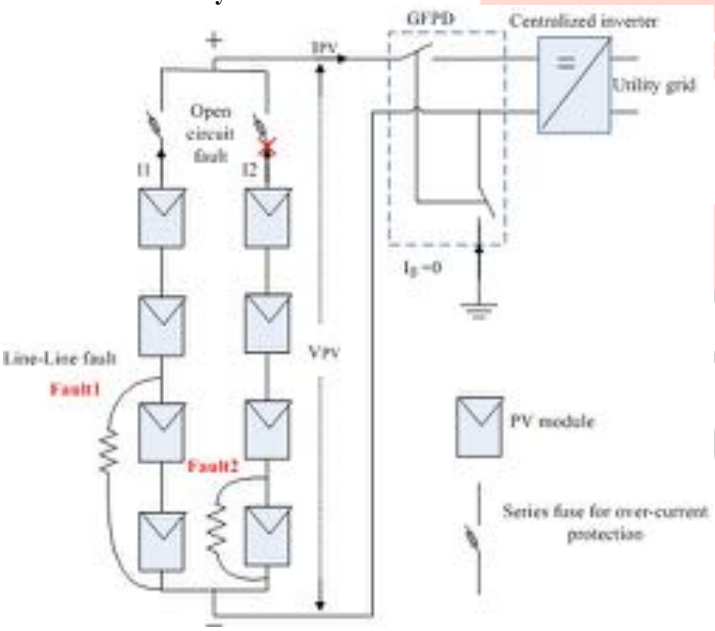


Fig. 8. LL faults and open circuit fault in simulated PV system.

In this paper the PV system of KW, 4 × 2 PV modules (monocrystalline silicon), is builds in MATLAB/Simulink using the widely used single-diode model [10] for each solar cell of module/panel that is capable of analysing faults among the modules. The schematic diagram is shown in Fig. 8. The number of series modules per string is $N_{MOD} = 4$, and the number of parallel strings is $N_{STR} = 2$. The main parameters of each PV module at standard test conditions (STC) are as follows: the maximum power $P_{MPP} = 250$ W, the open circuit voltage $V_{OC} = 36$ V, the maximum power voltage $V_{MPP} = 29$ V, the short-circuit current $I_{SC} = 9.25$A, the maximum power current $I_{MPP} = 8.62$ A.

A. Faults in photovoltaic Systems

As seen from Fig. 8, there are two categories of faults in the photovoltaic systems: line to line (LL) faults and open-circuit faults (Open). The normalized value of MPPs under a different normal and fault conditions have been plotted in Fig. 9.

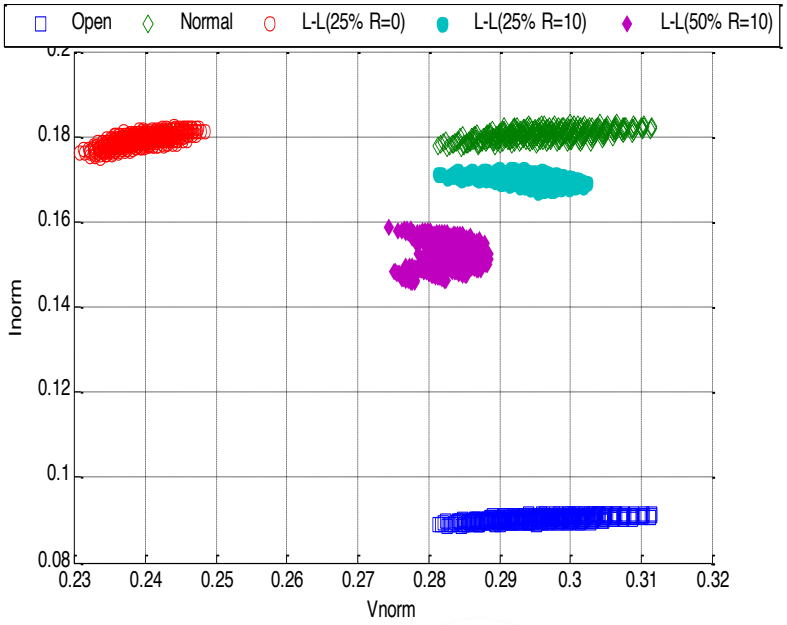


Fig. 9. V_{NORM} versus I_{NORM} of PV array for wide range of temperature and irradiance (Better for FDC).

1) *Normal condition:* Under various environmental conditions of changing solar temperature and irradiance, the normalized value of MPPs have the following operating range: V_{NORM} (0.28, 0.0.313) and $I_{NORM} \in$ (0.175, 0.185).
2) *Line–line fault:* A variety of LL faults without or with fault resistance (R_f = 0 Ω or 10 Ω) are studied. The fault resistance for typical line–line PV faults is considered as 0 Ω at solid faults, or as 10 Ω which may be caused by poor connections or dc arcs [30]. The fault between the fault point "Fault1" and negative conductor (Fault1-Neg) in Fig. 8 is defined as "50% location mismatch (LL 50%)," since it involves two-module mismatch between the fault points in the faulted string of PV system (normally 4 modules per string). Similarly, the Fault2-Neg fault is defined as "25% location mismatch (LL 25%)." Compared with NORMAL, I_{NORM} under LL is slightly reduced to a range of (0.148, 0.185), but V_{NORM} is significantly decreased, which lies in a wide range of (0.23, 0.301). The reason is that the MPPT tends to reduce V_{MPP} to reach the suboptimal MPP under LL faults, leading to a reduced V_{NORM}. Meanwhile, I_{NORM} may be reduced as well since the fault has a negative effect on I_{MPP}.
3) *Open-circuit fault:* The open circuit fault on string (OPEN) are included in Fig. 9. Notice that V_{NORM} of OPEN faults and NORMAL conditions are identical. But I_{NORM} is reduced proportionally by the number of lost strings, due to the parallel connection of PV strings.

B. Simulation of GBSSL for FDC

1) Initial Labels: As shown in Fig. 9, the normal condition is labelled as "Normal"; the LL faults include "L-L 25% R_f = 0," "L-L 25% R_f = 10," and "L-L 50% R_f = 10"; the open-circuit faults include "Open". The number of initial labels is only 10 (approximately 2.85% of 350 instances in each specific fault) for each of the previously mentioned PV conditions. The initial label data are PV MPPs under limited environmental conditions: solar irradiance ranging from 550 to 1000 W/m^2 and ambient temperature fixed at 25 °C. The advantage of GBSSL is to use of a few initial labels (labelled data set) to classify a large amount of new photovoltaic operating data, which are known as test data as given below.

2) Test Data: In order to test the robustness of the GBSSL algorithm for PV systems FDC, two types of photovoltaic data are used that are not included in the initial labels, namely new weather conditions and new type of fault conditions.

- First, we utilize the photovoltaic data under wide-ranging weather conditions that the GBSSL has hardly seen before. To cover all possible working conditions, the PV simulation covers the combinations of module-plane solar irradiance (GT) extensively varying from 550 to 1000 W/m^2 with step change of 50 W/m^2, and ambient temperature (T_{amb}) changing from *25* to 59 °C by 1 °C. It is useful to test the robustness of the proposed GBSSL [32].
- Second, new fault conditions (could be more challenging) are added into the test data, including "LL 75% R_f = 0," "LL 75% R_f = 5," "LL 75% R_f = 7," and "LL 75% R_f = 10". These new types of faults have not included in the initial lebelled data so they are "hidden" for the GBSSL algorithm.

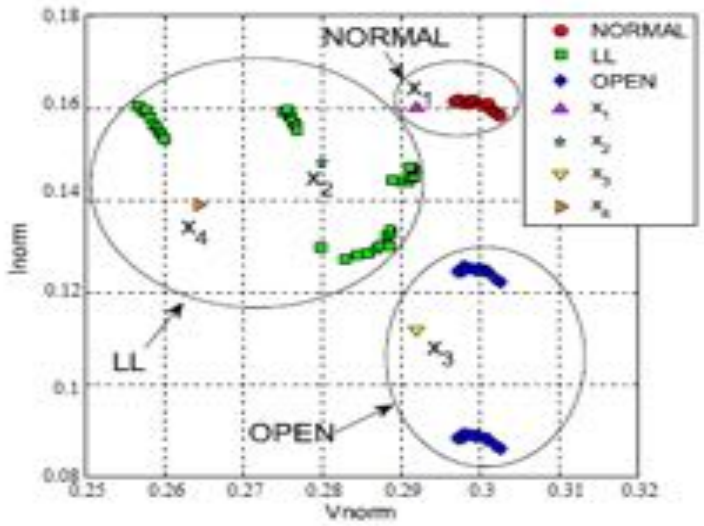


Fig. 10. Example of GBSSL model: new data points x_1, x_2, x_3, and x_4 will be classified as Normal, LL, Open and LL respectively.

3) How GBSSL Works at PV Test Data: Test data are fed into the GBSSL algorithm one by one as real-time operation. Therefore, only one new data instance at every GBSSL iteration, which implies that $n = l + 1$ in (5). The initial labels for the GBSSL are illustrated in the normalized parameters shown in Fig. 10. The FDC procedure is explained as the following example. At time $t = t_1$, there is a new instance (x_1) coming into the proposed model, which indicates a sudden change of PV operating point. Based on the given initial labels as well as the similarity between x_1 and various classes, it is clear to see that x_1 is closer to "NORMAL" than any other clusters. Therefore, GBSSL algorithm will classify x_1 as class "NORMAL" and store it as a new label. Thus, the size of labelled data can increase as more data are collected and identified demonstrating the ability to self-learning by the GBSSL. Similarly, new data points x_2 at $t = t_2$ and x_3 at $t = t_3$ will be correctly classified as "LL" class and "OPEN" class, respectively. As previously discussed, a gradual and uniform PV aging over years is likely to be classified as "NORMAL," since it is a slow deviation from "NORMAL" working MPPs. Therefore, the GBSSL algorithm may avoid unwanted tripping on such events. In the previous discussion, we considered that the GBSSL algorithm always utilizes the PV systems MPP points under both normal and fault conditions, which can be done by MPPT of the photovoltaic inverter. It is observed that the GBSSL can also successfully identify the fault even when the photovoltaic array deviates from its true MPP to a "sub-MPP." The new data point x_4 at $t = t_4$ is closer to the LL than other clusters. Therefore, GBSSL will correctly classify x_4 as "LL" class.

D. Simulation of SVM for FDC

The daily working states simulated by the simulation system are include: (1) normal, (2) one of the strings open circuit (Open1), (3) one module in a string short circuit (Short1). We mix data from four states and define labels for each state so that they can be trained and classified by SVM. In addition, a certain amount of data is randomly selected from each state as the test data, and these test data are mixed to form the data matrix. And the rest is the training data, which consists of the training data matrix. The simulation irradiance is 550-1000 W/m2.

E. Simulation Results

To test the effectiveness of the GBSSL and avoid the effects from data occurring sequences, the simulation focuses on the worst-case scenario of FDC, the GBSSL only uses initial labels for FDC and does not update the calculated labelled data (the Step 4 in Fig. 7 is neglected). Thus, any particular test data can be the first input data that the GBSSL learns in the PV system. The detection accuracy and classification accuracy are defined in (12) and (13). The simulation results of FDC for a variety of PV conditions are summarized in Table I

$$\text{Detection Accuracy} = \frac{\text{correct \# of detection}}{\text{total \# of instances}} \tag{11}$$

$$\text{Classification Accuracy} = \frac{\text{correct \# of classification}}{\text{total \# of instances}} \tag{12}$$

By using only 2.85% (= 60/2100) of the total data as initial labels, the proposed GBSSL model can detect and classify PV data under various classes, including NORMAL, LL, and OPEN. Specifically, the GBSSL is able to identify LL faults successfully that are undetectable by conventional OCPD, such as "LL 25% Rf = 0," "LL 25 % Rf = 10," and "LL 50% Rf = 10." In addition, that the proposed model can work with new types of faults, even if they are not listed in the initial label set. Forb new "Open" PV faults, the proposed model can identify them successfully, since they are more similar to previously "Open" in the initial labels similar V_{NORM} but reduced I_{NORM}. However, for new "LL" faults, the proposed model may misclassify them as

NORMAL, because they are overlapping with NORMAL in the initial labels. The more overlapping between the LL fault and NORMAL, the more likely the LL is misclassified as NORMAL, their detection accuracy and classification accuracy are the same.

TABLE I
Simulation Result of Fault Detection and Classification.

Fault Category	Detailed PV Conditions	GBSSL Results		
		Initial labels of total data	Detection Accuracy	Classification Accuracy
NORMAL	Normal	10 labels of 350	100%	100%
LL	L-L 25% $R_f = 0$ L-L 25% $R_f = 10$ L-L 50% $R_f = 10$	10 labels of 350 of each type	100%	100%
OPEN	Open	10 labels of 350	100%	100%

F. Discussion

The success of fault detection and classification relies on the distinguishable data points in the 2-D normalized parameters (i.e., V_{NORM} versus I_{NORM}). The simulation results show that photovoltaic array's size, fault type, and fault resistance R_f can make the operating point difficult or even impossible to identify. The detection and classification errors are caused by overlapping between NORMAL and LL. This brings difficulty to the GBSSL model to identify the new PV data near their border. For this reason, the LL fault may be misclassified as NORMAL by the GBSSL, leading to a protection challenge in the PV array, especially when the fault resistance R_f becomes significant. To eliminate this "blind spot" one can achieve increasing the number of current or voltage measurement in the PV array, such as using additional current sensors in strings [19]. Another limitation is that the GBSSL classification may become vulnerable to bad data, which could be caused by measurement noise, communication error or poor electrical connections. If the GBSSL uses a bad data, it may have incorrect FDC result (i.e., wrong class label) and cause false alarms. During the GBSSL's self-learning process, the incorrect class labels may be propagated to the next data point. To mitigate this problem, data pre-treatment, outlier detection rules, and digital filter have been studied to eliminate the unwanted data [17], [19] and [33]. These types of technical challenges are typical of many machine learning methods, particularly for GBSSL. According to the simulation results and analysis SVM can successfully classify the faults type. The FDC using SVM requires large amount of labelled data, therefore the cost required for SVM classification is higher than GBSSL algorithm.

VII. Conclusion

The solar photovoltaic system exhibits nonlinear current versus voltage characteristics, so it is difficult to identify and cleared the fault during the low irradiance weather condition by the conventional OCPD as this fault have small fault current. Thus, the faults can remain unseen in the photovoltaic system [3], resulting in possible hazards associated with PV system (suboptimal performance, dc arcing, fire hazard, etc.). To identify these hidden faults, GBSSL has been proposed for FDC in solar PV system. It increases the PV's safety and reliability by detecting these faults (unnoticeable by OCPD). In addition, the GBSSL method have ability to identify the specific type of fault so that PV users can expedite the system restoration procedure. To better visualization of the PV data under both normal and fault conditions, the paper first proposes new attributes by using the normalized voltage and current. In comparison to other type of machine learning methods, the GBSSL algorithm only requires a few number of the expensive labelled data (~3% among all data set), while making use of unlabelled data which requires no cost. The GBSSL algorithm and SVM is analysed and explained in details by using simulation results. Furthermore, the ability to self-learning of GBSSL algorithm is proved as the label set can be updated over time during weather changes or PV arrays degrade. In addition to that, the GBSSL and SVM method does not depend on any particular PV inverter topologies and only uses easily available measurements from the existing photovoltaic systems, such as PV-array voltage, array current, operating temperature, and irradiance requiring no additional hardware installations.

References

[1] Md Tohfael Ahmed, T. Goncalves and M. Tlemcani, "Single Diode Model Parameter Analysis of Photovoltaic Cell", 5th International conf. on Renewable Energy Research and Applications, pp. 396-400, Nov 2016.

[2] Ye Zhao, R. Ball, J. Mosesian, J. de Palma and B. Lehman, "Graph-based semi-supervised learning for fault detection and classification in solar photovoltaic arrays", *IEEE Trans. Power Electron*., vol.30, no.5, pp. 2848-2858, may 2015.

[3] Y. Zhao, J. de Palma, J. Mosesian, R. Lyons, and B. Lehman, "Line-line fault analysis and protection challenges in solar photovoltaic arrays," *IEEE Trans. Ind. Electron.*, vol. 60, no. 9, pp. 3784–3795, Sep. 2013.

[4] S. Harb and R. S. Balog, "Reliability of candidate photovoltaic module integrated- inverter (PV-MII) topologies—A usage model approach," *IEEE Trans. Power Electron.*, vol. 28, no. 6, pp. 3019–3027, Jun. 2013.

[5] B. Brooks, "The bakers field fire—A lesson in ground-fault protection," *Solar Pro Mag.*, pp. 62–70, 2011.

[6] *Article 690 - Solar Photovoltaic Systems*, NFPA70, National Electrical Code, 2014.

[7] J. Flicker and J. Johnson, "Analysis of fuses for blind spot ground fault detection in photovoltaic power systems," Sandia National Laboratories Albuquerque, NM, USA, Tech. Rep.

[8] S. Jing Jun and L. Kay-Soon, "Photovoltaic model identification using particle swarm optimization with inverse barrier constraint," *IEEE Trans. Power Electron.*, vol. 27, no. 9, pp. 3975–3983, Sep. 2012.

[9] E. Ribeiro, A. J. M. Cardoso, and C. Boccaletti, "Fault-tolerant strategy for a photovoltaic DC–DC converter," *IEEE Trans. Power Electron.*, vol. 28, no. 6, pp. 3008–3018, Jun. 2013.

[10] E. Koutroulis and F. Blaabjerg, "Design optimization of transformer less grid-connected PV inverters including reliability," *IEEE Trans. Power Electron.*, vol. 28, no. 1, pp. 325–335, Jan. 2013.

[11] C. Shih-Ming, L. Tsorng-Juu, Y. Lung-Sheng, and C. Jiann-Fuh, "A safety enhanced, high step-up DC-DC converter for AC photovoltaic module application," *IEEE Trans. Power Electron.*, vol. 27, no. 4, pp. 1809–1817, Apr. 2012.

[12] S. M. MacAlpine, R. W. Erickson, and M. J. Brandemuehl, "Characterization of power optimizer potential to increase energy capture in photovoltaic systems operating under non-uniform conditions," *IEEE Trans. Power Electron.*, vol. 28, no. 6, pp. 2936–2945, Jun. 2013.

[13] C. Olalla, D. Clement, M. Rodriguez, and D.Maksimovic, "Architectures and control of submodule integrated DC-DC converters for photovoltaic applications," *IEEE Trans. Power Electron.*, vol. 28, no. 6, pp. 2980–2997, Jun. 2013.

[14] A. Chouder and S. Silvestre, "Automatic supervision and fault detection of PV systems based on power losses analysis," *Energy Convers. Manag.*, vol. 51, pp. 1929–1937, 2010.

[15] S. Silvestre, A. Chouder, and E. Karatepe, "Automatic fault detection in grid connected PV systems," *Sol. Energy*, vol. 94, pp. 119–127, 2013.

[16] A. Drews, A. C. de Keizer, H. G. Beyer, E. Lorenz, J. Betcke, W. G. J. H. M. van Sark, W. Heydenreich, E. Wiemken, S. Stettler, P. Toggweiler, S. Bofinger, M. Schneider, G. Heilscher, and D. Heinemann, "Monitoring and remote failure detection of grid-connected PV systems based on satellite observations," *Sol. Energy*, vol. 81, pp. 548–564, 2007.

[17] K. Byung-Kwan, K. Seung-Tak, B. Sun-Ho, and P. Jung-Wook, "Diagnosis of output power lowering in a PV array by using the Kalman-filter algorithm," *IEEE Trans. Energy Convers.*, vol. 27, no. 4, pp. 885–894, Dec. 2012.

[18] Y. Zhao, B. Lehman, R. Ball, J. Mosesian, and J.-F. de Palma, "Outlier detection rules for fault detection in solar photovoltaic arrays," in *Proc. 28th IEEE Appl. Power Electron. Conf.*, 2013, pp. 2913–2920.

[19] Y. Zhao, F.Balboni,T.Arnaud, J. Mosesian,R.Ball, andB.Lehman, "Fault experiments in a commercial-scale PV laboratory and fault detection using local outlier factor," in *Proc. IEEE 40th Photovoltaic Spec. Conf.*, 2014, pp. 3398–3403.

[20] D. Riley and J. Johnson, "Photovoltaic prognostics and heath management using learning algorithms," in *Proc. IEEE 38th Photovoltaic Spec. Conf.*, 2012, pp. 1535–1539.

[21] A. Massi Pavan, A. Mellit, D. De Pieri, and S. A. Kalogirou, "A comparison between BNN and regression polynomial methods for the evaluation of the effect of soiling in large scale photovoltaic plants," *Appl. Energy*, vol. 108, pp. 392–401, 2013.

[22] Y. Zhao, L. Yang, B. Lehman, J. F. de Palma, J. Mosesian, and R. Lyons, "Decision tree-based fault detection and classification in solar photovoltaic arrays," in *Proc. 27th IEEE Annu. Appl. Power Electron. Conf.*, 2012, pp. 93–99.

[23] W. A. Omran, M. Kazerani, and M. M. A. Salama, "A clustering-based method for quantifying the effects of large on-grid PV systems," *IEEE Trans. Power Del.*, vol. 25, no. 4, pp. 2617–2625, Oct. 2010.

[24] X. Zhu and A. B. Goldberg, *Introduction to Semi-Supervised Learning*. San Rafael, CA, USA: Morgan & Claypool, 2009.

[25] J. A. D. Cueto and S. R. Rummel, (2010). "Degradation of photovoltaic modules under high voltage stress in the field," *Proc. SPIE*, vol. 7773, pp. 77730J-1–77730J-11.

[26] E. L. Meyer and E. E. van Dyk, "Assessing the reliability and degradation of photovoltaic module performance parameters," *IEEE Trans. Rel.*, vol. 53, no. 1, pp. 83–92, Mar. 2004.

[27] R. Bharti, J. Kuitche, and M. G. Tamizh Mani, "Nominal operating cell temperature (NOCT): Effects of module size, loading and solar spectrum," in *Proc. 34th IEEE Photovoltaic Spec. Conf.*, 2009, pp. 001657–001662.

[28] D. Zhou and B. Sch¨olkopf, "A regularization framework for learning from graph data," presented at the Workshop Statistical Relational Learning at Int. Conf. Mach. Learning, Banff, Canada, 2004.

[29] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Sch¨olkopf, "Learning with local and global consistency," in *Proc. 22nd Int. Conf. Advances in Neural Information Processing Systems 16, (2003). [Online]. Available at: http://papers.nips.cc/paper/2506-learning-with-local-andglobal- consistency*

[30] C. Strobl and P. Meckler, "Arc faults in photovoltaic systems," in *Proc. 56th IEEE Holm Conf. Electr. Contacts*, 2010, pp. 1–7.

[31] R. H. Dieck, *Measurement Uncertainty: Methods and Applications*, 4th ed. Research Triangle Park, NC, USA: ISA, 2007.

[32] E. Skoplaki and J. A. Palyvos, "Operating temperature of photovoltaic modules: A survey of pertinent correlations," *Renew. Energy*, vol. 34, pp. 23–29, 2009.

[33] R. K. Pearson, *Mining Imperfect Data: Dealing with Contamination and Incomplete Records.* Philadelphia, PA, USA: SIAM, 2005.

[34] A. M. Tayade, N. J. Phadkule, "Transmission line faults detection and classification by using support vector machine technique", *JETIR*, May-2019.

[35] L. C. Chen, "Fault diagnosis and classification for photovoltaic arrays based on principal components analysis and support vector machine", *IOP Conf. Ser.: Earth Environ. Sci.,*2018.