Design of Hybrid Artificial Bee Colony Algorithm and Semi-Supervised Extreme Learning Machine for PV Fault Diagnoses by Considering Dust Impact

Jun-Ming Huang, Rong-Jong Wai, *Senior Member, IEEE*, and Geng-Jie Yang

***Abstract*—Photovoltaic (PV) systems operating in the outdoor environment are vulnerable to various factors, especially dust impact. Abnormal operations lead to massive power losses, and severe faults as short circuit may cause safety problems and fire hazards. Therefore, monitoring the operation status of PV systems for timely troubleshooting potential failure and effective cleaning scheme are the focus of current research works. In this study, *I–V* characteristics of PV strings under various fault states are analyzed, especially soiling condition. Because labeled data for PV systems with specific faults are challenging to record, especially in the large-scale ones, a novel algorithm combining artificial bee colony algorithm and semi-supervised extreme learning machine is proposed to handle this problem. The proposed algorithm can diagnose PV faults using a small amount of simulated labeled data and historical unlabeled data, which greatly reduces labor cost and time-consuming. Moreover, the monitoring of dust accumulation can warn power plant owners to clean PV modules in time and increase the power generation benefits. PV systems of 3.51 and 3.9 kWp are used to verify the proposed diagnosis method. Both numerical simulations and experimental results show the accuracy and reliability of the proposed PV diagnostic technology.**

***Index Terms*—Artificial bee colony (ABC) algorithm, dust impact, fault diagnosis, photovoltaic (PV), semi-supervised extreme learning. machine.**

I. Introduction

PHOTOVOLTAIC (PV) power generated by solar energy is one of the most promising renewable technologies. Due to the virtues of being free from geographical constraints, flexible in scale application and pollution-free, PV generation capacity has increased exponentially in the past decade. However, complex outdoor environment leads to PV systems prone to suffer electrical faults as short-circuit or ground faults, internal faults of PV cells as abnormal aging or hot spots, and partial shading faults caused by external objects [1], [2]. The dust deposition is an inevitable problem for outdoor PV systems. In real PV systems, dust adhere on the surface of PV modules would cause two consequents. First, the direct impact is the losses of the PV power generation. Due to different environments around the world, dust deposition reduces the power generation of a PV system by 20, even to 80% decay [3]–[6]. Moreover, the different degree of dust deposition on the surface of each PV module makes the corresponding degree of mismatch failure in a PV system. On the other hand, the influence of dust on the single PV module is mainly reflected in the different degree of dust accumulation on the upper and lower edges of the inclined modules. Because PV modules needs to be tilted for receiving more irradiance, the dust is easy to accumulate at the lower area of PV modules especially after the raining day, which results in the partial shading effect. Therefore, long-term uncleaned PV modules may cause hot spots and irreversible damage in the lower edge of PV modules. Therefore, monitoring the status of dust deposition and diagnosing PV faults under the soiling condition is an essential task for improving the reliability of PV systems.

Severe dust deposition, as well as other failures mentioned above, need manpower maintenance to reduce the harm and loss, and they can be regarded as the faulty state of PV systems. Due to the sensitivity of PV modules to irradiance, the variety of PV faults and the influence of dust deposition, traditional protective devices are hard to detect PV faults. Although PV systems can continue to operate in the faulty state, the power generation efficiency is deeply affected, and long-term failure causes irreversible damage to the module even results in fire disaster [7]. To monitor the operation status of PV systems and detect potential faults, many advanced diagnostic methods for PV faults have been proposed in recent years. As for diagnostic features among them, *I–V* curves can reflect the operation status of PV systems preferably. According to output *I–V* curves, various faults of PV systems at the dc side can get a more accurate classification [8]–[10]. With the development of online *I–V* tracker in smart inverters recently, *I–V* curves of each PV string can be obtained without interrupting the power flow to the load in a PV system [11], [12], which can massively reduce the cost and is of great significance for PV fault diagnoses.

As for PV fault detection and classification methods, many advanced techniques have been developed for various PV faults

Manuscript received October 6, 2019; accepted November 27, 2019. Date of publication November 27, 2019; date of current version March 13, 2020. This work was financially supported in part by the Ministry of Science and Technology of Taiwan under Grant MOST 108-2221-E-011-080-MY3. Recommended for publication by Associate Editor K.-B. Lee. *(Corresponding author: Rong-Jong Wai.)*

J.-M. Huang is with the College of Electrical Engineering and Automation, Fuzhou University, Fuzhou 350108, China, and also with the Department of Electronic and Computer Engineering, National Taiwan University of Science and Technology, Taipei 106, Taiwan (e-mail: 994809688@qq.com).

R.-J. Wai is with the Department of Electronic and Computer Engineering, National Taiwan University of Science and Technology, Taipei 106, Taiwan (e-mail: rjwai@mail.ntust.edu.tw).

G.-J. Yang is with the College of Electrical Engineering and Automation, Fuzhou University, Fuzhou 350108, China (e-mail: ygj23802@126.com).

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

Digital Object Identifier 10.1109/TPEL.2019.2956812

in recent years. The spread spectrum time-domain reflectometry in [13] was applied to determine the impedance variation of a PV system for the line-ground fault detection. However, this signal-injection-based method needs a specific external signal function generator, and its accuracy is easily affected by different configurations of PV systems. Dhimish and Badran [14] proposed a current limiter circuit to eliminate of various types of hot spots by decreasing the temperature level of faulty PV modules, and improved the power generation under partial shading conditions. Based on the impact of the maximum-power-point-tracking (MPPT) technology, Pillai and Rajasekar [15] investigated a statistical-based method to detect line-line and line-ground faults via specific detection rules. Statistical tools, such as the principal component analysis in [16] and the wavelet packets in [17], were also utilized to detect failures in PV systems. Moreover, a mathematical analysis technique based on *t*-test in [18] was studied for identifying faulty conditions in both dc and ac sides of PV power generation systems. Even though these statistical and mathematical-based approaches are validated to be effective and computation efficient, they commonly rely on manual threshold values through rigorous analysis of faulty systems, which may limit the performance and applications.

In addition, with the development of artificial intelligence, machine learning technologies have been widely used, which can be generally divided into three categories including supervised learning, unsupervised learning, and semi-supervised learning. Supervised learning trains diagnostic models through completely labeled data to detect potential PV faults. By measuring the total voltage and the string current of PV systems, a random forest (RF) algorithm was used to identify PV faults in [19]. Although the RF training model can avoid the overfitting problem, the convergence time of the method in [19] increases with the number of decision trees. Chine *et al*. [8] introduced a diagnostic algorithm to combine the thresholding method with an artificial neural network (ANN) for identifying six types of PV faults. However, the ANN method may suffer slow learning and lack of generalization. In [9], an optimized kernel extreme learning machine was investigated to identify PV faults based on *I–V* curve measurements. Aiming to detect line-to-line faults under low irradiation with an active MPPT algorithm, Yi and Etemadi [20], [21] extracted fault features based on multiresolution signal decomposition, and proposed a fuzzy inference system in [20] and a two-stage support vector machine in [21]. However, methods [20], [21] based on transient signals may fail to detect faults happened in no irradiance condition. Belaout *et al*. [22] proposed a variable dimensionality reduction technique based on the area and the slopes at different points of the *I–V* curve, and developed a multiclass adaptive neuro-fuzzy classifier to discriminate five different PV faults. However, the method in [22] requires the calculations of too many features for the optimal selection, and needs to train several models for multiclass problems. Dhimish *et al*. [23] used the power and voltage ration as input features for artificial neural networks and fuzzy logic systems to detect faulty module and partial shading condition. Although the method in [23] has a high detection accuracy during normal and partial shading scenarios, it cannot detect internal fault of abnormal aging or sever dust accumulation via only two parameters. Moreover, the detection accuracies in [22] and [23] are easily affected by the PV simulation model, and the performance is limited to their feature normalization methods.

Unsupervised machine learning focuses on the inherent data characteristics, and needs extensive analysis and post-processing of reference training data. The unsupervised machine learning applied for PV fault diagnoses mainly calculates the clustering centers of reference data, e.g., density peak-based clustering [24], Gaussian kernel fuzzy C means-based clustering [25], Dilation and Erosion-based clustering [26], etc. Semi-supervised learning uses both labeled and unlabeled data, which can deal with the shortcomings of supervised learning to be ineffective for unlabeled data. Graph-based semi-supervised learning (GBSSL) is the semi-supervised learning applied to PV fault diagnoses until now [27], [28]. The GBSSL can diagnose PV faults via only a small amount of labeled data without training model. However, the stability of GBSSL methods is susceptible to noise, and the speed of sample testing slows down with the accumulation of historical data. Besides, the methods in [27] and [28] have not yet been able to utilize simulated labeled data instead of measured ones.

Among the aforementioned machine learning for PV fault diagnoses in [8]–[10], [19]–[28], supervised learning dominates the research trend, and its classification model has superior reliability. Unfortunately, supervised learning often requires a large amount of expensive labeled data, which is limited by the difficulty of obtaining the faulty data from real PV power plants. In the real PV systems, operational & maintenance (O&M) companies tend to store a large number of unlabeled historical data in the cloud to be not sufficiently utilized. Semi-supervised learning algorithm can use these unlabeled historical data with a small amount of labeled data for classification, which has a good prospect for PV fault diagnoses. Until now, the research works on semi-supervised learning for PV faults are still inadequate.

The core assumptions of a semi-supervised learning algorithm are summarized as follows. 1) Clustering assumption: data from the same class are likely to lie in the same cluster. 2) Manifold assumption: data points to be locally close to each other are likely to have the same class label. Clustering assumption makes decision boundary located in a low-density region of both labeled and unlabeled data, such as transductive support vector machine (TSVM) in [29], semi-supervised support vector machines in [30]. The manifold assumption forces the decision boundary sufficiently smooth with respect to the intrinsic structure, such as leaning with local and global consistency (LGC) algorithm in [31] and Laplacian support vector machines (LapSVM), which fully learns the local structure by labeled and unlabeled data. Combining with the manifold assumption, Huang *et al.* [32] improved the loss function of extreme learning machine (ELM) in [33], and proposed the semi-supervised ELM (SSELM) to show excellent performance compared with other algorithms. However, the selection of hyperparameters directly affects the classification accuracy, and the generalization of training model in the SSELM can be further improved.

In this study, *I–V* curves of PV strings under different fault states are analyzed. Characteristic parameter normalization

equations of *I–V* curves are tuned via low-cost data under normal operation of PV strings. A hybrid artificial bee colony optimization and semi-supervised extreme learning machine (ABC-SSELM) is investigated as a pattern recognition method for faults diagnoses of PV strings. The proposed PV fault diagnostic technology can effectively identify short circuit, two types of partial shading, abnormal aging, non-uniformed soiling condition, and faults under soiling condition. The proposed ABC-SSELM diagnostic model only needs a small amount of labeled data and can utilize historical unlabeled data from PV systems. Moreover, the simulated fault labeled samples are used instead of experimental ones to further save the labor cost and time-consumption. In this study, two different PV modules are used to validate simulated and experimental data. In addition, the performance of the proposed ABC-SSELM is evaluated in comparisons with other machine learning methods to show the reliability and accuracy of the proposed technology.

As our knowledge goes, there are no methods to deal with the problems of PV faulty diagnoses by considering dust impact simultaneously. The advantages of the proposed method over other strategies are recited as follows.

1) The requirement of less labeling data than [8]–[10], [19] and [22], [23] for PV fault diagnoses.
2) The labeled data in numerical simulations can replace experimental data without extra labeled verification set, which is superior to [8]–[10], and [19]–[28].
3) The performance of PV fault classification is better than the ones of the algorithms in [10], [27], [28] and [32], [33].
4) The generalization ability of the proposed method is verified by different PV modules.

The rest of this study is organized as follows: Section II introduces the MATLAB/Simulink-based PV modeling, and analyzes faults impact on the *I–V* curves. In Section III, the parameter normalization and the proposed ABC-SSELM for the PV fault diagnostic model are expressed in detail. The performance of the fault diagnostic method is verified by numerical simulations and experimental results in Section IV. Section V concludes this study.

II. Modeling and Fault Analyses of PV Strings

In a real PV system, the outdoor irradiance and temperature are uncontrollable. In general, the precise control for environmental conditions is required to analyze *I–V* characteristics of PV strings. Besides, a wide range of controllable irradiance and temperature is necessary in order to obtain fault samples under any environmental factors. Fortunately, an online *I–V* tracker incorporated into a smart inverter can get *I–V* curves of each PV string without power interruption [8]. Therefore, one uses the MATLAB/Simulink software to model PV strings for obtaining the *I–V* curve data. By setting various faults under standard test condition (STC), variation rules of output *I–V* curves and electrical characteristics of PV strings are analyzed to extract PV fault diagnostic features.

A. PV String Modeling Via MATLAB/Simulink

In this analysis, a PV system is formed by a string with 13 series PV modules. Each module consists of 60 cells connected in series, which evenly gathered into three substrings with three bypass diodes. The *I–V* testing circuit and the PV module via the MATLAB/Simulink simulation are depicted in Fig. 1. By controlling the output value of the voltage source to linearly increase, the output current and voltage of the PV string are recorded, and then input the corresponding data into the MATLAB workspace to obtain the final *I–V* curve. A rectifier diode is employed in the output of the PV string to avoid the occurrence of negative currents. Irradiances and temperatures are set by the gain amplifier of each substring.

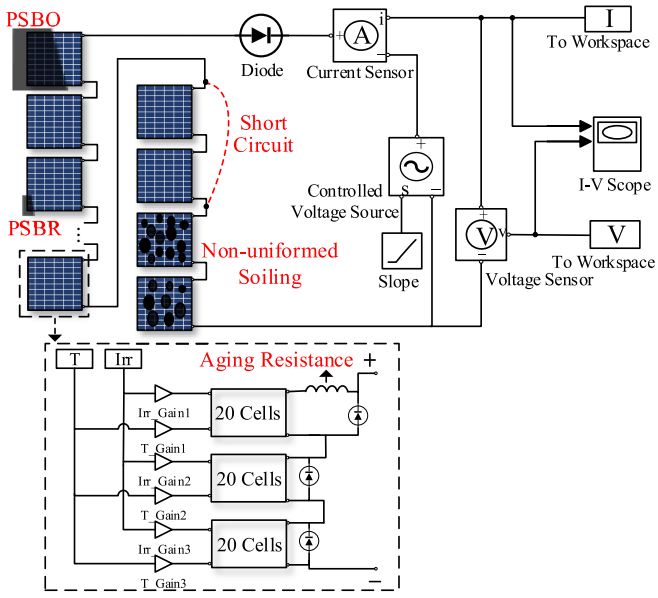

Fig. 1. *I–V* testing circuit and PV module modeling via MATLAB/ Simulink.

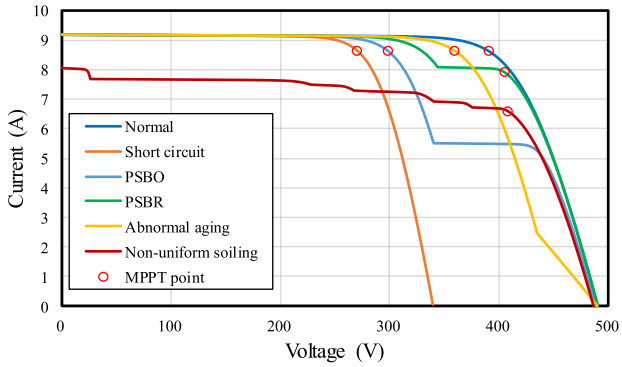

Fig. 2. Typical *I–V* curves in single fault condition under STC.

B. PV Fault Analysis

The PV string faults in this study include short circuit, abnormal aging, two types of partial shading, and non-uniformed soiling. Typical *I–V* curves at a single fault condition under STC are depicted in Fig. 2. Short-circuit is an accidental connection between two points of different potential among PV modules or PV cables. Abnormal aging distorts the lower end of the *I–V* curve, and the aging resistance [34] can be defined as (1) and (2). Based on the activation of bypass diodes inside

TABLE I
VARIATIONS OF FAULTY FEATURES AT STC

Fault Type	V_{oc}	I_{sc}	V_m	I_m	R_s
Short circuit	↓	—	↓	—	—
PSBO	—	—	↓	—	—
PSBR	—	—	↑	↓	—
Abnormal aging	—	—	↓	—	↑
Non-uniform soiling	—	↓	↑	↓	—

shaded PV modules at the global MPPT point, shading faults are divided into two types, which contain the partial shading with the bypass-diode reversed (PSBR) and the partial shading with the bypass-diode on (PSBO). In the soiling condition, dust adheres to the surface of PV modules, which reduces the amount of incident radiation on the panel. This phenomenon results in a great effect on the current in *I–V* curve. In this study, dust accumulation is regarded as a particular form of shading, which occurs in all PV modules. Note that the output current of a PV module is determined by the degree of shading. Therefore, the short-circuit current ($I_{\rm sc}$) is an important index to measure the severity of PV string dust deposition [3]–[6]. In a real PV string, the short-circuit current ($I_{\rm sc}$) more specifically represent the least soiling PV module in the string because each PV module has a different degree of dust deposition. The dust deposition in this study is non-uniform soiling to cause more than 20% power loss, which can be regarded as the hybrid fault of the uniform soiling and the PSBR

$$R_s = -\frac{dV}{dI}\bigg|_{V\cong V_{\rm oc}} = \frac{V_{\rm oc}-V_1}{I_1} \tag{1}$$

$$R_s = \frac{1}{3}\left(\frac{V_{\rm oc}-V_1}{I_1}+\frac{V_{\rm oc}-V_2}{I_2}+\frac{V_{\rm oc}-V_3}{I_3}\right) \tag{2}$$

where (I_1, V_1), (I_2, V_2), and (I_3, V_3) are three closest *I–V* points to (0, $V_{\rm oc}$). In order to suppress external interference and measurement noise in experimentations, one can modify (1) to (2) by averaging three estimated values of R_s.

Characteristic parameters of *I–V* curve at single faulty state under STC are summarized in Table I. The open-circuit voltage $V_{\rm oc}$, short-circuit current $I_{\rm sc}$, maximum power point voltage V_m and current I_m, equivalent series resistance R_s are considered as PV fault diagnostic features in this study. These features with different variations perform the characteristics of different faults under the STC. Therefore, the fault type of a PV system can be accurately discriminated when the selected features of the PV strings can be converted into the ones under the STC. At the STC, the characteristics of the short circuit, the PSBO, and the abnormal aging under the non-uniformed soiling can be regarded as the superposition of single faulty state.

III. NOVEL FAULT DIAGNOSTIC TECHNIQUES FOR PV SYSTEMS

The proposed fault diagnostic method for PV systems in this study includes the data normalization process and the pattern-recognition theory. The parameter normalization method eliminates environmental effect based on parts of low-cost normal operated PV data. As for the pattern-recognition theory, a hybrid ABC-SSELM is proposed via unlabeled data to improve the performance of the diagnostic model. The procedure is elaborated in the following sections.

A. Parameter Normalization

In order to eliminate the influence of irradiance and temperature on PV systems and sensors placement, the feature normalization method in [10] is also applied for accurate identification. As for the normalization process, the low-cost normal operating data of PV strings under different irradiances are used to tune unknown coefficients (*a*, *b*, *c*, *d*, *e*) of output *I–V* curve characteristic equation as shown in (3)–(7). These characteristic coefficients can be solved by the nonlinear least-squares fitting method. After shifting the output equation and dividing by the corresponding reference value, one can obtain the normalized equations as (8)–(12). The characteristic parameters ($V_{\rm oc}$, $I_{\rm sc}$, V_m, I_m, R_s) of *I–V* curves are normalized by (8)–(12) to form a diagnostic feature of five dimensions

$$V_{{\rm oc}.f} = V_{\rm oc.stc} + a_1\cdot\ln\frac{G}{G_{\rm stc}} + a_2\cdot dT + a_3\cdot\frac{G}{G_{\rm stc}}dT \tag{3}$$

$$I_{{\rm sc}.f} = b_1\cdot I_{\rm sc.stc}\frac{G}{G_{\rm stc}} + b_2\cdot dT + b_3\cdot\frac{G}{G_{\rm stc}}dT \tag{4}$$

$$V_{m.f} = V_{m.\rm stc} + c_1\cdot\ln\frac{G}{G_{\rm stc}} + c_2\cdot dT + c_3\cdot\frac{G}{G_{\rm stc}}dT \tag{5}$$

$$I_{m.f} = d_1\cdot I_{m.\rm stc}\frac{G}{G_{\rm stc}} + d_2\cdot dT + d_3\cdot\frac{G}{G_{\rm stc}}dT \tag{6}$$

$$R_{s.f} = R_{s.\rm stc}\left(\frac{G}{G_{\rm stc}}\right)^{e_1} + e_2\cdot dT + e_3\cdot\frac{G}{G_{\rm stc}}dT \tag{7}$$

where G is the measured irradiance; $G_{\rm stc}$ is a constant of 1000 W/m^2; dT is the measured temperature minus the STC temperature; $V_{{\rm oc}.f}$, $I_{{\rm sc}.f}$, $V_{m.f}$, $I_{m.f}$, and $R_{s.f}$ represent the open-circuit voltage, the short-circuit current, the voltage and current at the MPPT point, and the equivalent series resistance under different irradiances and temperatures for parameter fitting, respectively. $V_{\rm oc.stc}$, $I_{\rm sc.stc}$, $V_{m.\rm stc}$, $I_{m.\rm stc}$, and $R_{s.\rm stc}$ represent the open-circuit voltage, the short-circuit current, the voltage and current at the MPPT point, and the equivalent series resistance under STC, respectively

$$V_{\rm oc.norm} = \frac{1}{V_{\rm oc.stc}}\left(V_{\rm oc} - a_1\cdot\ln\frac{G}{G_{\rm stc}} - a_2\cdot dT - a_3\cdot\frac{G}{G_{\rm stc}}dT\right) \tag{8}$$

$$I_{\rm sc.norm} = \frac{1}{I_{\rm sc.stc}}\left(\frac{G_{\rm stc}}{b_1\cdot G}\left(I_{\rm sc} - b_2\cdot dT - b_3\cdot\frac{G}{G_{\rm stc}}dT\right)\right) \tag{9}$$

$$V_{m.\rm norm} = \frac{1}{V_{m.\rm stc}}\left(V_m - c_1\cdot\ln\frac{G}{G_{\rm stc}} - c_2\cdot dT - c_3\cdot\frac{G}{G_{\rm stc}}dT\right) \tag{10}$$

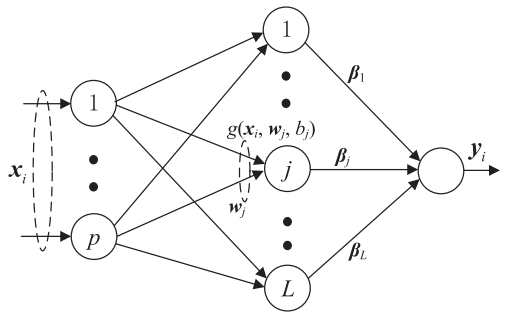

Fig. 3. ELM network.

$$I_{m.\mathrm{norm}} = \frac{1}{I_{m.\mathrm{stc}}}\left(\frac{G_{\mathrm{stc}}}{d_1 \cdot G}\left(I_m - d_2 \cdot dT - d_3 \cdot \frac{G}{G_{\mathrm{stc}}}dT\right)\right) \tag{11}$$

$$R_{s.\mathrm{norm}} = \frac{1}{R_{s.\mathrm{stc}}}\left(\left(\frac{G}{G_{\mathrm{stc}}}\right)^{-e_1}\left(R_s - e_2 \cdot dT - e_3 \cdot \frac{G}{G_{\mathrm{stc}}}dT\right)\right). \tag{12}$$

where V_{oc}, I_{sc}, V_m, I_m, and R_s represent the measured values of the open-circuit voltage, the short-circuit current, the voltage and current at the MPPT point, and the equivalent series resistance, respectively. $V_{\mathrm{oc.norm}}$, $I_{\mathrm{sc.norm}}$, $V_{m.\mathrm{norm}}$, $I_{m.\mathrm{norm}}$, and $R_{s.\mathrm{norm}}$ are the corresponding normalized parameters.

B. Semi-Supervised Extreme Learning Machine

ELM is similar to a single-hidden layer feedforward network including an input layer, a hidden layer and an output layer [33]. The ELM has the ability of fast training speed, and its simple structure is depicted in Fig. 3. The key of ELM is to find a mapping space from the input to the output with a minimum error. For N samples $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, where $\boldsymbol{x}_i \in \mathbf{R}^p$ and $\boldsymbol{y}_i \in \mathbf{R}^q$, where p and q represent individual dimensions. Given hidden nodes L and the activation function g (*), the connective weights (w_j) and the hidden biases (b_j) are randomly generated according to a continuous probability distribution. The output matrix of the hidden layer ($\mathbf{H}$) can be defined as follows:

$$\mathbf{H} = \begin{bmatrix} \boldsymbol{h}(\boldsymbol{x}_1) \\ \vdots \\ \boldsymbol{h}(\boldsymbol{x}_N) \end{bmatrix} = \begin{bmatrix} g(\boldsymbol{w}_1^T\boldsymbol{x}_1 + \boldsymbol{b}_1) & \cdots & g(\boldsymbol{w}_L^T\boldsymbol{x}_1 + b_L) \\ \vdots & \ddots & \vdots \\ g(\boldsymbol{w}_1^T\boldsymbol{x}_N + \boldsymbol{b}_1) & \cdots & g(\boldsymbol{w}_L^T\boldsymbol{x}_N + b_L) \end{bmatrix} \tag{13}$$

where the superscript "T" is the transpose operator. Then, the outputs layer of the ELM network can be represented as follows:

$$\hat{\boldsymbol{y}}_i = \mathbf{H}(\boldsymbol{x}_i)\boldsymbol{\beta}, \quad i = 1, \ldots, N \tag{14}$$

where β are the weights between the hidden layer and the output layer; Huang *et al.* [33] have proved that one can obtain the unique minimum norm least squares solution via the following Moore–Penrose inverse

$$\hat{\boldsymbol{\beta}} = \mathbf{H}^{\dagger}\mathbf{Y} \tag{15}$$

where † is the Moore–Penrose inverse of a matrix.

As it is generally accepted that ELM is not well suited to deal with data that are outside of the range of the data used for calibration. As a supervised learning algorithm, ELM requires large labeled samples to be hard to obtain, and cannot use unlabeled samples. Based on the manifold assumption, the manifold regularization framework [32] is introduced into the ELM to improve the loss function of the ELM for forming the SSELM. The regularization term of manifold can be expressed as follows:

$$L_m = \frac{1}{2}\sum_{i,j} w_{ij}\left\|\hat{y}_i - \hat{y}_j\right\|^2 = \mathrm{Tr}(\hat{\mathbf{Y}}^T\mathbf{L}\hat{\mathbf{Y}}) \tag{16}$$

where w_{ij} is the pair-wise similarity between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ as shown in (17). $\mathbf{L} \in \mathbf{R}^{(l+u)\times(l+u)}$ is the *graph Laplace* defined as (18), in which *l* and *u* represent the number of labeled and unlabeled training samples, respectively, and ten-nearest-neighbor graph is selected in this study. Tr (*) denotes the trace operator of a matrix

$$w_{ij} = e^{-\frac{|x_i - x_j|^2}{2\sigma^2}} \tag{17}$$

$$\mathbf{L} = \mathbf{D} - \mathbf{W} \tag{18}$$

where the matrix $\mathbf{D}$ is a diagnostic matrix with the following elements:

$$d_{ii} = \sum_{j=1}^{l+u} w_{ij}. \tag{19}$$

The objective function of the SSELM is defined as follows:

$$\begin{aligned} \min_{\beta\in\mathbb{R}^{L\times d}} & \frac{C_i}{2}\sum_{i=1}^{l}\|\boldsymbol{\varepsilon}_i\|^2 + \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \frac{\lambda}{2}\mathrm{Tr}\left(\hat{\mathbf{Y}}^T\mathbf{L}\hat{\mathbf{Y}}\right) \\ \text{s.t.} \quad & \boldsymbol{h}(x_i)\boldsymbol{\beta} = \boldsymbol{y}_i^T - \boldsymbol{\varepsilon}_i^T, \quad i = 1, \ldots, l \\ & \hat{\boldsymbol{y}}_j = \boldsymbol{h}(x_j)\boldsymbol{\beta}, \quad j = 1, \ldots, l+u \end{aligned} \tag{20}$$

where λ is the penalty coefficient of the manifold term; $\boldsymbol{\varepsilon}_i$ is the error vector caused by the *i*th labeled training sample. Similar to the weighted ELM [35], C_i is a penalty coefficient with respect to patterns from different classes for solving the problem of imbalance data, which is defined as follows:

$$C_i = \frac{C_0}{N_i} \tag{21}$$

where C_0 is a user-defined parameter; N_i is the number of training samples of labeled $\boldsymbol{y}_i$.

According to [32], when the number of labeled training data is greater than or equal to the number of neurons in the hidden layer, the solution can be obtained by

$$\boldsymbol{\beta} = \mathbf{H}^T\left(\mathbf{I}_l + \mathbf{H}^T\mathbf{C}\mathbf{H} + \lambda\mathbf{H}^T\mathbf{L}\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{C}\mathbf{Y}. \tag{22}$$

When the number of labeled training data is less than the number of neurons in the hidden layer, the solution should be computed by the following alternative way:

$$\boldsymbol{\beta} = \mathbf{H}^T\left(\mathbf{I}_{l+u} + \mathbf{C}\mathbf{H}\mathbf{H}^T + \lambda\mathbf{L}\mathbf{H}\mathbf{H}^T\right)^{-1}\mathbf{C}\mathbf{Y}. \tag{23}$$

The advantage of the SSELM in comparisons with the TSVM and the LapSVM is that the SSELM naturally handles the multiclassification problems. The major implementation of the

SSELM is to calculate the matrix of $\mathbf{H}$ and solve the output weight, β. However, the selection of penalty coefficients λ and C_0 is related to the performance of the SSELM, and empirical values by manual setting are always adopted. Besides, the lack of labeled verification data set may result in ill-posed SSELM model, which suffers overfitting problems. Therefore, the artificial bee colony (ABC) algorithm is utilized to optimize penalty coefficients and improve the generalization of the SSELM model.

C. Hybrid Artificial Bee Colony Algorithm and Semi-Supervised Extreme Learning Machine

ABC algorithm is one of the swarm intelligence algorithms inspired by foraging behavior of bee colony [36]. The performance of the ABC algorithm with less control parameters is better than or similar to other population-based algorithms, such as particle swarm optimization (PSO) and genetic algorithm [37]. The search model of the ABC consists of four basic parts: food source, employed, onlooker, and scout bees. The objective of optimization is to search the best nectar around food source. Generally, the fitness function of the ABC represents the space of all food sources. Employed bees investigate food source and share information with onlooker bees, which further explored the food source with a certain probability. If employed and onlooker bees cannot find a better nectar from a food source for a long time, it would be abandoned by scout bees and find a new food source, which avoid falling into a local optimum. The implementation of the ABC algorithm in the SSELM can be explained as follows.

Initialization: The population number S is set to 10, and the max cycle number is set to 100 in this study. The real coding method for parameters (λ and C_0) of the SS-ELM to be optimized is used to reduce the dimension

$$\lambda = 10^{x_1}, \quad C_0 = 10^{x_2}. \tag{24}$$

Moreover, the location of each food source in the ABC can be represented by the following two-dimensional space:

$$\boldsymbol{x}_i = [\, x_{i1} \; x_{i2} \,]\big|_{i=1,\ldots,S} \tag{25}$$

where S is the number of bee populations. The upper and lower bounds of the food source position is constrained as follows:

$$\mathrm{UB} = [-10, -10], \quad \mathrm{LB} = [10, 10]. \tag{26}$$

In addition, the location of the initial food source can be randomly generated as follows:

$$x_{\mathrm{id}} = \mathrm{LB}_d + (\mathrm{UB}_d - \mathrm{LB}_d) \times \mathrm{rand}(0,1) \qquad d = 1, 2; \quad i = 1, \ldots, S. \tag{27}$$

Employed bees: Each food source $\boldsymbol{x}_i$ is correspondingly sent to an employed bee for searching nectar as (28). If better nectar is found, then the food source updates to the new position, i.e., $\boldsymbol{x}_i$ updates to $\boldsymbol{v}_i$. Otherwise, the food source $\boldsymbol{x}_i$ remains. In this study, the objective of the ABC is to search the minimum of the fitness functions as shown in (29)

$$v_{\mathrm{id}} = x_{\mathrm{id}} + \phi_{\mathrm{id}} \times (x_{\mathrm{id}} - x_{kd}) \tag{28}$$

where ϕ_{id} is a random number distributed evenly on $[-1, 1]$

$$\begin{aligned} \mathrm{Fit} = {} & a_f \cdot \frac{\sum_{k=1}^{l} I(\hat{\boldsymbol{y}}_k \neq \boldsymbol{y}_k)}{l} \\ & + b_f \cdot \frac{\sum_{i=1}^{u} \sum_{j=1}^{l+u} I(\hat{\boldsymbol{y}}_i \neq \hat{\boldsymbol{y}}_j) \frac{1}{w_{ij}}}{u \times \sum_{i=1}^{u} \sum_{j=1}^{l+u} (\frac{1}{w_{ij}})} \\ & + c_f \cdot \|\boldsymbol{\beta}\| \\ \text{s.t.} \quad & w_{ij} \neq 0, \; a_f = \frac{u}{l+u}, \; b_f = \frac{l}{l+u}, \; a_f \gg c_f, \; b_f \gg c_f \end{aligned} \tag{29}$$

where $I(^*)$ equals to unity when the condition $(^*)$ can be satisfied; $\hat{\boldsymbol{y}}_k$ and $\boldsymbol{y}_k$ represent prediction and original labels of labeled samples; $\hat{\boldsymbol{y}}_i$ denotes the prediction of unlabeled samples, $\boldsymbol{x}_i$; $\hat{\boldsymbol{y}}_j$ expresses the prediction of ten-nearest-neighbor of $\boldsymbol{x}_i$. The first item on the right-hand side of (29) is the training error of labeled data. The intermediate term of (29) represents the clustering of unlabeled samples, which means that unlabeled samples with the same structure belong to the same class. The third term of (29) is the norm of the output weights in the SSELM. The coefficients (a_f and b_f) are complementary weights, which means that the importance of labeled and unlabeled data is inversely proportional to their number. The value of c_f to be far less than a_f and b_f represents the best generalization ability of the SSELM model to be found from that of the most satisfied the first two terms of the fitness function.

Onlooker bees: Based on a new food source information from employed bees, onlooker bees are sent to further explore the food source with a certain probability according to the fitness value (30). The search strategy of onlooker bees is the same as employed bees as (28)

$$P(\boldsymbol{x}_i) = \frac{1/\mathrm{Fit}(\boldsymbol{x}_i)}{\sum_{m=1}^{S} 1/\mathrm{Fit}(\boldsymbol{x}_m)}. \tag{30}$$

Scout bees: Some food sources may remain unchanged after several generations of employed and onlooker bees, which is possibly trapped in local optima. Therefore, the scout bee operator discards the unchanged food source and find a new one instead according to (27).

With sufficient labeled data, the optimal penalty parameters λ and C_0 can be determined by the ABC algorithm based on the training error of the verification set. Unfortunately, in real PV systems, the fault labeled data are difficult to obtain. The training model of the SSELM may result in ill-posed models based on the training error of insufficient labeled data. In this study, the clustering degree of unlabeled data, and the output weight of the SSELM are also taken into account. According to [38], the smaller the norms of weights for feedforward neural networks tend to have better generalization performance. Since the fitness function (29) of the ABC is to optimize the corresponding parameters and search for the best generalization ability of the SSELM model, the number hidden nodes in the SSELM is selected by the tradeoff between the system performance and the computation time.

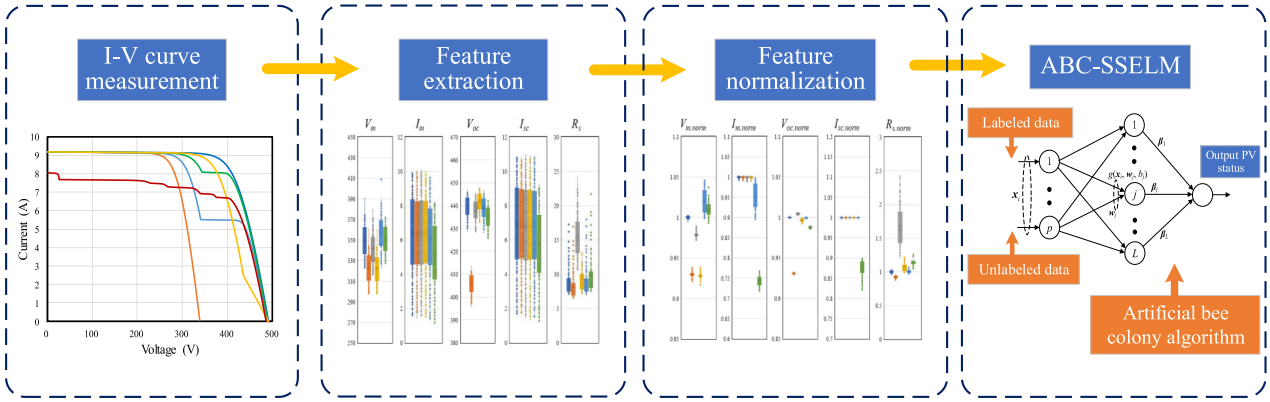

Fig. 4. Framework of the proposed PV fault diagnostic method.

After the parameter optimization by the ABC, the optimal PV fault diagnostic model can be obtained. The proposed PV fault diagnostic technology is depicted in Fig. 4. It is worthy to note that using the parameters of normal I-V curves to regularly tune the normalized equation can adapt to the natural aging of PV modules and maintain the long-term reliability of the PV fault diagnostic model.

The innovative points of the proposed method in this study are recited as follows:

Based on the characteristics of dust deposition in a PV system, the influence of dust on the output characteristics of a PV module is analyzed, and non-uniformed soiling and faults under the occurrence of non-uniformed soiling in a PV system are considered in the fault identification types. As our knowledge goes, there are no literatures to deal with the problems of PV faulty diagnoses by considering dust impact simultaneously.

Fig. 4 is a general framework of PV fault diagnosed technology based on machine learning including data acquisition, data pre-procession and diagnostic model establishment. Compared with other literatures, one of the innovations in the proposed method is the semi-supervised learning algorithm of ABC-SSELM at the step 4 of Fig. 4. The proposed algorithm only needs a small amount of labeled data, and can take advantage of the historical unlabeled data of a PV system to establish a fault diagnostic model. In the previous research works, supervised learning algorithms in [19]–[23] only can use expensive labeled data to build the corresponding models.

Compared with [8]–[10] and [19]–[28], the combination with parameter normalization method into the proposed method can make the simulated labeled data to be used to replace the fault labeled data of a real PV system, which greatly reduces the human and time costs to reprocess the information of the PV plant.

IV. Verification of PV Diagnostic Techniques

As described in Section II, the normal operation and five fault types including the short circuit, the PSBR, the PSBO, the abnormal aging and the non-uniformed soiling are considered. Moreover, hybrid faults of the short circuit, the PSBO, and the abnormal aging under non-uniformed soiling are also examined. Therefore, this study has nine types of PV operating states including normal and un-normal operational modes to be discriminated totally. This section introduces the process of data acquisition, and shows the performance of parameters normalization. Moreover, the performance and superiority of the proposed PV fault diagnostic technology can be verified by numerical simulations, experimental results, and hybrid simulation and experiment cases in comparisons with other methods.

TABLE II
PV Module Parameters of PVM1 and PVM2

PV Module	PVM1	PVM2
P_{max}/Wp	260	300
V_m/V	30.02	31.51
I_m/A	8.66	9.52
V_{oc}/V	37.78	39.24
I_{sc}/A	9.12	9.93
k_p/%/k	-0.4631	-0.4003
k_v/%/k	-0.3315	-0.2906
k_i/%/k	0.0443	0.053
N_s	60	60

A. Data Acquisition

Two types of modules including the PVM1 manufactured by the polycrystalline silicon and the PVM2 manufactured by the monocrystalline silicon as shown in Table II are used to form two PV systems (3.51 and 3.9 kWp) with 13 modules in series for both simulated and experimental verifications.

1) Acquisition of Simulated Data: In Section II, the *I–V* testing circuit is set up with different conditions to obtain the corresponding *I–V* curves. The value *of the irradiance gain amplifier is randomly set during the* range of [0.3, 0.6] to simulate the PSBO condition, while set randomly during the range of [0.88, 0.95] to simulate the PSBR condition. As a special shading, the irradiance gain amplifier of all modules is set during the range of [0.7, 0.9] to simulate the situation under non-uniformed soiling. In the abnormal aging fault, the value of the aging resistance is randomly set during the range of [3 Ω, 10 Ω]. In order to cover a wide range of operating environment, the irradiance ranges from 100 to 1200 W/m^2, while the temperature varies from 35 to 65 °C synchronously. The varied steps

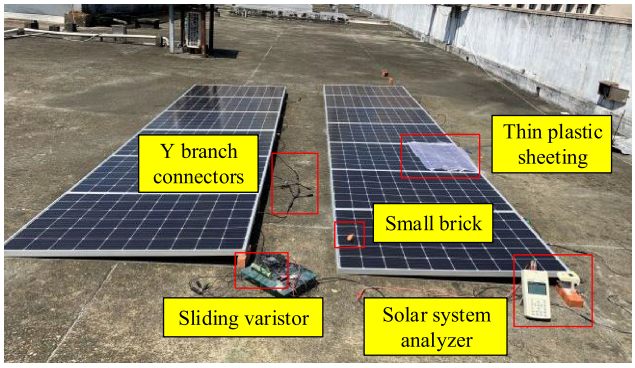

Fig. 5. Experimental hardware platform and faults creation.

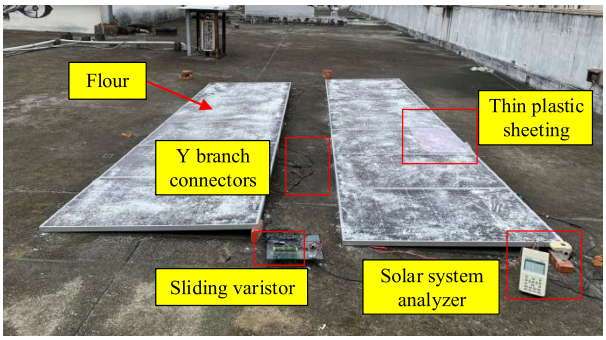

Fig. 6. Experimental faults setup under non-uniform soiling condition.

are determined by a certain value (A) and a randomly varying value [random(B)] in (31) to truly reflect the real environment. There are 600 simulated data samples in every nine categories, and the total number of simulated data samples is 5400 for a PV string

$$\text{Step} = A + \text{random}(B). \tag{31}$$

2) Acquisition of Experimental Data: The experimental site as shown in Fig. 5 is located at National Taiwan University of Science and Technology in Taiwan. In experimental cases, *I–V* curves of PV systems are collected by a solar system analyzer (PROVA1011) manufactured by TES Electrical Electronic Corp., and real-time irradiances and temperatures of a PV panel are measured by matching sensors. As shown in Fig. 5, the short-circuit fault is created by Y-branch connectors. Small pieces, such as small bricks or discarded cigarette boxes, are used to simulate the PSBR condition. Thin plastic sheeting or paper sheets are employed to simulate the PSBO condition. The thin plastic sheeting can reduce irradiance on the shaded area of a PV module, and it can be referred to the partial shading experiment in the previous research works [9], [10], [19], which result in obvious multipeak in *I–V* curves. Moreover, cardboard is also used as external objects to simulate partial shading condition in this study. The abnormal aging fault utilizes a sliding varistor as the aging resistor to be connected in series with the PV substring. The experiment of non-uniform soiling and its hybrid faults are depicted in Fig. 6. In this study, flour is used to simulate dust deposition, and each module is sprinkled with 50 g of flour.

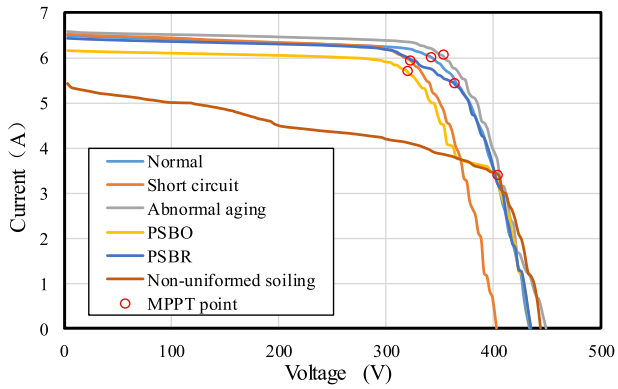

Fig. 7. Experimental *I–V* curves at irradiance 700 W/m^2 under the occurrence of single fault.

Owing to the output current of each module's substring to be limited by the most shaded cell, some cells without enough shaded area do not affect the overall output property. However, it would increase the possibility of hot spots formed by severely shaded cells. Artificial spraying in this experimental setup is hard to determine the same degree of the most severe shaded cell in each substring, which naturally creates the equivalent output characteristics of non-uniformed dust deposition. Irradiances in experimental environment range from 100 to 1000 W/m^2. The experimental *I–V* curves at irradiance 700 W/m^2 under the occurrence of single fault are depicted in Fig. 7. The characteristics in Fig. 7 are similar to the ones in Fig. 2. The total number of measured data from PVM1 and PVM2 are 3064 and 3013, respectively. The data selection criterion in experiments is that the weather during the measurement of *I–V* curves is required to be stable. In other words, experimental data under the weather varied dramatically during the measurements are excluded.

B. Parameters Normalization

According to the setup of numerical simulations and experimental platform, *I–V* curves of PV strings under different irradiances can be obtained, and characteristic parameters of open-circuit voltage V_{oc}, short-circuit current I_{sc}, maximum power point voltage V_m and current I_m, and equivalent series resistance R_s can be extracted as diagnostic features to be introduced in Section III. It is worthy to mention that objective errors in the experiment can be eliminated by the parameters normalization, e.g., the temperature difference between the PV panel and the measured backplane; the inconsistency of irradiance between cells and measured ones, and errors caused by measured equipment. Therefore, measured values can be taken as reference values in this study. The boxplot is formed by setting the single fault to show the performance of parameter normalization visually. Taking PVM1 as an example, the statistical distribution of the normalized simulated and experimental data is depicted in Fig. 8. The features show clustering and unification as the same characteristic at STC to be summarized in Section II. Although the experiment samples are disturbed by various environmental factors and existing outliers, the distribution is similar to the simulation one. This result verifies

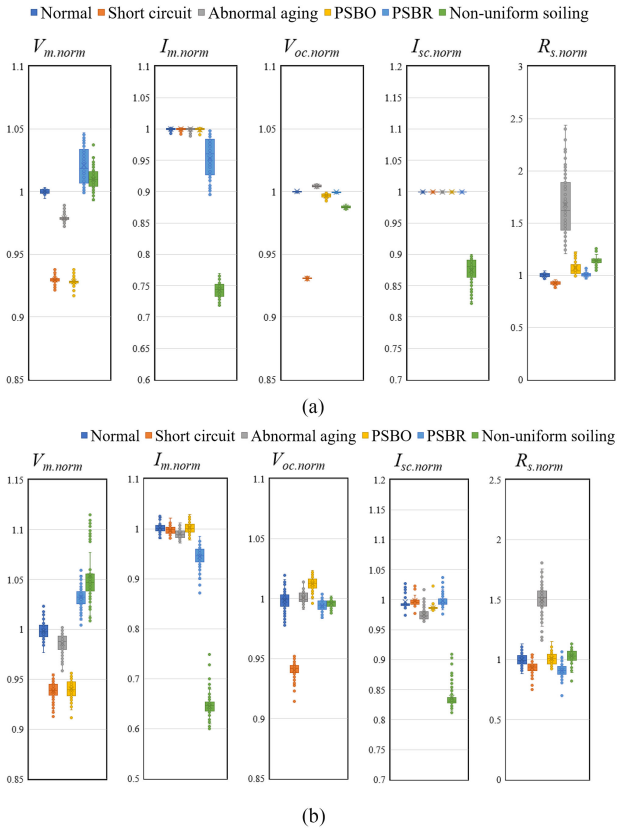

Fig. 8. Boxplot of five normalized feature variables. (a) Simulation samples distribution of PVM1. (b) Experimental samples distribution of PVM1.

the effectiveness of the parameter normalization and provides evidence for the feasibility of replacing the measured data with simulated ones under the absence of data.

C. Performance Verification of the Proposed ABC-SSELM

In order to verify the performance of the proposed ABC-SSELM algorithm for nine types of PV status classification, different numbers of unlabeled historical data are randomly obtained from the corresponding dataset, and the number of labeled data is with an increment step in each training. The remaining data are used to examine training models. Moreover, 50 times are run in each case, and the average accuracy is used to measure the performance of the proposed algorithm.

1) Case 1. Simulated Verification: In this case, labeled/unlabeled training and testing data are created from simulated datasets. The examined results of the proposed ABC-SSELM in different cases for PVM1 and PVM2 in Figs. 9(a) and 10(a) show that the test accuracy increases rapidly with the increase of the number of labeled samples. Note that, "UL" represents the number of unlabeled data to be used in the training case. Moreover, with the rise of unlabeled data, the stability of training model can be increased, and the test accuracy can be further improved. When the number of labeled data reaches 0.67% of the total data number, the average recognition accuracy is more than 98% in all cases of different unlabeled data numbers. This result shows that a large number of unlabeled data can be used to enhance the generalization and precision of training models. In addition, the number of labeled data account for more than 0.67% of the total data number, which can achieve a good performance of classifying nine types of PV status in the simulation verification.

2) Case 2. Experimental Verification: In experimental cases, labeled/unlabeled training and testing data are created from experimental platform. The examined results of the proposed ABC-SSELM in different cases for PVM1 and PVM2 are shown in Figs. 9(b) and 10(b), respectively. The number of unlabeled and labeled data has similar effects on the test accuracy as that of the simulated ones. Note that, unlike the simulation data, many interference factors exist in the measured experimental data, which will decrease the overall accuracy. However, with the increase of unlabeled and labeled samples, the proposed ABC-SSELM still performs well. When the number of labeled samples increases to be 45, which is only 1.5% of the total samples number, the average accuracy in all cases of PVM1 and PVM2 is higher than 96%. It should be noted that the quality of few labeled data directly affects the examined results. In other words, the diagnostic model would be affected by the labeled data with considerable noise.

3) Case 3. Hybrid Simulation and Experiment Verification: Because labeled faulty data of PV systems are difficult to obtain, one uses simulated data to replace the measured one and verifies the corresponding performance in this case. Labeled samples are taken from the simulation dataset, while measured datasets are divided into unlabeled samples and testing samples. As can be seen from Figs. 9(c) and 10(c), the influence of unlabeled data on the test accuracy is higher than that of both two earlier cases. The major reason is the distribution of simulated and measured data is different. Fortunately, the proposed ABC-SSELM can learn from the distribution of unlabeled data and enhance the generalization ability of the training model. With the increase of unlabeled samples, the distribution of experiment data is more definite, the performance of the training model is more stable, and the test accuracy is markedly improved.

Among these three verifications mentioned above, classification results (mean $\pm$ variance) of 90 labeled data and 500 unlabeled data are summarized in Table III. Labeled data account for corresponding datasets less than 3% of the total dada number in all cases. Note that the test accuracy of PVM2 in Case 3 is even better than the ones in Case 2, which means using the simulated labeled data to replace the measured one with sufficient historical unlabeled data could have superior performance. The reason is that the measured labeled data may with large noise called outliers which affects the model establishment, while the simulated data show better clustering in Fig. 8(a).

D. Comparison With Other Machine Learning Methods

Diagnostic features under STC in Table I are necessary for the establishment of PV diagnostic models. The downscaling methods maybe more suitable for continuous data or high-dimensional data, which are hard to apply in the proposed framework. For example, a multivariate statistical method, named as the principal component analysis, was used in [16] to classify

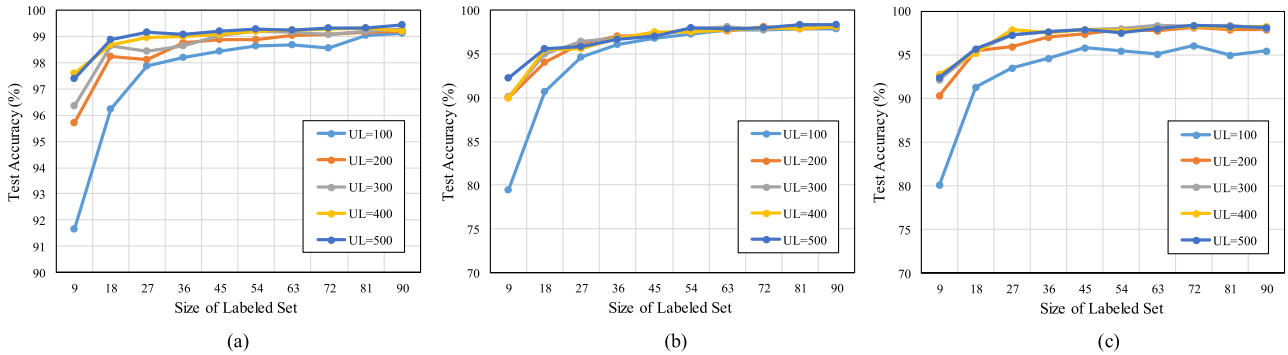

Fig. 9. Performance of proposed ABC-SSELM in different cases via PVM1. (a) Simulated verification. (b) Experimental verification. (c) Hybrid simulation and experiment verification.

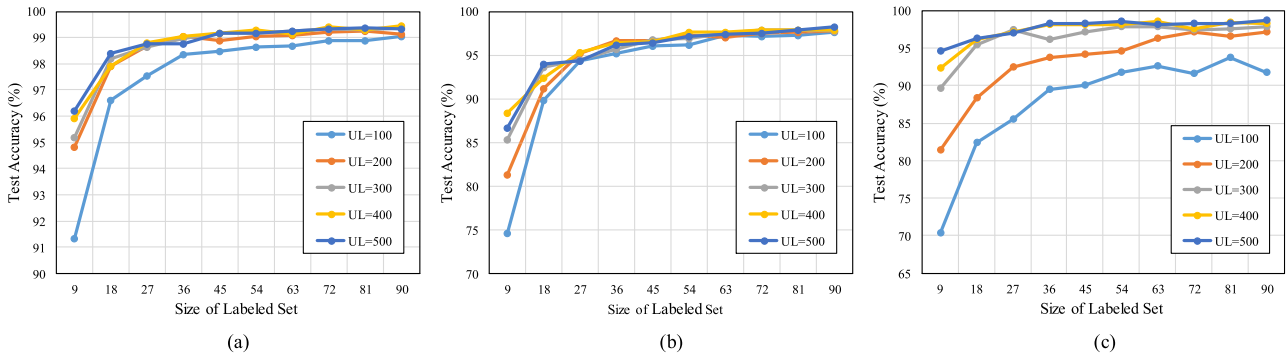

Fig. 10. Performance of proposed ABC-SSELM in different cases via PVM2. (a) Simulated verification. (b) Experimental verification. (c) Hybrid simulation and experiment verification.

TABLE III
PERFORMANCE OF THE PROPOSED ABC-SSELM IN DIFFERENT CASES

Data Set	PV Module	Accuracy	ABC-SSELM
Simulated Verification (Case 1)	PVM1	Average	99.44 ± 0.09
		Best	99.85
	PVM2	Average	99.34 ± 0.14
		Best	99.86
Experimental Verification (Case 2)	PVM1	Average	98.36 ± 0.58
		Best	99.59
	PVM2	Average	98.19 ± 0.62
		Best	99.42
Hybrid Simulation and Experiment Verification (Case 3)	PVM1	Average	98.17 ± 1.14
		Best	99.38
	PVM2	Average	98.74 ± 0.42
		Best	99.48

PV faults. In [16], input data are all sample points on *I–V* curves and it had different data preprocessing procedure. Due to different research ideas, downscaling methods cannot be directly compared with the proposed ABC-SSELM technology. In this section, performance comparisons between the proposed ABC-SSELM and other machine learning methods in [10], [27], [28], [32], and [33] by taking PVM1 as an example are presented. In this comparisons, the learning with LGC algorithm applied in [27] and [28], the original SSELM based on parameter-grid-search in [32], the stage-wise additive modeling using multiclass exponential loss function based on the classification and regression tree (SAMME-CART) in [10], the ELM in [33], and the PSO-based SSELM (PSO-SSELM) are introduced to discuss. The number of unlabeled data is set to be 300, and the size of labeled dataset is gradually increased in all comparisons. Two cases including experimental case (Case 2), and hybrid simulation and experiment case (Case 3) are used to verify the proposed ABC-SSELM in comparisons with other methods. Ten times are running in each case, and the average accuracy is used to measure the corresponding performance.

The classification results by the supervised learning including the ELM in [33] and the SAMME-CART in [10] to be compared with the proposed ABC-SSELM are depicted in Fig. 11(a) and (b), respectively. For Case 2, the proposed ABC-SSELM exhibits competitive accuracy with ELM and SAMME-CART under sufficient data condition. However, the supervised learning needs a large number of labeled data for training, and are insensitive to unlabeled data. Therefore, the ELM in [33] and the SAMME-CART in [10] have poor performance under the absence of labeled data in Case 2, especially for the ELM, while the proposed ABC-SSELM keeps excellent performance. Owing to different distribution of training and testing data for Case 3, even the increase of training label cannot improve the

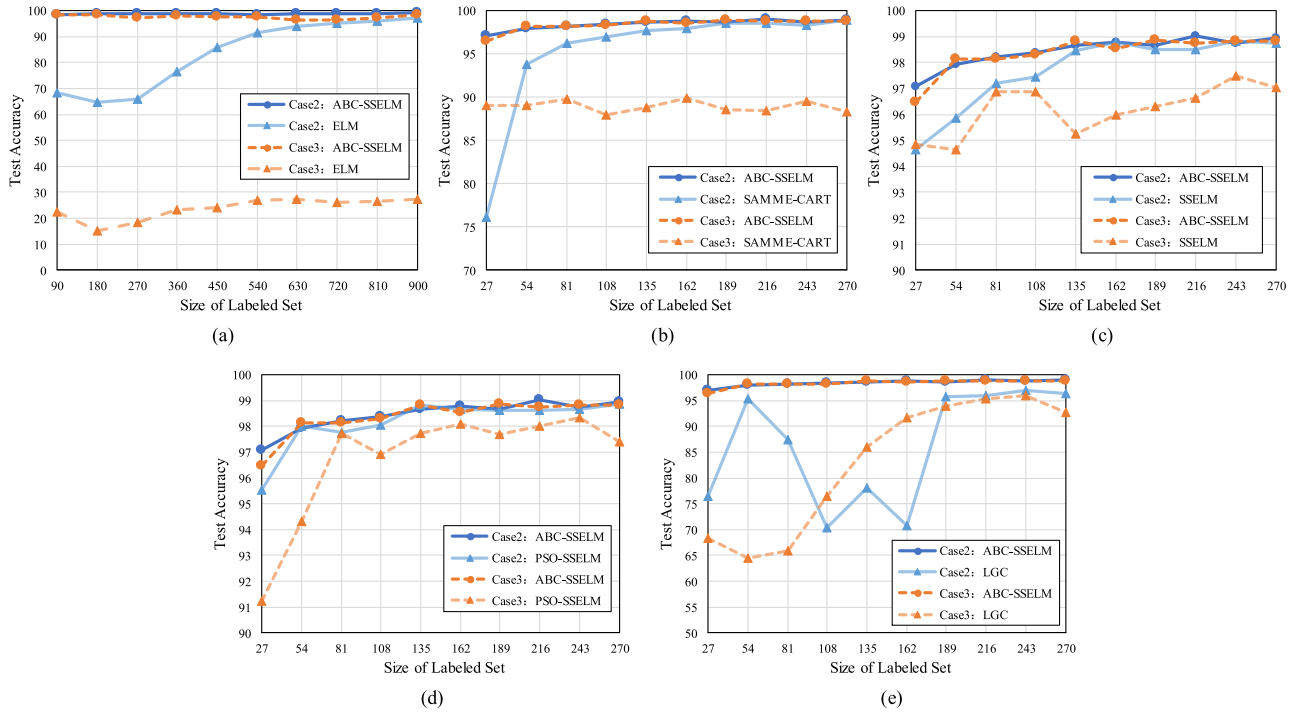

Fig. 11. Performance of proposed ABC-SSELM in comparison with other machine learning methods. (a) ELM in [33]. (b) SAMME-CART in [10]. (c) SSELM in [32]. (d) PSO-SSELM and (e) LGC in [27] and [28].

performance of the ELM in [33] and the SAMME-CART in [10], while the proposed ABC-SSELM still works well in this situation. Note that, with the increase of simulated labeled data, the test accuracy of the proposed ABC-SSELM in identifying measured samples decreases slightly for Case 3 in Fig. 11(a). The training model would be significantly dominated by simulated labeled data but less learning from the unlabeled data when the number of simulated labeled data is far more than the measured one. Therefore, the number of simulated label data should not exceed that of unlabeled historical one in practical applications.

The classification results by the proposed ABC-SSELM in comparisons with the original SSELM in [32] and the PSO-SSELM are depicted in Fig. 11(c) and (d). The original SSELM in [32] easily generates ill-posed models under the absence of labeled data. In addition, the PSO is used to optimize penalty coefficients in the SSELM to form the PSO-SSELM with the same objective function (29) to compare the performance of the proposed ABC-SSELM (i.e., the ABC algorithm in combination with the SSELM). These two swarm optimization algorithms are both set to have the same generation number as 100. As can be seen from Fig. 11(d), they perform similarly for Case 2. Moreover, the performance of the proposed ABC-SSELM is much better than the one of the PSO-SSELM for Case 3, and the ABC algorithm has fewer parameters to be determined than the PSO.

In [27] and [28], the graph-based semi-supervised learning algorithm is adapted to diagnose PV faults. The method used in [27] and [28] refers to the LGC algorithm in [31]. The LGC is a label propagation algorithm, which recognizes samples without initial training model. In [27], the test data are fed one after another into the LGC, and the corresponding model is updated in real time. It would increase the testing time because it has the time complexity of $O(n^3)$. In order to reduce the computational complexity and compare with the proposed ABC-SSELM fairly, the LGC algorithm is examined by using 300 random unlabeled data, and different number of labeled data without updating the model via the testing data, which is the same setting as the proposed ABC-SSELM. As can be seen from Fig. 11(e), the classification results of the proposed ABC-SSELM for Cases 2 and 3 are both superior to the ones of the LGC algorithm. Moreover, each testing sample fed into the LGC needs to retrain the model for predicting, and the time consuming is remarkable. Besides, the practical use of the LGC is more susceptible to the impact of outliers leading to the deterioration of performance. Obviously, the proposed PV fault diagnosis technology solves the problems in [27] and [28] to replace the measured data with the simulated one.

By taking the PVM1 module with 90 labeled data and 500 unlabeled data as an example, the computation time comparisons of different diagnostic methods are summarized in Table IV. The results show the sample testing time of the proposed ABC-SSELM is comparable to other ELM-based methods [32], [33], and is much shorter than the ones of the SAMME-CART supervised learning method in [10] and the LGC semi-supervised method in [27] and [28]. Moreover, the testing time of the LGC method in [27] and [28] would slow down with the data increased, which will increase the computational complexity for large data. Note that the online testing speed of the proposed ABC-SSELM method is the fastest one due to its simple structure.

TABLE IV
COMPUTATION TIME COMPARISONS OF DIFFERENT DIAGNOSTIC METHODS

Diagnostic method	Sample testing time
ELM in [33]	3.43 ms
SAMME-CART in [10]	29.40 ms
SSELM in [32]	3.47 ms
PSO-SSELM	3.63 ms
LGC in [27]-[28]	446 ms
ABC-SSELM	**3.33 ms**

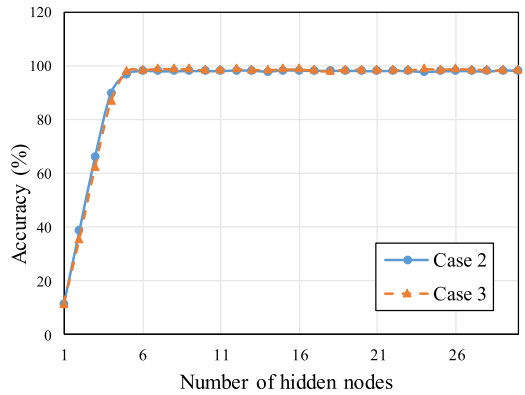

Fig. 12. Impact of number of hidden nodes in ABC-SSELM.

E. Discussion

By taking the PVM1 module with 90 labeled data and 500 unlabeled data as an example, 100 times are running in Cases 2 and 3. The average accuracy affected by the number of hidden nodes in the proposed ABC-SSELM method is depicted in Fig. 12. As can be seen from Fig. 12, the average accuracy can be gradually improved when the number of hidden nodes is increased until seven hidden nodes. Owing to the diagnostic features normalization and the parameters optimization, just a few numbers of hidden nodes can satisfy the requirement of this method. Therefore, the number of hidden nodes in the proposed ABC-SSELM method is set to be 7 in this study.

In Section IV-C and IV-D, the dynamic changes of labeled and unlabeled samples are used to show the superior performance of the proposed ABC-SSELM methodology. Each point in Figs. 9–11 represents the average classification accuracy at nine PV operating states. If all the accuracies for identifying nine kinds of PV operational states are expressed entirely, the length of this article will become too long. By taking the PVM1 module with 90 labeled data and 500 unlabeled data as an example, 10 times are running in each case, and the average accuracies for nine types of PV operational states are summarized in Table V. As can be seen from Table V, the average classification accuracy of the proposed ABC-SSELM methodology for nine kinds of PV operational states is over 98.44%.

From the above verification of three cases and the comparisons with other methods, the superiority of the proposed ABC-SSELM is obvious. In practice, the proposed PV fault diagnosis technology can make full use of a large number of historical data stored by PV O&M companies. Moreover, a small amount of labeled data can be replaced by simulated data, which further saves human and time costs. Although an online *I–V* tracker and related sensors are required in the practical implementation, the proposed PV fault diagnosis technology can monitor the running state of each PV string to inform potential faults and bring economic benefits. According to different environment around the world, effective cleaning policies for PV plates can be carried out by combining dust accumulation status with local weather forecast. For example, when the present power drops 20% of the theoretical value due to the soiling condition, an early warning can be carried out.

TABLE V
DETAILED CLASSIFICATION ACCURACY OF PVM1 AT THREE CASES

Types of PV operational states	Case 1 accuracy (%)	Case 2 accuracy (%)	Case 3 accuracy (%)
Normal	99.94	96.37	97.40
Short circuit	100	99.60	96.91
PSBR	100	96.16	96.93
PSBO	100	99.95	99.91
Abnormal aging	99.89	99.70	99.79
Non-uniformed soiling	98.28	98.78	98.41
Short circuit under non-uniformed soiling condition	100	99.89	99.71
PSBO under non-uniformed soiling condition	98.18	96.95	97.06
Abnormal aging under non-uniformed soiling condition	99.65	99.64	99.94
Average accuracy	**99.55**	**98.52**	**98.44**

The normalization of parameters requires a part of the normal operation data from PV strings for fitting. In the future study, it can recognize the normal data in the historical data first. As a result, characteristic parameters can be normalized directly via the normal historical data, and the cost of technology initialization can be further saved. More experiment of quantitative analysis on the degree of dust accumulation can be taken into account. Even if the sticky flour is used in the test, the power would still rise sharply after heavy rains. According to the local weather forecast, the optimal economic analysis would be further study to contribute to the best cleaning policies for PV plates.

V. CONCLUSION

In this study, an ABC-SSELM has been successfully designed for PV fault diagnoses. The normal operation and five fault types including the short circuit, the PSBR, the PSBO, the abnormal aging, and the non-uniformed soiling are considered. Moreover, hybrid faults of the short circuit, the PSBO, and the abnormal aging under non-uniformed soiling are also examined. The effectiveness of the proposed method is verified by practical PV strings with the power capacities of 3.51 and 3.9 kWp.

The main contributions of this study are summarized as follows. Electrical characteristics and *I–V* curves of PV faults under STC are analyzed, especially for the dust deposition impact. The short-circuit current is of great guiding significance in the soiling condition, which represents the output of the cleanest module in the PV string. The distribution of real PV string normalized data further verifies diagnostic features with different variations, and performs characteristics of various faults under the STC.

Different from supervised machine learning, the proposed ABC-SSELM algorithm can make use of unlabeled historical data, and require only 1%–3% labeled data of the total dataset. Moreover, the generalization ability of the diagnostic model is optimized. In the hybrid simulation and experiment verification, the average accuracy improvement of the proposed ABC-SSELM is over 2.94% than the LGC in [27] and [28], over 1.26% than SSELM in [32], over 7.37% than the SAMME-CART in [10], over 69.28% than the ELM in [33], over 0.42% than the PSO-SSELM.

As for the proposed PV fault diagnostic technology, labeled fault data to be difficult to obtain can be replaced by the simulated ones. It even performs better classification accuracy, and avoids potential safety issue and additional labor cost in a large-scale PV system. Moreover, the soiling degree can be monitored by the proposed scheme for effective module cleaning. Other factors of dust that directly affect PV performance are reducing irradiation and increasing the temperature of the panel. These factors can be considered in the future research works.

REFERENCES

[1] M. K. Alam, F. Khan, J. Johnson, and J. Flicker, "A comprehensive review of catastrophic faults in PV arrays: Types, detection, and mitigation techniques," *IEEE J. Photovolt.*, vol. 5, no. 3, pp. 982–997, May 2015.

[2] D. S. Pillai and N. Rajasekar, "A comprehensive review on protection challenges and fault diagnosis in PV systems," *Renewable Sustain. Energy Rev.*, vol. 91, pp. 18–40, Aug. 2018.

[3] J. J. John, V. Rajasekar, S. Boppana, S. Chattopadhyay, A. Kottantharayil, and G. TamizhMani, "Quantification and modeling of spectral and angular losses of naturally soiled PV modules," *IEEE J. Photovolt.*, vol. 5, no. 6, pp. 1727–1734, Nov. 2015.

[4] C. Schill, S. Brachmann, and M. Koeh, "Impact of soiling on IV-curves and efficiency of PV-modules," *Sol. Energy*, vol. 112, pp. 259–262, Feb. 2015.

[5] P. D. Burton, A. Hendrickson, S. S. Ulibarri, D. Riley, W. E. Boyson, and B. H. King, "Pattern effects of soil on photovoltaic surfaces," *IEEE J. Photovolt.*, vol. 6, no. 4, pp. 976–980, Jul. 2016.

[6] J. J. John, S. Warade, G. Tamizhmani, and A. Kottantharayil, "Study of soiling loss on photovoltaic modules with artificially deposited dust of different gravimetric densities and compositions collected from different locations in India," *IEEE J. Photovolt.*, vol. 6, no. 1, pp. 236–243, Jan. 2016.

[7] 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.

[8] W. Chine, A. Mellit, V. Lughi, A. Malek, G. Sulligoi, and A. M. Pavan, "A novel fault diagnosis technique for photovoltaic systems based on artificial neural networks," *Renewable Energy*, vol. 90, pp. 501–512, May 2016.

[9] Z. Chen, L. Wu, S. Cheng, P. Lin, Y. Wu, and W. Lin, "Intelligent fault diagnosis of photovoltaic arrays based on optimized kernel extreme learning machine and IV characteristics," *Appl. Energy*, vol. 204, pp. 912–931, Oct. 2017.

[10] J. M. Huang, R. J. Wai, and W. Gao, "Newly-designed fault diagnostic method for solar photovoltaic generation system based on IV-curve measurement," *IEEE Access*, vol. 7, pp. 70919–70932, 2019. doi: 10.1109/ACCESS.2019.2919337.

[11] A. V. Joglekar and B. Hegde, "Online I-V tracer for per string monitoring and maintenance of PV panels," in *Proc. 44th Annu. Conf. IEEE Ind. Electron. Soc.*, Oct. 2018, pp. 1890–1894.

[12] C. B. Jones, B. H. Ellis, J. S. Stein, and J. Walters, "Comparative review of high resolution monitoring versus standard inverter data acquisition for a single photovoltaic power plant," in *Proc. IEEE 7th World Conf. Photovolt. Energy Convers.*, Jun. 2018, pp. 715–720.

[13] S. Roy, M. K. Alam, F. Khan, J. Johnson, and J. Flicker, "An irradiance independent, robust ground-fault detection scheme for PV Arrays based on spread spectrum time-domain reflectometry (SSTDR)," *IEEE Trans. Power Electron.*, vol. 33, no. 8, pp. 7046–7057, Aug. 2018.

[14] M. Dhimish and G. Badran, "Current limiter circuit to avoid photovoltaic mismatch conditions including hot-spots and shading," *Renewable Energy*, vol. 145, pp. 2201–2216, Jan. 2020.

[15] D. S. Pillai and N. Rajasekar, "An MPPT based sensorless line-line and line-ground fault detection technique for PV systems," *IEEE Trans. Power Electron.*, vol. 34, no. 9, pp. 8646–8659, Sep. 2019.

[16] S. Fadhel *et al.*, "PV shading fault detection and classification based on IV curve using principal component analysis: application to isolated PV system," *Sol. Energy*, vol. 179, pp. 1–10, Feb. 2019.

[17] B. P. Kumar, G. S. Ilango, M. J. B. Reddy, and N. Chilakapati, "Online fault detection and diagnosis in photovoltaic systems using wavelet packets," *IEEE J. Photovolt.*, vol. 8, no. 1, pp. 257–265, Jan. 2018.

[18] M. Dhimish, V. Holmes, and M. Dales, "Parallel fault detection algorithm for grid connected photovoltaic plants," *Renewable Energy*, vol. 113, pp. 94–111, Dec. 2017.

[19] Z. Chen *et al.*, "Random forest based intelligent fault diagnosis for PV arrays using array voltage and string currents," *Energy Convers. Manage.*, vol. 178, pp. 250–264, Dec. 2018.

[20] Z. Yi and A. H. Etemadi, "Line-to-line fault detection for photovoltaic arrays based on multiresolution signal decomposition and two-stage support vector machine," *IEEE Trans. Ind. Electron.*, vol. 64, no. 11, pp. 8546–8556, Nov. 2017.

[21] Z. Yi and A. H. Etemadi, "Fault detection for photovoltaic systems based on multi-resolution signal decomposition and fuzzy inference systems," *IEEE Trans. Smart Grid*, vol. 8, no. 3, pp. 1274–1283, May 2017.

[22] A. Belaout, F. Krim, A. Mellit, B. Talbi, and A. Arabi, "Multiclass adaptive neuro-fuzzy classifier and feature selection techniques for photovoltaic array fault detection and classification," *Renewable Energy*, vol. 127, pp. 548–558, Nov. 2018.

[23] M. Dhimish, V. Holmes, B. Mehrdadi, and M. Dales, "Comparing Mamdani Sugeno fuzzy logic and RBF ANN network for PV fault detection," *Renewable Energy*, vol. 117, pp. 257–274, Mar. 2018.

[24] P. Lin, Y. Lin, Z. Chen, L. Wu, L. Chen, and S. Cheng, "A density peak-based clustering approach for fault diagnosis of photovoltaic arrays," *Int. J. Photoenergy*, vol. 2017, no. 9, 2017, Art. no. 4903613.

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

[26] S. Liu, L. Dong, X. Liao, Y. Hao, X. Cao, and X. Wang, "A dilation and erosion-based clustering approach for fault diagnosis of photovoltaic arrays," *IEEE Sensors. J.*, vol. 19, no. 11, pp. 4123–4137, Jun. 2019.

[27] Y. Zhao, R. Ball, J. Mosesian, J. F. 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.

[28] H. Momeni, N. Sadoogi, M. Farrokhifar, and H. F. Gharibeh, "Fault diagnosis in photovoltaic arrays using GBSSL method and proposing a fault correction system," *IEEE Trans. Ind. Informat.*, to be published, doi: 10.1109/TII.2019.2908992.

[29] T. Joachims, "Transductive inference for text classification using support vector machines," in *Proc.16th Int. Conf. Mach. Learn.*, San Francisco, CA, USA, 1999, pp. 200–209.

[30] O. Chapelle, M. Chi, and A. Zien, "A continuation method for semi-supervised SVMs," in *Proc. 23rd Int. Conf. Mach. Learn.*, New York, NY, USA, 2006, pp. 185–192.

[31] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Proc. 16th Adv. Neural Inf. Process. Syst.*, Dec. 2003, pp. 321–328.

[32] G. Huang, S. Song, J. N. D. Gupta, and C. Wu, "Semi-supervised and unsupervised extreme learning machines," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2405–2417, Dec. 2014.

[33] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Trans. Syst., Man., Cybern. B, Cybern.*, vol. 42, no. 2, pp. 513–529, Apr. 2012.

[34] J. S. C. M. Raj and A. E. Jeyakumar, "A novel maximum power point tracking technique for photovoltaic module based on power plane analysis of I-V characteristics," *IEEE Trans. Ind. Electron.*, vol. 61, no. 9, pp. 4734–4745, Sep. 2014.

[35] W. Zong, G.-B. Huang, and Y. Chen, "Weighted extreme learning machine for imbalance learning," *Neurocomputing*, vol. 101, pp. 229–242, Feb. 2012.

[36] D. Karaboga, "An idea based on honey bee swarm for numerical optimization," Erciyes Univ., Kayseri, Turkey, Tech Rep. TR06, 2005.

[37] D. Karaboga and B. Basturk, "A comparative study of artificial bee colony algorithm," *Appl. Math. Comput.*, vol. 214, no. 1, pp. 108–132, Aug. 2009.

[38] P. L. Bartlett, "The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network," *IEEE Trans. Inf. Theory*, vol. 44, no. 2, pp. 525–536, Mar. 1998.

Jun-Ming Huang received the B.S. degree in electrical engineering and automation from Fujian University of Technology, Fuzhou, China, in 2017. He is currently working toward the joint M.S. degree with Fuzhou University, China, and National Taiwan University of Science and Technology, Taiwan.

His research interests include machine learning and photovoltaic fault diagnosis.

Rong-Jong Wai (M'99–SM'05) was born in Tainan, Taiwan, in 1974. He received the B.S. degree in electrical engineering and the Ph.D. degree in electronic engineering from Chung Yuan Christian University, Chung Li, Taiwan, in 1996 and 1999, respectively.

From August 1998 to July 2015, he was with Yuan Ze University, Chung Li, Taiwan, where he was the Dean of General Affairs from August 2008 to July 2013, and the Chairman of the Department of Electrical Engineering from August 2014 to July 2015. Since August 2015, he has been with National Taiwan University of Science and Technology, Taipei, Taiwan, where he is currently a Distinguished Professor, the Dean of General Affairs, and the Director of the Energy Technology and Mechatronics Laboratory. He is a chapter-author of *Intelligent Adaptive Control: Industrial Applications in the Applied Computational Intelligence Set* (CRC Press, 1998) and the co-author of *Drive and Intelligent Control of Ultrasonic Motor* (Tsang-Hai, 1999), *Electric Control* (Tsang-Hai, 2002) and *Fuel Cell: New Generation Energy* (Tsang-Hai, 2004). He has authored more than 170 conference papers, 190 international journal papers, and 57 inventive patents. His research interests include power electronics, motor servo drives, mechatronics, energy technology, and control theory applications. The outstanding achievement of his research is for contributions to real-time intelligent control in practical applications and high-efficiency power converters in energy technology.

Dr. Wai received the Excellent Research Award in 2000, and the Wu Ta-You Medal and Young Researcher Award in 2003 from the National Science Council, R.O.C. In addition, he was the recipient of the Outstanding Research Award in 2003 and 2007 from the Yuan Ze University, R.O.C.; the Excellent Young Electrical Engineering Award and the Outstanding Electrical Engineering Professor Award in 2004 and 2010 from the Chinese Electrical Engineering Society, R.O.C.; the Outstanding Professor Award in 2004 and 2008 from the Far Eastern Y. Z. Hsu-Science and Technology Memorial Foundation, R.O.C.; the International Professional of the Year Award in 2005 from the International Biographical Centre, U.K.; the Young Automatic Control Engineering Award in 2005 from the Chinese Automatic Control Society, R.O.C.; the Yuan-Ze Chair Professor Award in 2007, 2010 and 2013 from the Far Eastern Y. Z. Hsu-Science and Technology Memorial Foundation, R.O.C.; the Electric Category-Invent Silver Medal Award in 2007, the Electronic Category-Invent Gold and Silver Medal Awards in 2008, the Environmental Protection Category-Invent Gold Medal Award in 2008, the Most Environmental Friendly Award in 2008, the Power Category-Invent Bronze Medal Award in 2012, and the Electronic Category-Invent Gold and Silver Medal Awards in 2015 from the International Invention Show and Technomart, Taipei, R.O.C.; the University Industrial Economic Contribution Award in 2010 from the Ministry of Economic Affairs, R.O.C; the Ten Outstanding Young Award in 2012 from the Ten Outstanding Young Person's Foundation, R.O.C.; the Taiwan Top 100 MVP Managers Award in 2012 from MANAGER today magazine, R.O.C; the Outstanding Engineering Professor Award in 2013 from the Chinese Institute of Engineers, R.O.C., the Green Technology Category-Scientific Paper Award in 2014 from the Far Eastern Y. Z. Hsu-Science and Technology Memorial Foundation, R.O.C., the Scopus Young Researcher Lead Award-Computer Science in 2014 from Taiwan Elsevier, the Outstanding Research Award in 2016 and 2018 from the National Taiwan University of Science and Technology, R.O.C., and the Most Cited Researchers Award in 2016 (Field: Electrical & Electronics Engineering). He is a Fellow of the Institution of Engineering and Technology (U.K.) and a senior member of the Institute of Electrical and Electronics Engineers (USA).

Geng-Jie Yang was born in Wuyishan, China, in 1966. He received the B.S. and M.S. degrees in electrical engineering from Fuzhou University, Fuzhou, China, in 1985 and 1988, respectively.

Since 1988, he has been with Fuzhou University, where he is currently a Professor with the Department of Electric Power Engineering. His research interests include power system analysis and control.