Solar Energy 237 (2022) 340–351

Contents lists available at ScienceDirect

Solar Energy

journal homepage: www.elsevier.com/locate/solener

Intelligent fault diagnosis of photovoltaic array based on variable predictive models and I–V curves

Yongjie Liu [a], Kun Ding [a,*], Jingwei Zhang [a], Yinghu Lin [c], Zenan Yang [a], Xiang Chen [a], Yuanliang Li [b], Xihui Chen [a]

[a] College of Mechanical and Electrical Engineering, Hohai University, Changzhou, Jiangsu 213022, China
[b] Changzhou Key Laboratory of Photovoltaic System Integration and Production Equipment Technology, Changzhou, Jiangsu 213022, China
[c] Huaneng Nanjing Jinling Power Generation Co., Ltd, Nanjing, Jiangsu 210000, China

ARTICLE INFO

Keywords:
Photovoltaic array
Current–voltage curves
Fault diagnosis
Variable predictive model
Feature selection

ABSTRACT

Photovoltaic (PV) arrays are installed in outdoors and continuously exposed to harsh environmental conditions, which are susceptible to suffer from abnormal and various kinds of faults, reduce the efficiency of power conversion and the lifetime of PV modules, and even cause electric shock and fire. In recent years, many methods of machine learning and deep learning have been successfully applied in the field of fault diagnosis for PV array. However, these methods still exist some issues: neural network training requires a large number of data samples and is prone to overfitting and underfitting; network structure and hyperparameters are difficult to determine, mainly rely on extensive experiments; The kernel extreme learning machine (KELM) and support vector machine (SVM) require kernel functions and hyper parameter adjustment, and the selection of kernel functions and hyperparameter will directly affect the accuracy of fault diagnosis; To solve the above issues, an intelligent fault diagnosis method for PV array based on the variable predictive models and I–V curves is proposed. Firstly, the impact of different operating states on the I–V curve of PV array under standard test condition (STC) is analyzed. Then, the five-dimensional effective feature variables are extracted from the key points of the measured I-V curve as input of the fault diagnosis model. Since the output of the electrical characteristics of the PV array is affected by faults and environmental factors, the extracted feature variables are normalized to eliminate the influence of environmental factors. Finally, the normalized feature variables are used as the input of variable predictive model to identify the fault types of the PV array, which reduces the computational cost and avoids the problem that the kernel function and hyperparameter are difficult to determine. Moreover, the proposed method is verified by simulation and measured data and compared with other machine learning algorithms. The results indicate that the proposed fault diagnosis method has high accuracy and reliability.

1. Introduction

In the recent report of the international energy agency (IEA), the annual renewable energy generation increased by 45% to almost 280GW in 2020 and is expected to account for 90% of the world's new installed capacity between 2021 and 2022 (IEA, 2020a). With the shortage of fossil energy and environmental pollution issues, solar energy as renewable energy is widely concerned because of its cleanliness, sustainability, and huge development potential. Photovoltaic (PV) power generation becomes one of the main ways of solar energy conversion and utilization (Gonzalo et al., 2020). Solar PV development will continue to break the records, with an expected 162GW annual incremental capacity by 2022 (IEA, 2020b).

As the essential components of the PV systems, PV arrays are installed outdoors and exposed to harsh environments, suffering from strong wind, heavy rain, high temperature, soiling, etc., during long-term operation (Mellit et al., 2018; Triki-Lahiani et al., 2018). Thus, PV arrays are susceptible to various kinds of faults or abnormalities, e.g., short-circuit or ground faults, partial shading, bypass diode fault, open-circuit fault, degradation (increased series-resistance losses or potential included degradation), etc., which may seriously reduce the power generation efficiency and a lifetime of the PV systems, and even lead to fire hazards (Pillai and Rajasekar, 2018). Therefore, the effective fault detection and diagnosis (FDD) of the PV arrays play a vital role in the operation and maintenance of the PV plants.

* Corresponding author.
E-mail address: dingk@hhu.edu.cn (K. Ding).

https://doi.org/10.1016/j.solener.2022.03.062
Received 23 July 2021; Received in revised form 15 March 2022; Accepted 25 March 2022
Available online 13 April 2022

Nomenclature

Abbreviations

AI	Artificial Intelligence
ABC	Artificial Bee Colony
ANN	Artificial Neural Network
BP	Back Propagation Network
BOA	Bayesian Optimization Algorithm
CNN	Convolutional Neural Network
DL	Deep Learning
FDD	Fault Detection and Diagnosis
EWMA	Exponential Weighted Moving Average
FDD	Fault Detection and Diagnosis
IEA	International Energy Agency
I–V	Current–Voltage
ICFFSM	Improved Code-based Fast Fault Simulation Model
KELM	Kernel Extreme Learning Machine
L	Linear model
LI	Linear Interaction Model
MPP	Maximum Power Point
PV	Photovoltaic
PNN	Probabilistic Neural Network
PSBO	Partial Shading with the Bypass Diode Open-circuit Fault
PSSC	Partial Shading with the Bypass Diode Short-circuit Fault
Q	Quadratic Model
QI	Quadratic Interaction Model

Symbols

P_{m_stc}	Maximum Power under STC
V_{m_stc}	Voltage at Maximum Power Point under STC
I_{m_stc}	Current at Maximum Power Point under STC
α	Temperature Coefficient of Current
β	Temperature Coefficient of Voltage

In order to effectively detect and diagnose PV array faults, a variety of FDD methods have been proposed in recent years. According to the type of data collected from the monitoring system, they are mainly divided into three conditions. The first category is the FDD method based on infrared images obtained by using the unmanned aerial vehicle (UAV) with infrared equipment. The infrared images are effective to detect and locate defects such as hot-spots in PV modules (Claudia et al., 2016; Márquez and Segovia, 2019; Akram et al., 2020). However, it is not suitable for small PV systems because of the high cost of UAVs and high-precision infrared imagers. The second category of the FDD method is based on the operational data including current, voltage, or power at the maximum power point (MPP) of the PV array. The deviation between the reference operating current and voltage of the mathematical model and the measured data is usually calculated. The corresponding threshold is set and compared with the deviation to detect faults or abnormalities (Ding et al., 2019; Fezai et al., 2019; Harrou et al., 2019; Zhu et al., 2019; Yc et al., 2019). However, the operational data contains less information on the PV array, which is difficult to identify various fault types and the fault diagnosis accuracy is poor. Besides, the corresponding threshold is difficult to determine. If the threshold is set too high, the faults may be undetected. If the alarm threshold is set too low, the probability of misdiagnosis would increase. The third category is based on the measured current–voltage (I–V) curves of the PV array, which is also the main research object of this paper. Compared with the second category of fault diagnosis method, the I–V curves can provide more operation information of the PV array, including the short-circuit current, open-circuit voltage, current of maximum power point, the voltage of maximum power point, etc.

In recent years, more and more PV inverters have integrated the I–V curve scanning function to make the method feasible for real-time fault diagnosis. With the development of artificial intelligence (AI) technology, machine learning (ML) and deep learning (DL) have been widely used. More and more researchers focus on extracting feature variables from I–V curves as input of ML models for FDD (Chen et al., 2017; Guo and Zetao, 2019; Chine et al., 2016; Harrou et al., 2018; Chen et al., 2019; Sun et al., 2019; Eskandari et al., 2020; Appiah et al., 2019). In Zhu et al. (2019), the fault diagnosis model is established by optimizing the kernel extreme learning machine (KELM) to identify four typical faults, including degradation, short-circuit fault, open-circuit fault, and partial shading. In Guo and Zetao (2019), feature variables are extracted from the I–V curve as the input of the support vector machine (SVM) to establish the fault diagnosis model. In Chine et al. (2016), the three-layer feed-forward artificial neural network (ANN) is established to diagnose short-circuit fault and partial shading conditions, based on the I–V curves of a PV array. In Chen et al. (2019), the deep residual network (ResNet) is explored as a simplified multi-classifier to detect and diagnose the normal and different faults of the PV array by using the output I–V curves and corresponding ambient conditions. In Appiah et al. (2019), the original current and voltage signals of the PV array are converted into two-dimensional sequence diagrams and used as the input of a convolutional neural network (CNN) to train the neural network for fault diagnosis. It is noteworthy the above methods have the following limitations:

(1) Neural network requires a number of data samples to train, and the generalization of networks is difficult to guarantee. Besides, the computing cost is increased due to the neural network's iterative learning process;

(2) The network structure and hyperparameters are difficult to determine, which mainly relies on expert experience and extensive experiments;

(3) KELM and SVM both require strict kernel function and hyper-parameter adjustment, which will directly affect the accuracy of fault diagnosis.

To solve the above problems, the optimization algorithms are used to automatically adjust the network structure and hyperparameters. In Xla et al. (2019), the artificial bee colony (ABC) algorithm is applied to optimize the semi-supervised extreme learning machine (SSELM) as a new pattern recognition method for PV array fault diagnosis based on the I–V curves. In Huang et al. (2019), the Bayesian optimization algorithm (BOA) is adopted to achieve the hyperparameter optimization of the semi-supervised ladder network (SSLN) to identify various faults, i.e., line–line faults, open-circuit fault, and partial shading. Unfortunately, these methods greatly increase the computational cost during the iterative optimization process. Moreover, the above-mentioned fault diagnosis methods only discuss the single fault type. The concurrent faults, e.g., the partial shading with the bypass diode open-circuit fault (PSBO), and the partial shading with the bypass diode short-circuit fault (PSSC), etc., are not considered.

Addressing the above issues, a fault diagnosis method of PV array based on the variable prediction model-based class discrimination (VPMCD) by using the output I–V curves is proposed in this paper for the first time to automatically identify the PV array faults, including three typical faults, e.g., degradation (increased series resistance loss), short-circuit fault, partial shading and two concurrent faults including PSBO and PSSC. The proposed VPMCD diagnostic model is a new AI technique based on the interaction between feature variables, which

Table 1
Specification of PV module TMS-240.

Parameters	Symbol	Value
Maximum power	P_{m_stc}	240 W
Voltage at maximum power point	V_{m_stc}	29.7 V
Current at maximum power point	I_{m_stc}	8.1 A
Open circuit voltage	V_{oc_stc}	37.3 V
Short circuit current	I_{sc_stc}	8.62 A
Temperature coefficient of current	α	0.047%/°C
Temperature coefficient of voltage	β	−0.32%/°C

has no preset hyperparameters and avoids the complex iterative process of neural networks. The determinations of hyperparameters and network structure are avoided. Firstly, the impact of different faults on the I–V characteristic curves is analyzed and the corresponding feature variables are extracted from the measured I–V characteristic curves. Secondly, the extracted feature variables are normalized to eliminate the influence of irradiance and temperature, so that the electrical features of the PV array are only affected by different faults. Then, the extracted feature variables are used to establish VPMCD for FDD of PV array. Moreover, the simulation and measured data samples are used to verify the fault diagnosis accuracy of the proposed model. In addition, the performance of the proposed fault diagnosis method is evaluated in comparisons with other conventional ML methods, including back propagation network (BP), SVM, ELM, probabilistic neural network (PNN).

The rest of this study is organized as follows. In Section 2, the impact of different faults on the I–V characteristic curves is analyzed and the corresponding feature variables are extracted from the measured I–V characteristic curves. In Section 3, the feature variables normalization and the new pattern recognition algorithm VPMCD are described, then the diagnostic procedure of the proposed PV array FDD model based on the VPMCD is expressed in detail. In Section 4, the performance of the proposed fault diagnosis method is verified by simulation and experimental data, and compared with other ML algorithms. In Section 5, some significant results are concluded.

2. Simulation and analysis of different faults on the I-V curves

In the real PV array, the outdoor irradiance and temperature are uncontrollable. It is impossible to measure the I–V curves of different faults under arbitrary environmental conditions for analysis. The fault experimental data are also insufficient to verify the proposed fault diagnosis model. Therefore, in this paper, the improved code-based fast fault simulation model (ICFFSM) of the PV array is used to generate the fault data samples (Liu et al., 2020). Then, the influence of different faults on the shape of the I–V curve and electrical characteristics of the PV array is analyzed and used to extract the feature variables.

The capacity of the PV array investigated in this paper is 5.28 kW. This PV array is a string composed of 22 multi-crystalline silicon PV modules (TSM-240) connected in series. The specification of the PV module TSM-240 under STC provided by the manufacturer is shown in Table 1. The PV module consists of 3 bypass diodes, and each bypass diode is connected with 20 solar cells in anti-parallel. The I–V curves of the PV array are measured by the inverter GW20KN-DT and stored with the corresponding measured co-plane irradiance and temperature data in 2 min time interval. In this study, three typical faults, including short-circuit fault, partial shading, and degradation (increased series resistance loss), and two concurrent faults, including PSBO, PSSC, are investigated. The influences of independent and concurrent faults on the I–V curve under STC are shown in Fig. 1.

Short-circuit faults in a PV string are commonly caused by the accidental connections between PV modules or cables at two different potentials, which will cause the significant reduction of the open-circuit voltage (V_{oc}), as is shown in Fig. 1. Partial shading refers to the nonuniform of input irradiance on the plane of the PV array

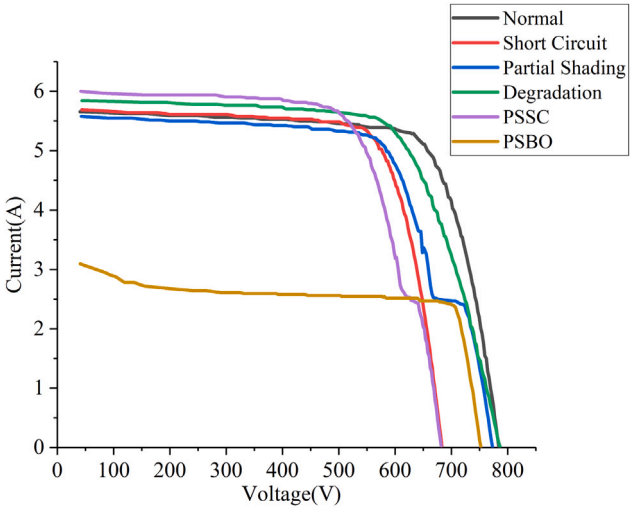


Fig. 1. I–V curves of independent and concurrent fault conditions under STC.

caused by bird droppings, dust, trees, and buildings, which will cause the I–V curve to be distorted and the current at maximum power point (MPP) to decrease. Besides, the partial shading may lead to an abnormal increase in the temperature of the shaded part to form hot-spots, which will seriously affect the life of PV modules (Luz et al., 2020). Degradation may be caused by many factors such as the degradation of series-connected cable, mechanical damage, harsh operating conditions, e.g., strong ultraviolet, high temperature, and oxidation, which will cause discoloration and delamination of the PV modules. These degradation phenomena will cause the series resistance (R_s) to increase and the parallel resistance (R_{sh}) to decrease (Hocine et al., 2021). In this paper, the degradation phenomenon of the increased series resistance R_s is mainly focused on, since it usually occurs and has a greater influence on the efficiency of the PV modules. The increased series resistance losses are caused by many factors, e.g. the degradation of contact material on the front and back surfaces of PV modules, the degradation of series cables, the contact resistance at each connection, the solder corrosion or disconnections (Spataru et al., 2015; Kohno et al., 2019). These factors will lead to an increase of the equivalent series resistance in the PV array. As the increase of R_s, the slope of the I–V curve between the open-circuit voltage V_{oc} and the MPP is changed, and maximum power is reduced, as is shown in Fig. 1. The concurrent fault of the PSBO is the partial shading with the open-circuit bypass diode, which will cause the reduced short-circuit current (I_{sc}) of the I–V curve. Besides, the remaining concurrent faults are combined by the above individual fault. After the above analysis, the open-circuit voltage V_{oc}, short-circuit current I_{sc}, series resistance R_s, current at the maximum power point I_m, and voltage at the maximum power point V_m of the I–V curve are used as feature variables of fault diagnosis. The influence of different faults on the feature variables of the I–V curve is analyzed, as shown in Table 2. The different faults have different effects on the selected five feature parameters, which further reveals the feasibility of using the selected feature parameters as the input to the fault diagnosis model. Furthermore, the feature parameters can improve the accuracy of fault diagnosis and the simple calculation of the feature parameters can speed up the fault diagnosis and improve the real-time performance of the fault diagnosis model.

The series resistance R_s is equivalently calculated as shown in (1) (Christy et al., 2014):

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

where (V_1, I_1) is the point closest to the open-circuit point (V_{oc}, 0). Due to the measurement error in the experiments, the average value of

Table 2
Effects on feature variables of I–V curve under different fault conditions.

Fault type	V_{oc}	I_{sc}	V_m	I_m	R_s
Normal	–	–	–	–	–
Short circuit fault	↓	–	↓	–	–
Partial shading	–	–	↓	↓	–
Degradation	–	–	↓	–	↑
PSSC	↓	–	↓	↓	–
PSBO	↑	↓	↑	↓	–

the three estimated values by (1) is used instead:

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

where (V_1, I_1), (V_2, I_2), and (V_3, I_3) are the three closest points of the I–V curve to the open-circuit point (V_{oc}, 0).

3. Variable prediction model-based class discrimination for fault detection and diagnosis of PV array

The proposed fault diagnosis method of the PV array based on the variable prediction model-based class discrimination (VPMCD) and I–V curves in this study mainly includes feature parameter normalization and the diagnosis approach of the VPMCD. Feature variables are extracted from the measured I–V curves and normalized to eliminate the effect of irradiance and temperature. The VPMCD is used for PV array fault diagnosis, which simplifies the determination of the network structure and the hyperparameters.

3.1. Normalization of feature variables

The electrical characteristic of the PV array is directly affected by the ambient conditions. At first, the feature variables are normalized to dimensionless numbers between [0-1] to eliminate the effects of the irradiance, temperature. Then, the normalized feature variables only represent the corresponding electrical characteristic of the PV array under different operational conditions with the same weight for expressing the character of the I–V curve. The feature variables V_{oc}, I_{sc}, I_m, V_m, R_s of the I–V curves are normalized to form five-dimensional fault diagnostic features, as shown in (3)-(7) (Huang et al., 2019):

$$V_{oc\_n} = V_{oc}/V_{oc\_stc}\left[1+\beta\left(T-T_{stc}\right)\right] \tag{3}$$

$$I_{sc\_n} = I_{sc}\frac{G_{stc}}{G}/I_{sc\_stc}\left[1+\alpha\left(T-T_{stc}\right)\right] \tag{4}$$

$$I_{m\_n} = I_m\frac{G_{stc}}{G}/I_{m\_stc}\left[1+\alpha\frac{I_{m\_stc}}{I_{sc\_stc}}\left(T-T_{stc}\right)\right] \tag{5}$$

$$V_{m\_n} = V_m/V_{m\_stc}\left[1+\beta\frac{V_{m\_stc}}{V_{oc\_stc}}\left(T-T_{stc}\right)\right] \tag{6}$$

$$R_{s\_n} = \frac{R_s G}{R_{s\_stc}G_{stc}} \tag{7}$$

where G is the measured in-plane irradiance. G_{stc} is the reference irradiance under STC, i.e., 1000 W/m^2. α is the temperature coefficient of short-circuit current. β is the temperature coefficient of open-circuit voltage. T is the measured temperature of the back sheet of PV module. T_{stc} is the reference temperature under STC, i.e., 25 °C. V_{oc_stc}, I_{sc_stc}, I_{m_stc}, V_{m_stc} represent the open-circuit voltage, short-circuit current, current and voltage at the maximum power point under STC, respectively. R_{s_stc} is the equivalent series resistance of the PV array under STC, and the value can be obtained by the specification variables in Table 1 based on the PV module simulation module in MATLAB/Simulink (Zhang et al., 2020). V_{oc}, I_{sc}, I_m, V_m, R_s represent the open-circuit voltage, short-circuit current, maximum power point current, maximum power point voltage, and equivalent series resistance under any measured conditions, respectively. V_{oc_n}, I_{sc_n}, I_{m_n}, V_{m_n}, R_{s_n} represent the normalized feature variables.

3.2. Variable prediction model-based class discrimination

VPMCD is a new pattern recognition algorithm, which establishes a predictive model using the internal relationship between the feature values. Then, the established predictive model is used to predict and discriminate unclassified data samples. The established predictive model contains nonlinear models, so it can be effectively applied to solve nonlinear and small sample problems (Rao and Lakshminarayanan, 2008).

VPMCD does not require preset hyperparameters, the determination of the network structure, kernel function, and hyperparameters of conventional CNN, SVM, or KELM is avoided. Additionally, the VPMCD method establishes the prediction model without iteration and parameter optimization by estimating parameters during the training process. Thus, the cost of computation is significantly reduced. At present, VPMCD has been widely adopted in protein structure identification and mechanical fault diagnosis. However, the application in the fault diagnosis of PV arrays is not investigated.

VPMCD mines the internal relationships between different feature variables by establishing variable predictive models (VPMs). Because these correlations vary among different conditions of failure, so the established VPMs are used to predict and classify data samples in unknown states. Assuming that the d-dimensional feature $\boldsymbol{X} = [X_1, X_2, X_3, \ldots, X_d]$ is extracted from the I–V curve to describe the operating state of the PV array. The one-to-one or many-to-one mapping among the d feature values may exist. Then, the VPMs are established based on the corresponding relationship between the feature variables of each fault type. In VPMCD, the arbitrary feature variable $X_i (i = 1, 2, \ldots, d)$ can be represented by the following four linear or nonlinear regression models VPM_i, including linear model (L), linear interaction model (LI), quadratic model (Q) and quadratic interaction model (QI). These models are expressed as follows (Bo et al., 2020):

(1) L VPM:

$$X_i = b_0 + \sum_{j=1}^{r} b_j X_j \tag{8}$$

(2) LI VPM:

$$X_i = b_0 + \sum_{j=1}^{r} b_j X_j + \sum_{j=1}^{r}\sum_{k=j+1,k\neq i}^{r} b_{jk} X_j X_k \tag{9}$$

(3) Q VPM:

$$X_i = b_0 + \sum_{j=1}^{r} b_j X_j + \sum_{j=1}^{r} b_{jj} X_j{}^2 \tag{10}$$

(4) QI VPM:

$$X_i = b_0 + \sum_{j=1}^{r} b_j X_j + \sum_{j=1}^{r} b_{jj} X_j^2 + \sum_{j=1}^{r}\sum_{k=j+1,k\neq i}^{r} b_{jk} X_j X_k \tag{11}$$

where r is the model order, $r \le d-1$, and b_0, b_j, b_{jj}, b_{jk} are the variables of the VPM. X_i is the predicted variable. $X_j (j \neq i)$ is the predictor variable. Using the above four models, the feature variable $X_j (j \neq i)$ is used to predict X_i, which can be expressed as follows:

$$X_i = f\left(X_j, b_0, b_j, b_{jj}, b_{jk}\right) + e \tag{12}$$

where e represents the deviation between the measured value and the predicted value.

3.3. Fault diagnosis approach based on the VPMCD for PV array

In this paper, VPMCD is utilized to diagnose the independent and concurrent faults of the PV array. The overall framework of fault diagnosis is shown in Fig. 2, which mainly includes data acquisition of I–V curves, feature variables extraction, feature variables normalization, fault diagnosis model establishment and application. The detailed fault diagnosis procedures are as follows:

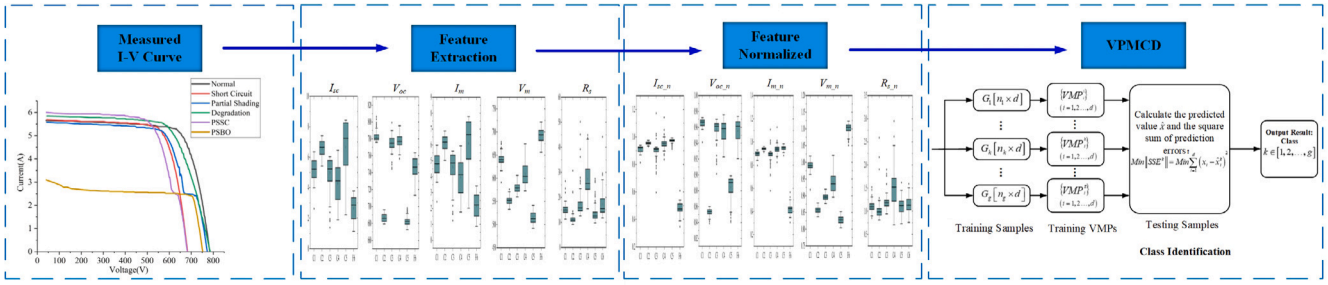


Fig. 2. Framework of proposed VPMCD-based approach for fault diagnosis of PV array.

(1) Feature variables extraction: The original I–V curves are obtained from the ICFFSM simulation model or from a real PV array based on the inverter I–V scan function. . Then,effective feature variables are extracted from the I–V curve to characterize the operating state of the PV array, including the open-circuit voltage V_{oc}, short-circuit current I_{sc}, series resistance R_s, current, and voltage at the maximum power point I_m, V_m. The feature variables (V_{oc}, I_{sc}, I_m, V_m) can be directly searched from the I–V curves. The feature variable R_s can be obtained by simple analytical computation. The detail is described in Section 2.

(2) Feature variables normalization: The original feature variables are not only affected by different faults but also by the environment. Thus, extracted feature variables are normalized to eliminate the influences of irradiation, temperature and the method is presented in Section 3.1. Only the effects of different faults on the output electrical characteristics of the PV array are considered.

(3) Fault diagnosis model establishment (VPMs training): The above normalized five-dimensional feature variables are used as input to the VPMCD model and the training is completed to obtain the VPM matrix. If g fault types may occur, the sample number of each fault type is $n_k(k = 1, 2, \dots, g)$, and the total number of samples is $n = \sum n_k$; $\boldsymbol{X} = [X_1, X_2, X_3, \dots, X_d]$ is the sample feature, and d is the sample feature dimension; $N[n \times d]$ is the total training sample set, and $N_k[n_k \times d]$ is the training sample set of each fault type. For each feature value X_i, the L, LI, Q, and QI models and the model order r are used to establish the VPM, and C_{d-1}^{r} types of VPM_i^k can be obtained. Then, VPMs are used to predict X_i to get the predicted $\overline{X_i}$. The minimum quadratic sums of deviation between the predicted and measured feature values are used to determine the optimal VPM. Then, the VPM matrix $VPM[g \times d]$ of all training data samples are calculated as follows:

$$VPM = \begin{bmatrix} VPM_1^1 & VPM_2^1 & \cdots & VPM_d^1 \\ VPM_1^2 & VPM_2^2 & \cdots & VPM_d^2 \\ \vdots & \vdots & & \vdots \\ VPM_1^g & VPM_2^g & \cdots & VPM_d^g \end{bmatrix} \tag{13}$$

(4) Fault diagnosis model application (Fault diagnosis): the feature variables $\boldsymbol{X} = [X_1, X_2, X_3, \dots, X_d]$ of test samples are extracted. Using the established matrix $VPM[g \times d]$ described above, test sample features $X_i(i = 1, 2, \dots, d)$ are predicted to obtain the predicted value $\overline{X_i}^k$. The sum of squares of prediction errors SSE^k of all feature values in the same type are calculated, and the minimum SSE is used as the discriminant function for classification. As shown in (14), when SSE^k is the smallest, the test sample is identified as the kth type.

$$\min \left\| SSE^k \right\| = \min \sum_{i=1}^{d} \left(X_i - \overline{X}_i \right)^2 \tag{14}$$

4. Experimental verification and analysis

In order to verify the performance of the proposed method, the normal and three typical abnormalities, including short-circuit fault, partial shading, degradation are investigated. Moreover, the two concurrent faults, PSSC and PSBO, are also considered, as described in Section 1. Firstly, the simulation and real datasets are obtained based on the simulation model and real experimental platform to evaluate the performance of the proposed VPMCD-based fault diagnosis method. In order to verify the performance of the fault diagnosis model under various ambient conditions, the simulation model is applied to obtain the data samples in the large range of ambient conditions for each fault. Then, the statistical distribution of normalized feature variables under different faults is analyzed. Finally, the proposed method is compared with other ML algorithms by the simulated and measured data sets, and the comparison results are discussed.

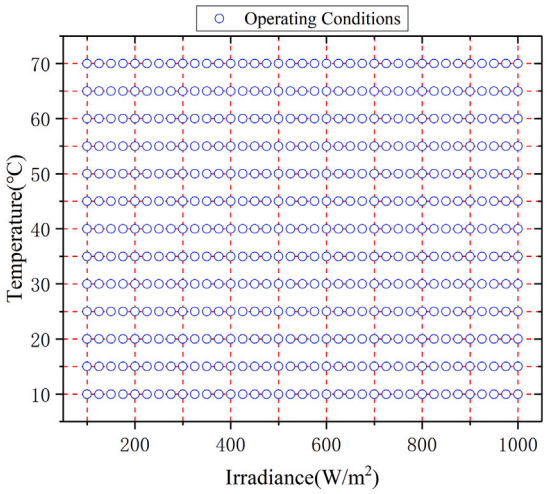
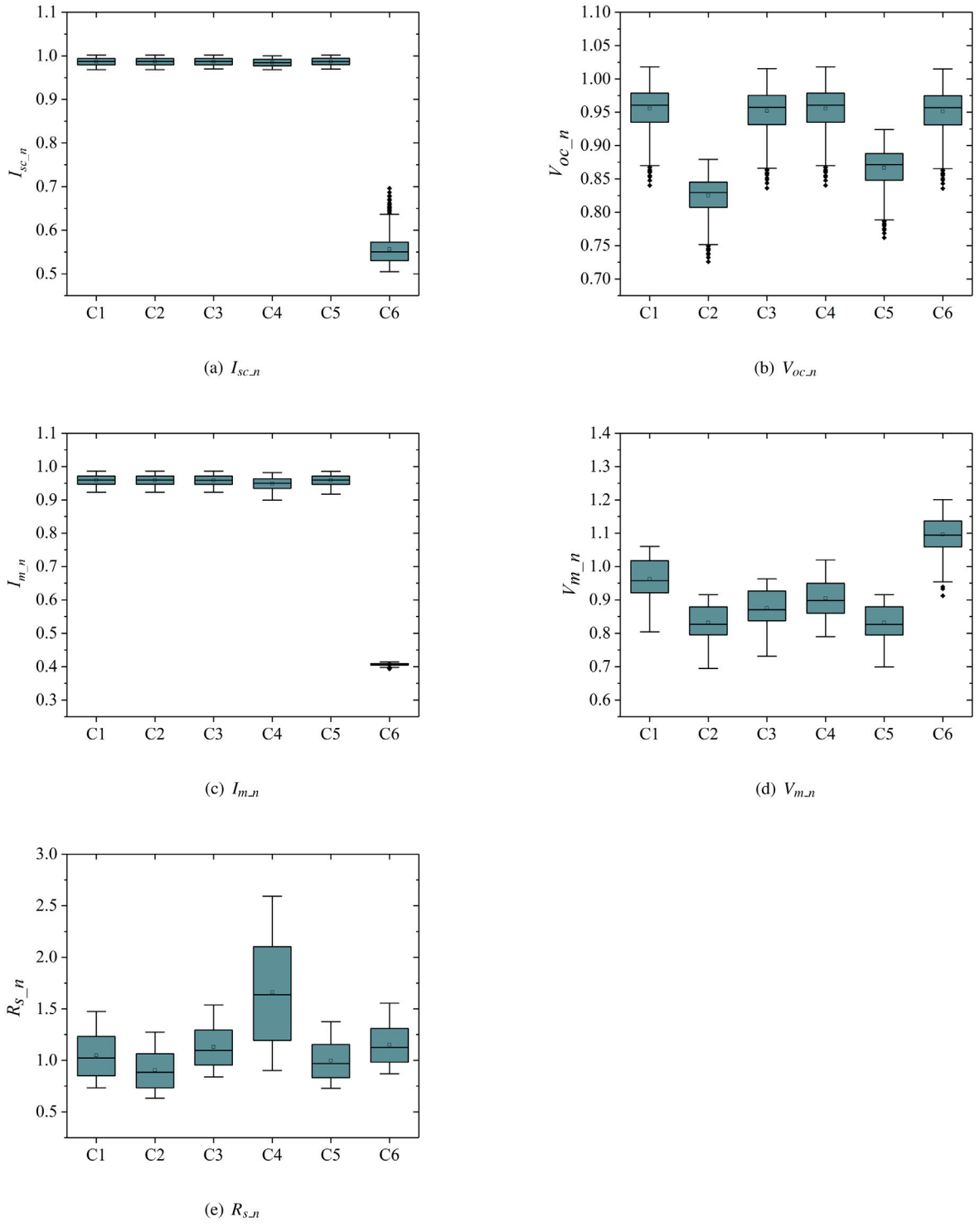


Fig. 3. Operation conditions of grid sampling.

Besides, in this study, the proposed model and other ML algorithms are completed using the MATLAB script and executed on the computer with Intel(R) Core(TM) i7-9750H CPU @ 2.60 GHz, 16GB(RAM), and 64 bits operating system.

4.1. Acquisition of the simulated data

At first, the fault simulations are performed to verify the generalization of the proposed FDD model. The operation data samples covering a wider range of ambient conditions are obtained by the simulation model. In the simulation, the irradiance range is set from 100 W/m^2 to 1000 W/m^2 with the interval of 25 W/m^2, and the temperature range is from 10 °C to 70 °C with the interval of 5 °C. The irradiance and temperature were combined in pairs and the environmental operating conditions are shown in Fig. 3. Then, the different faults are configured to obtain the corresponding I–V curves. In this work, the short-circuit of three PV modules is defined as short-circuit fault. The two PV modules with 45% irradiance reduction of the PV array is defined as partial shading. The series resistance increase of 10Ω is defined as Degradation. The concurrent fault of the PSBO is configured that the three PV modules have 50% irradiance reduction and the



(a) I_{sc_n} (b) V_{oc_n}

(c) I_{m_n} (d) V_{m_n}

(e) R_{s_n}

Fig. 4. Boxplot of five normalized feature variables on simulation data.

corresponding bypass diode open circuit. The PSSC fault is configured with the one PV modules 50% irradiance reduction and short-circuit fault of two PV modules. Finally, each fault type and normal operating state were paired to get 481 I–V curves and a total of 2886 I–V curves were obtained in the entire data sample sets.

The feature variables, including the open-circuit voltage V_{oc}, short-circuit current I_{sc}, current of maximum power point I_m, the voltage of maximum power point V_m, and equivalent series resistance R_s are extracted from I–V curves for the fault detection and diagnosis. In order to eliminate the discrepancy among irradiance, temperature and different physical quantities, the feature variables are normalized by using (3)–(7). Thus, five normalized feature variables, i.e., V_{oc_n}, I_{sc_n}, I_{m_n}, V_{m_n}, R_{s_n}, are obtained for each data sample. The statistical distribution of the normalized five feature variables is presented separately in the form of boxplot, as is shown in Fig. 4.

Fig. 5. Experimental configuration for faults simulation on the real PV array.

The horizontal label of the boxplot denotes the operating conditions of the PV array (C1: normal, C2: short-circuit, C3: partial shading, C4: degradation, C5: PSSC, C6: PSBO); The following phenomenon can be observed in Fig. 4. The normalized feature variables eliminate the effects of meteorological factors compared to the newly extracted features, and make the feature variables under the same operating conditions belong to the same class of clusters, and the distribution is more intensive. Then, the distribution characteristics of the feature variables under different operating conditions are analyzed. The I_{sc_n} of the PSBO fault is obviously less than that of the other five conditions. The V_{oc_n} of the short-circuit fault and PSSC concurrent fault is significantly lower than other four conditions and the I_{m_n} of the PSSC concurrent fault decreases slightly, while I_{m_n} of the short-circuit fault remains unchanged. The I_{m_n} of the PSBO is obviously lower than that of the other five conditions. The V_{m_n}, of the PSBO is higher than normal operating condition, the V_{m_n}, of the other operating conditions is reduced to varying degrees compared to normal operating condition. The R_{s_n} of the degradation fault is higher than the normal condition. The I_{m_n} and V_{m_n} of partial shading decreases slightly compared to normal condition, while other feature parameters basically remain unchanged. Therefore, it further indicates the effectiveness of using these five feature parameters for PV array fault diagnosis.

4.2. Acquisition of the experimental data

In order to verify the performance and effectiveness of the proposed fault diagnosis method, the fault experiments are implemented on the 5.28kWp PV array on the roof of the College of Mechanical and Electrical Engineering in Hohai University, Changzhou Campus (Latitude 31.82° N, Longitude 119.98° E). The PV array is a PV string formed by 12 240 W mc-Si PV modules connected in series. Different fault experiments of the PV array are shown in Fig. 5, which are the same as the simulation model mentioned in Section 3.1. The partial shading is implemented by shading the PV module with plastics films of different transmittances. In this work, the partial shading is emulated by two shaded PV modules with 45% irradiance reduction. The short-circuit fault is tested by short-circuiting PV modules with wires. The short circuit fault is emulated by short-circuiting three PV modules. The degradation fault is implemented by connecting a high-power sliding rheostat in series with the PV modules. The increase of the series resistance by 10 Ω is defined as degradation. The open-circuit fault of the bypass diodes is implemented by disconnecting the bypass diode in the junction box of the PV module. The corresponding concurrent faults are tested by combining the above faults. One of the concurrent faults PSBO is emulated by two shaded PV modules with 50% irradiance reduction and corresponding bypass diodes are open-circuit. Another concurrent fault PSSC was emulated by a single shaded PV module with 50% irradiation reduction and two modules being short-circuited. The fault experiments are carried out intermittently between January 10, 2021, and March 25, 2021. Then, a total of 660 data samples of the I–V curves are obtained, and 110 data samples are acquired for each fault or the normal condition. Similarly, the feature variables of the I–V curve are extracted and normalized, and the statistical distribution of the normalized feature variables, i.e., V_{oc_n}, I_{sc_n}, I_{m_n}, V_{m_n}, and R_{s_n} are also illustrated in the form of boxplot, as shown in Fig. 6. Although, the fact that the experiment data are interfered by various environmental factors, measurement error and existing outliers, the feature distributions of different operating conditions are similar to that of the simulated data.

4.3. Performance verification

In this section, the performance of the proposed PV array fault diagnosis method based on VPMCD is verified based on simulated and measured data for six different operating conditions. Firstly, 70% of data samples are randomly selected from each fault type of the dataset as the training set, and the remaining 30% are used as the testing set to evaluate the performance of the established fault diagnosis model. Moreover, the proposed fault diagnosis method based on VPMCD for PV array is compared with other machine learning algorithms based on simulated and measured datasets in diagnostic accuracy and computational time, including Back Propagation Neural Network (BP), Extreme Learning Machine (ELM), and Probabilistic Neural Network (PNN).

(1) Case 1: Simulation Verification

In this case, a total of 2886 simulated data samples are obtained from the improved code-based fast fault simulation model, as described in Section 4.1. In each operating state, 481 data samples are covering different irradiance and temperature ranges. To establish and test fault diagnosis model based on VPMCD for PV array, the data samples of the same operating state in the dataset are randomly divided into two subsets, including the training data set with 70% samples and the testing set with 30% samples. The training data set is used to train

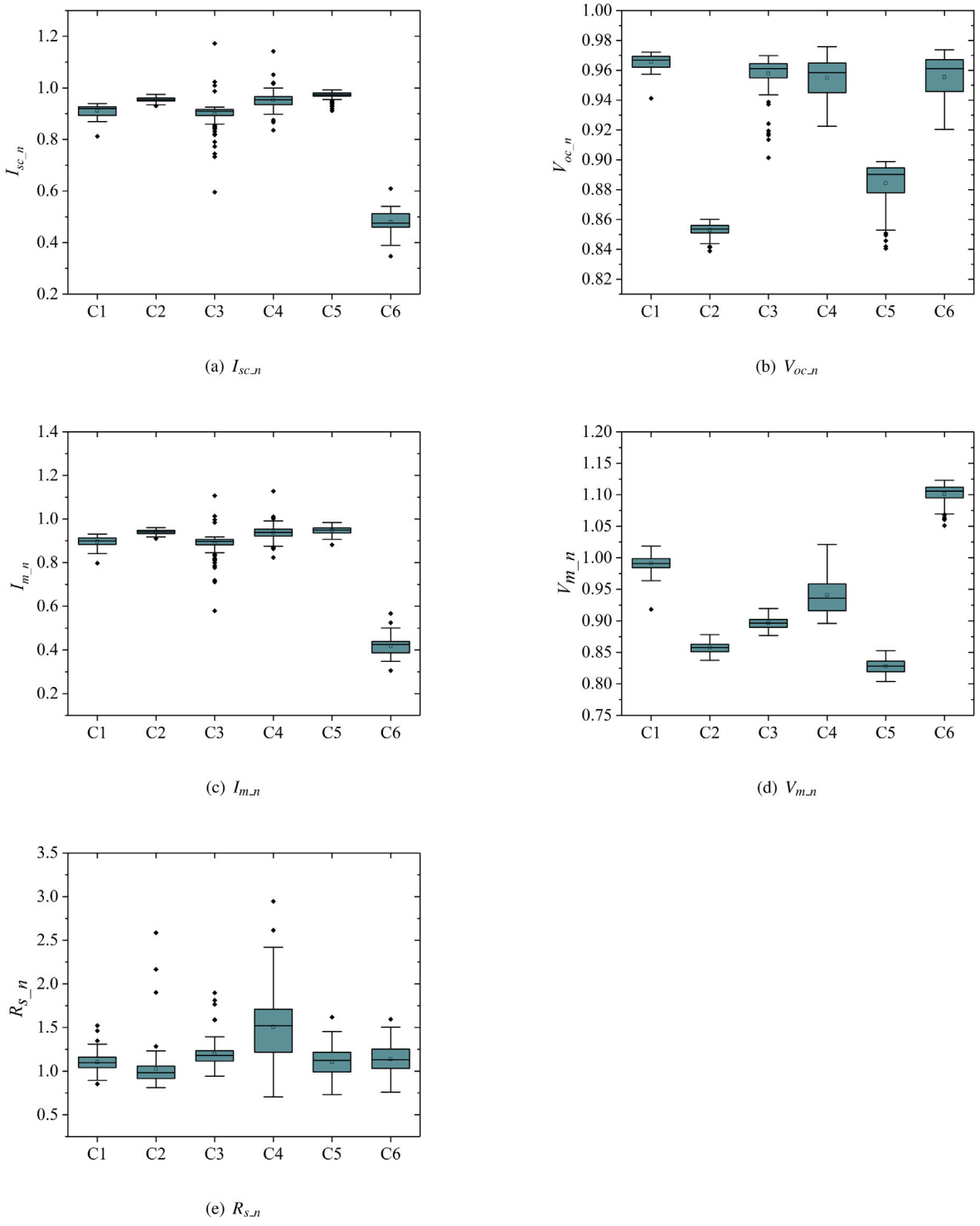


Fig. 6. Boxplot of five normalized feature variables on experimental data.

and establish variable predictive models, and the testing data set is considered as unknown new data samples to evaluate the performance of the fault diagnosis model.

Some diagnosis results based on VPMCD are demonstrated in Table 3, which presents the square sum of prediction errors and prediction labels based on the classification process of the VPMCD. It can be seen from Table 3 that the testing sample falls into the kth type when the sum of the prediction errors of all feature variables in the kth VPM is minimum. For example, the minimum square sum of prediction errors for the feature variables of the testing sample in the first-type VPM is 4.9804×10^{-4}, so the testing sample is labeled as "C1" which represents the normal operating state. Then, the predicted label is the same as the real label, indicating that the diagnosis result is correct. In addition, all data samples of the testing set are tested, and the comparison results between the real and predicted conditions are shown in Fig. 7. As illustrated in Fig. 7, the diagnosis accuracy of

Table 3
Evaluation results of prediction of non-stationary PPS in Spring.

Operation state	Real label	SSE^k						Predicted label
		VPM^1	VPM^2	VPM^3	VPM^4	VPM^5	VPM^6	
Normal	C1	**4.9804 × 10^{-4}**	0.7009	0.0216	0.0020	0.0611	3.2215 × 10^4	C1
Short circuit	C2	0.3860	**3.7201 × 10^{-4}**	0.0075	0.0110	0.0582	3.5428 × 10^4	C2
Partial shading	C3	1.9322	1.5476	**2.0673 × 10^{-4}**	0.7876	0.3170	3.5107 × 10^4	C3
Degradation	C4	0.0444	0.5270	0.9307	**9.7076 × 10^{-5}**	0.0227	3.2237 × 10^4	C4
PSSC	C5	1.2892	0.4821	0.1213	0.2208	**5.6556 × 10^{-4}**	3.5984 × 10^4	C5
PSBO	C6	8.1063 × 10^4	7.4765 × 10^4	6.7846 × 10^4	1.2779 × 10^4	5.9315 × 10^4	**1.0378 × 10^{-5}**	C6

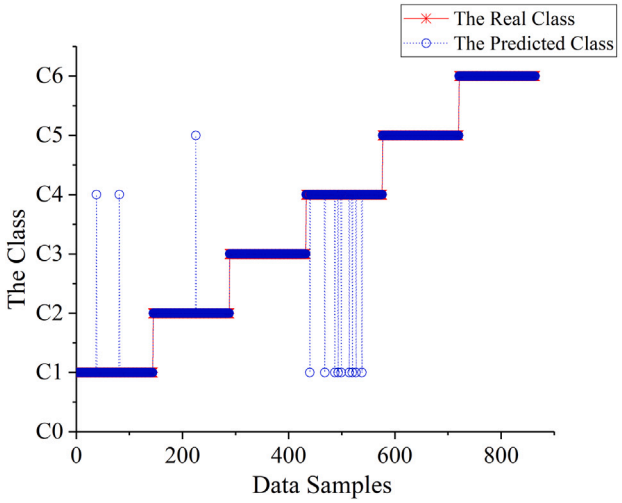


Fig. 7. The comparison of results on real and predicted conditions on the simulation dataset.

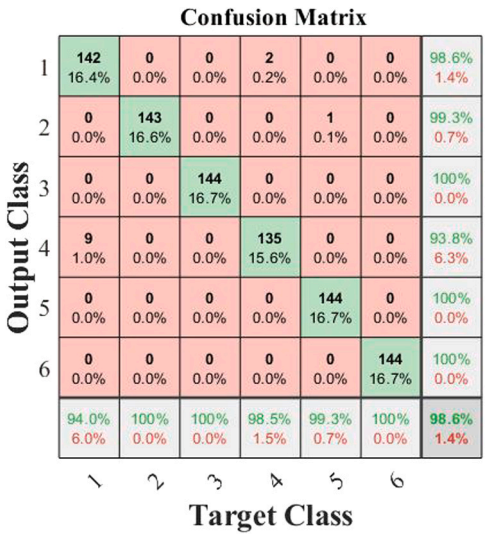


Fig. 8. The confusion matrix of the proposed method on the simulation dataset.

other operation states is 100% expect for normal (C1), short circuit fault (C2), and degradation (C4). In order to further analyze the performance of the proposed fault diagnosis method, the classification confusion matrix of the testing set is described in Fig. 8 to obtain the classification accuracy of each fault type, the misdiagnosed fault types, and the corresponding misdiagnosis rate. It can be seen from Fig. 8 that the diagnosis accuracy of degradation (C4) is the lowest 93.8%, and 1% of data samples are misdiagnosed as normal (C1). The degradation (C4) and normal (C1) are prone to be misdiagnosed because of the high similarity of the I–V curves between the normal and degradation operating states under low irradiance, which leads to the partial overlap of features, as shown in Fig. 9. It can be seen from the confusion matrix that the overall diagnostic accuracy of the testing set is 98.6%, with high diagnostic accuracy.

(2) Case 2: Experimental Verification

In the experiment, a total of 660 data samples were obtained from the real PV array, and each fault type was comprised of 110 data samples. Similar to the simulated data, 70% of the data samples were randomly selected as the training set, and 30% of the data samples were used as the test set. The comparison results of the actual and predicted conditions of the test sample set are shown in Fig. 10. Because of many interference factors in the measured data, the test samples were volatile, leading to misdiagnosis in various fault types. In order to further analyze the diagnostic performance of the fault diagnosis model, the confusion matrix of the measured data set was drawn, as shown in Fig. 11. In the measured data set, short circuit (C2) and degradation (C4) showed the lowest diagnostic accuracy of 78.6% and 81.8%, respectively. Short circuit (C2) is prone to be misdiagnosed as PSSC (C5). Similar to the simulated data, degradation (C4) and normal (C1) are susceptible to misdiagnosis. Due to the volatility of the measured data, the overall diagnostic accuracy of the measured data set is 91.4%, lower than the simulated data set.

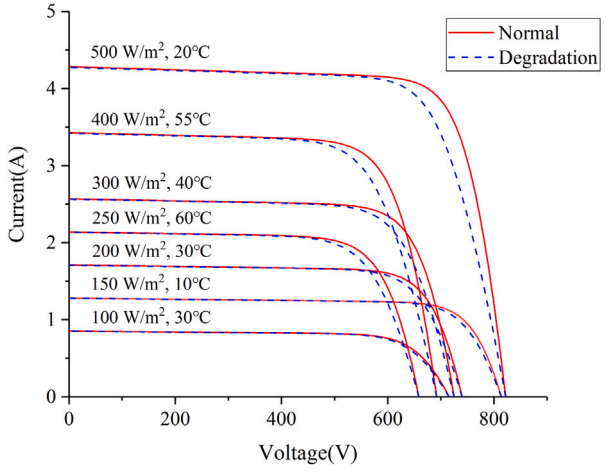


Fig. 9. The I–V curves of different environments under normal and degradation conditions.

(3) Comparative analysis

To further evaluate the performance of the proposed fault diagnosis method, the simulated and measured data sets were used to compare the diagnosis accuracy and computing time of BP, ELM, PNN, and VPMCD. In order to verify the efficiency of the fault diagnosis model under the condition of different training data samples, the training data samples were increased from 10% to 80% of the total data samples with a step size of 5%. Correspondingly, the remaining data samples were used as a testing set to examine the accuracy of the diagnostic model. Moreover, to ensure fairness and reliability, the compared three other machine learning methods are trained and tested using the same way,

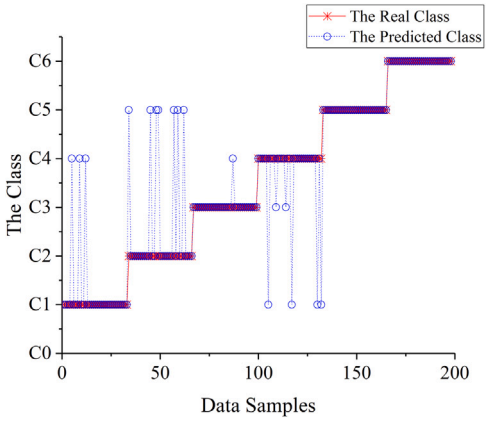


Fig. 10. The comparison of results on real and predicted conditions on the experimental dataset.

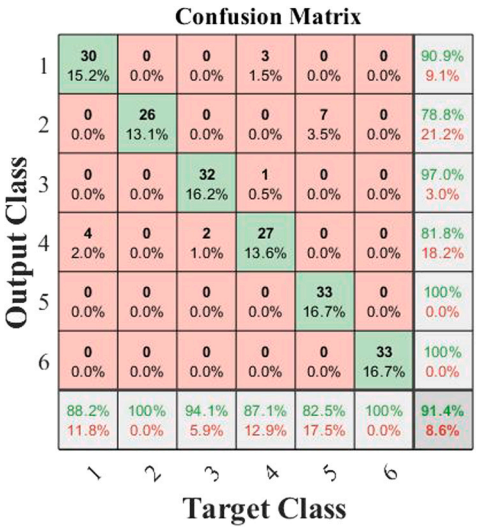


Fig. 11. The confusion matrix of the proposed method on the experimental dataset.

and each diagnostic model is independently trained and tested for 10 times, respectively. Then, the average accuracy and computing time of the testing set is used to measure the performance of the fault diagnosis method. In the simulated data set, the comparison results of the fault diagnosis accuracy and computing time of different diagnosis models are described in Fig. 12(a) and (b), respectively. Under the condition of different training samples, the average diagnostic accuracy of BP, ELM, PNN, and VPMCD is 82.24%, 92.83%, 63.77%, and 97.62%, respectively; The corresponding computing time is 0.2619s, 0.0059s, 0.9716s, and 0.2353s, respectively.

As can be seen from Fig. 12(a), the diagnostic accuracy curves of ELM, PNN, and VPMCD are relatively stable, indicating that the diagnostic accuracy of the model is stable and reliable. By contrast, the diagnosis accuracy of BP shows a more significant fluctuation. In the case of random training samples, the accuracy of VPMCD is higher than the other diagnosis models of ELM, BP, and PNN, with PNN showing the lowest fault diagnosis accuracy. It can be seen from Fig. 12(b) that the computing time of PNN is much higher than other models. The computing time of VPMCD is almost the same as the BP. The computing time of ELM is the shortest, but the diagnostic accuracy of ELM is lower than the VPMCD under the condition of random training samples on the simulated data set. As illustrated in Fig. 12(a) and (b), VPMCD outperforms other algorithms in terms of fault diagnosis performance and computation time on the simulated data set. In addition, VPMCD based fault diagnosis model is more stable and does not fluctuate significantly with the change of training data samples.

In order to verify the diagnostic performance of the fault diagnosis model in the case of measured data set and small samples, 10% to 80% of the data samples with a step size of 5% are randomly selected from the 110 data samples of each operating state on the measured data set. The remaining data samples are utilized as the testing set to evaluate the performance of various fault diagnosis models. The training set in each operating state consisted of at least 11 training samples and 99 testing samples. In the small measured data set, the comparison results of diagnostic accuracy and computing time for different fault diagnosis models are shown in Fig. 13(a) and (b), correspondingly. In the measured data sample, the average diagnostic accuracy of BP, ELM, PNN, and VPMCD is 87.93%, 90.65%, 81.97%, and 90.64%, respectively. The corresponding computing time is 0.1298s, 0.0021s, 0.5429s and 0.1475s, respectively. The overall diagnostic accuracies of VPMCD and ELM on the measured data set are lower than the simulated data set. This is due to the volatility of measured data, which reduces the overall diagnostic accuracy; Additionally, due to the decrease of data samples, the computing time is reduced compared with the simulated data, but the comparison results of computing time for different models are the same as the simulated results.

As shown in Fig. 13(a), on the measured data set, the diagnostic accuracy curve of BP still shows a comparatively large fluctuation, and PNN presents the lowest diagnostic accuracy compared with other algorithms. The diagnostic accuracy curves of ELM and VPMCD are relatively stable but have greater volatility compared with the simulation data set. The diagnostic accuracy of VPMCD on the measured data set is lower than the simulated results. However, it still has relatively high diagnosis accuracy under the case of small sample training. Moreover, the VPMCD also has a good diagnostic accuracy of 90% and a relatively short calculation time under the conditions of different training samples.

5. Conclusion

In this research, an intelligent fault diagnosis method for PV arrays based on VPMCD and I–V curves is proposed to detect and classify normal operating conditions and five fault types, including three typical faults, short circuit fault, partial shading, degradation, as well as two concurrent faults, PSSC, and PSBO. This method does not require preset variables and avoids the problem that the network structure, kernel function, and hyperparameters of CNN, SVM, and KELM are difficult to determine. There is no need for network iteration and parameter optimization during the training process, which reduces the computational cost and avoids the problems of under-fitting and over-fitting in the network training.

The impact of different operating states on the I–V curve of PV array under STC is analyzed. Then, 5-dimensional effective feature variables are extracted from the key points of the measured I–V curve as the input of the fault diagnosis model. Since the output of the electrical characteristics of the PV array is influenced by both faults and the environment, the extracted feature variables are normalized to eliminate the influence of environmental factors. The distribution characteristics of normalized feature variables are investigated for both measured and simulated data under various operational situations. Different operational conditions have different effects on feature variables, which further verifies the feasibility of using the extracted feature variables to establish the fault diagnosis model.

In the verification of simulated and measured data, the overall diagnostic accuracy of the proposed fault diagnosis model is 98.6% and 91.4%, respectively. Since the measured data is interfered with by many factors, its lower diagnostic accuracy can be understood. The

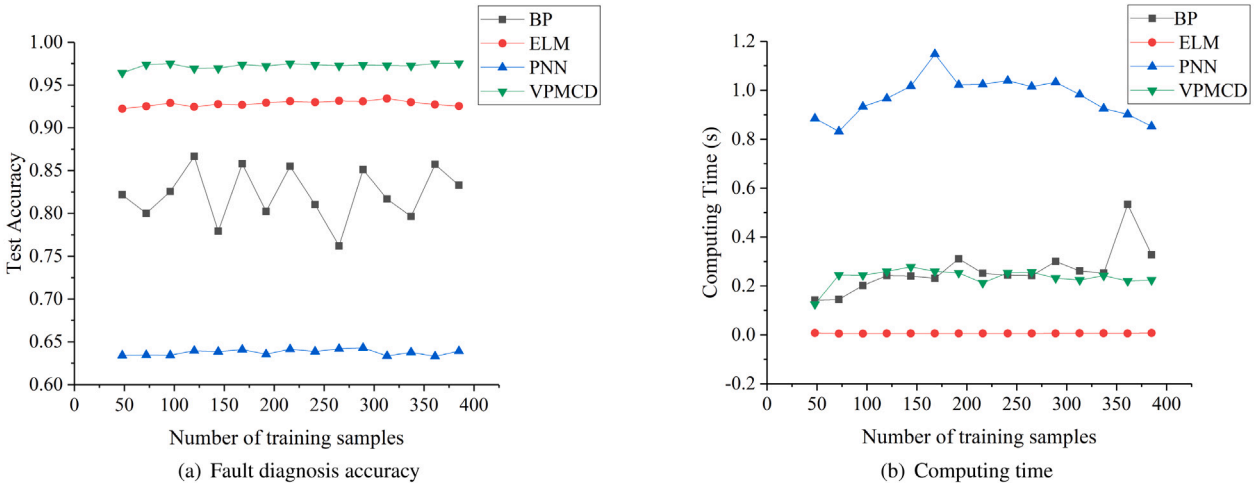


Fig. 12. Performance comparison with other machine learning algorithms on the simulation dataset.

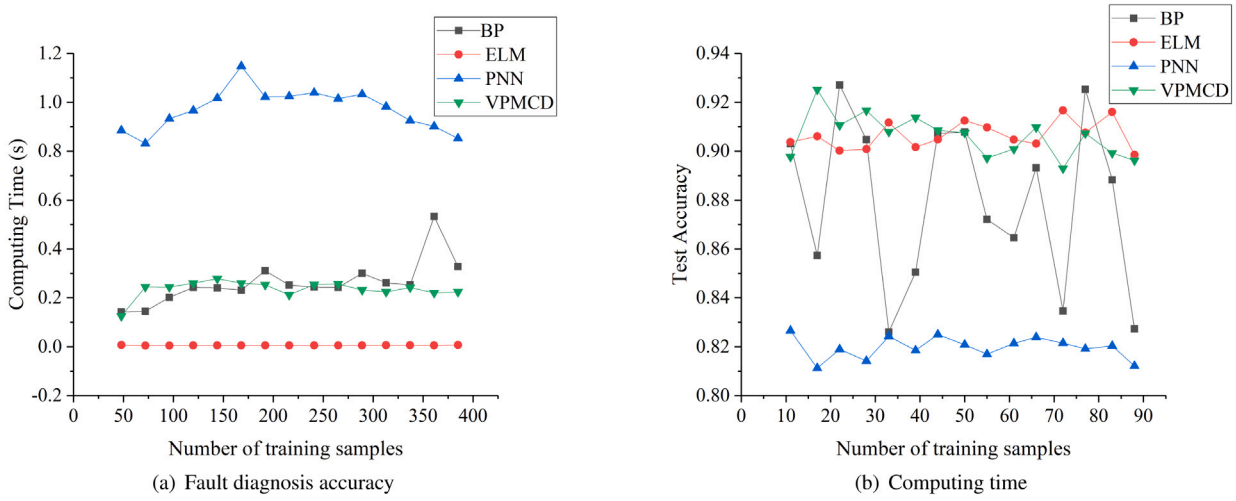


Fig. 13. Performance comparison with other machine learning algorithms on the experimental dataset.

proposed method is compared with BP, ELM, PNN, and other ML algorithms on simulated and measured data sets. The results indicate that the proposed method performs well in terms of diagnostic accuracy, small sample training, generalization, while also being computationally efficient.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX21_0464), the National Natural Science Foundation of China under Grant 51777059, Changzhou Sci & Tech Program, China (CJ20200074), the Qing Lan Project, China and the Fundamental Research Funds for the Central Universities, China (B200201051).

References

Akram, M.W., Li, G., Jin, Y., et al., 2020. Automatic detection of photovoltaic module defects in infrared images with isolated and develop-model transfer deep learning. Sol. Energy 198, 175–186.

Appiah, A.Y., Zhang, X., Ayawli, B., et al., 2019. Long short-term memory networks based automatic feature extraction for photovoltaic array fault diagnosis. IEEE Access 1.

Bo, L., Liu, X., Xu, G., 2020. Intelligent diagnostics for bearing faults based on integrated interaction of nonlinear features. IEEE Trans. Ind. Inf. 16 (2), 1111–1119.

Chen, Z., Chen, Y., Wu, L., et al., 2019. Deep residual network based fault detection and diagnosis of photovoltaic arrays using current–voltage curves and ambient conditions. Energy Convers. Manage. 198, 111793.

Chen, Z., Wu, L., Cheng, S., et al., 2017. Intelligent fault diagnosis of photovoltaic arrays based on optimized kernel extreme learning machine and I-V characteristics. Appl. Energy S0306261917305214.

Chine, W., Mellit, A., Lughi, V., et al., 2016. A novel fault diagnosis technique for photovoltaic systems based on artificial neural networks. Renew. Energy 90, 501–512.

Christy, S.J., et al., 2014. A novel maximum power point tracking technique for photovoltaic module based on power plane analysis of I-V characteristics. IEEE Trans. Ind. Electron..

Claudia, Buerhop, Long, et al., 2016. Faults and infrared thermographic diagnosis in operating c-Si photovoltaic modules: A review of research and future challenges. Renew. Sustain. Energy Rev..

Ding, K., Zhang, J., Ding, H., et al., 2019. Fault detection of photovoltaic array based on grubbs criterion and local outlier factor. IET Renew. Power Gener. 14 (4).

Eskandari, A., Milimonfared, J., Aghaei, M., 2020. Line-line fault detection and classification for photovoltaic systems using ensemble learning model based on IV characteristics. Sol. Energy 211, 354–365.

Fezai, R., Mansouri, M., Trabelsi, M., et al., 2019. Online reduced kernel GLRT technique for improved fault detection in photovoltaic systems. Energy 179 (JUL.15), 1133–1154.

Gonzalo, A.P., Marugán, A. Pliego, Márquez, F.P. García, 2020. Survey of maintenance management for photovoltaic power systems. Renew. Sustain. Energy Rev. 134.

Guo, H., Zetao, L.I., 2019. Research on fault diagnosis of photovoltaic array based on support vector machine optimized by genetic algorithm. Intell. Comput. Appl..

Harrou, F., Dairi, A., Taghezouit, B., et al., 2018. An unsupervised monitoring procedure for detecting anomalies in photovoltaic systems using a one-class support vector machine. Sol. Energy 179 (FEB.), 48–58.

Harrou, F., Taghezouit, B., Sun, Y., 2019. Improved *k*NN-based monitoring schemes for detecting faults in PV systems. IEEE J. Photovolt. 9 (3), 811–821.

Hocine, L., Samira, K.M., Tarek, M., et al., 2021. Automatic detection of faults in a photovoltaic power plant based on the observation of degradation indicators. Renew. Energy 164.

Huang, J.M., Wai, R.J., Yang, G.J., 2019. Design of hybrid artificial bee colony algorithm and semi-supervised extreme learning machine for PV fault diagnoses by considering dust impact. IEEE Trans. Power Electron. (99), 1.

IEA, 2020a. Renewables 2020 analysis and forecast to 2025, International Energy Agency [online].Available:https://www.iea.org/reports/renewables-2020/solar-pv#china.

IEA, 2020b. Renewables 2020, IEA, Paris. https://www.iea.org/reports/renewables-2020.

Kohno, T., Gokita, K., Shitanishi, H., et al., 2019. Fault-diagnosis architecture for large-scale photovoltaic power plants that does not require additional sensors. IEEE J. Photovolt. PP (3), 1–10.

Liu, Y., Ding, K., Zhang, J., Li, Y., 2020. An improved code-based fault simulation model for PV module. In: 2020 12th IEEE PES Asia-Pacific Power and Energy Engineering Conference. APPEEC, Nanjing, China, pp. 1–5.

Luz, C., Vicente, E.M., Tofoli, F.L., 2020. Experimental evaluation of global maximum power point techniques under partial shading conditions - ScienceDirect. Sol. Energy 196, 49–73.

Márquez, F.P.G., Segovia, I., 2019. Condition monitoring system for solar power plants with radiometric and thermographic sensors embedded in unmanned aerial vehicles. Measurement.

Mellit, A., Tina, G.M., Kalogirou, S.A., 2018. Fault detection and diagnosis methods for photovoltaic systems: A review. Renew. Sustain. Energy Rev. 91 (aug.), 1–17.

Pillai, D.S., Rajasekar, N.A., 2018. Comprehensive review on protection challenges and fault diagnosis in PV systems. Renew. Sustain. Energy Rev. 91 (aug.), 18–40.

Rao, R., Lakshminarayanan, S., 2008. Variable predictive model based classification algorithm for effective separation of protein structural classes. Comput. Biol. Chem. 32 (4), 302–306.

Spataru, S., Sera, D., Kerekes, T., et al., 2015. Diagnostic method for photovoltaic systems based on light I-V measurements. Sol. Energy 119 (sep.), 29–44.

Sun, X., Chavali, R.V.K., Alam, M.A., 2019. Real-time monitoring and diagnosis of photovoltaic system degradation only using maximum power point—the Suns-Vmp method. Progress Photovolt. Res. Appl. 27.

Triki-Lahiani, A., Abdelghani, Bennani-Ben, Slama-Belkhodja, I., 2018. Fault detection and monitoring systems for photovoltaic installations: A review. Renew. Sustain. Energy Rev. 82 (PT.3), 2680–2692.

Xla, B., Pla, B., Sca, B., et al., 2019. Fault diagnosis for photovoltaic array based on convolutional neural network and electrical time series graph. Energy Convers. Manage. 196, 950–965.

Yc, A., Mm, B., Ac, C., et al., 2019. Simple and efficient approach to detect and diagnose electrical faults and partial shading in photovoltaic systems. Energy Convers. Manage. 196, 330–343.

Zhang, J., Liu, Y., Li, Y., et al., 2020. A reinforcement learning based approach for online adaptive parameter extraction of photovoltaic array models. Energy Convers. Manage. 214, 112875.

Zhu, H., Wang, H., Kang, D., et al., 2019. Study of joint temporal-spatial distribution of array output for large-scale photovoltaic plant and its fault diagnosis application. Sol. Energy 181 (MAR.), 137–147.