Energy Conversion and Management 245 (2021) 114603

Contents lists available at ScienceDirect

Energy Conversion and Management

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

Fault diagnosis approach for photovoltaic array based on the stacked auto-encoder and clustering with I-V curves

Yongjie Liu [a], Kun Ding [a,*], Jingwei Zhang [a], Yuanliang Li [b], Zenan Yang [a], Wenming Zheng [c], Xiang 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] State Grid Jiangsu Intergated Energy Service Co., Ltd., Changzhou, Jiangsu 213022, China

ARTICLE INFO

Keywords:
Photovoltaic array
Fault diagnosis
Stacked auto-encoder
Clustering
Photovoltaic modeling
I-V and P-V curves

ABSTRACT

Photovoltaic arrays are usually installed outdoors in harsh environments and prone to various faults, which will seriously affect the efficiency of photovoltaic arrays. Therefore, the effective fault detection and diagnosis plays an important role in the safe, operation, and maintenance of the photovoltaic plant. In recent years, machine learning methods have made remarkable achievements in fault diagnosis. However, there still exist some limitations: (1) feature extraction relies on expert experience and lacks automation. (2) artificial feature extraction easily ignores some potential useful features. (3) the nonlinear characteristics of current–voltage curves cannot be effectively learned by the shallow network structure. In order to address the above issues, the supervised deep learning methods with automatic feature extraction capability are applied, but a lot of labeled data are required for pre-training. Therefore,a fault diagnosis method is proposed for photovoltaic array based on stacked auto-encoder and clustering algorithm in this paper, which can automatically extract features and use a small number of labeled data samples to mine data sample features for fault diagnosis. Firstly, the effective features are automatically extracted by the stacked auto-encoder from the current–voltage curves. Secondly, the dimension of the features is reduced and visualized by the t-distributed stochastic neighbor embedding to improve the performance of the clustering algorithm. Finally, clustering centers and clusters are obtained by clustering algorithm and membership function is used for fault diagnosis. Moreover, the simulation and experimental data are used to verify the performance of the proposed fault diagnosis method. The 97.3% and 98.3% classification accuracies are obtained in the simulation and experimental results.

1. Introduction

With shortage of fossil energy and environmental pollution, solar energy as renewable energy is widely concerned because of its cleanliness, sustainability, and great development potential.The photovoltaic (PV) power generation is one of the main ways of solar energy conversion and utilization. According to the latest statistics of the International Energy Agency (IEA), the global solar PV capacity increases are expected to nearly 107GW in 2020, which means steady growth from 2019. PV capacity in China will increase by more than 30% from 2019 [1]. The PV array is an important part of the PV system. However, it is installed outdoors and exposed in a harsh environment, influenced by the strong wind, heavy rain, high temperature, etc., for a long time, it is prone to occur abnormal and faults which will seriously affect the efficiency of the PV system [2,3]. Therefore, effective fault detection and diagnosis (FDD) plays an important role in the safe operation and maintenance of the PV power station.

To detect and diagnose PV array faults effectively, a variety of fault diagnosis methods have been proposed in recent years, which can be divided into three categories from the aspect of the data collected. The first category is based on the infrared thermogram, unmanned aerial vehicle (UAV). The infrared thermal imager are used to detect and locate defects such as hot-spots by analyzing the temperature characteristics of different PV modules by infrared thermogram [4,5]. However, the high cost of UAV and high-resolution infrared imager makes it unsuitable for small-scale PV systems.

The second category is based on detecting the operating voltage, current, or power data at the maximum power point (MPP) of the PV array. The deviations between the theoretical working current, voltage,

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

https://doi.org/10.1016/j.enconman.2021.114603
Received 9 April 2021; Accepted 1 August 2021
Available online 11 August 2021

Nomenclature

Abbreviations

AI	artificial intelligence
AE	auto encoder
BP	back propagation
CNN	convolution neural network
CFSFDP	clustering by fast search and find of density peaks
CFFSM	code-based fast fault simulation model
DL	deep learning
DF	degradation faults
EWMA	exponential weighted moving average
FDD	fault detection and diagnosis
IEA	international energy agency
ICA	independent component analysis
ICFFSM	improved code-based fast fault simulation model
KELM	kernel extreme learning machine
KPCA	kernel principal component analysis
LPP	locality preserving projections
LE	laplacian eigenmaps
LLE	locally linear embedding
LSTM	long short-term memory
ML	machine learning
MPP	maximum power point
MDS	multiple dimensional scaling
NMI	normalized mutual information
PS	partial shading
PV	photovoltaic
PNN	probabilistic neural network
PSO	particle swarm optimization
PSIM	power simulation software
PSBO	partial shading with the bypass diode open-circuit fault
PSSC	partial shading with the bypass diode short-circuit fault
RSDM	reverse-biased single-diode model
ResNet	residual network
SCF	short-circuit faults
SCD	short-circuit with the degradation fault
SAE	stacked auto-encoder
t-SNE	t-distributed stochastic neighbor embedding
UAV	unmanned aerial vehicle

Symbols

a	the correction factor for avalanche breakdown
G	measured irradiance
G_{stc}	irradiance under STC
I_{ph}	photocurrent
I_d	diode current
I_{sh}	shunt current
I_s	saturation current of the diode
V_{bd}	breakdown voltage

or power of the mathematical model and the measured results are calculated, and the corresponding threshold is set to judge the above deviations for the FDD [6–9]. In our previous study, the measured current is compared with the reference current from the simulation model, and fault detection is carried out by Grubbs criterion and local outlier factor [6]. In [7], the online reduced kernel generalized likelihood ratio test technique is used for fault detection with MPP operation data of the PV array. Harrou et al. compared the measured current and voltage with the reference values based on the power simulation software (PSIM), and then the comparison results are used as the input of the improved kNN algorithm [8]. The threshold value is calculated by the exponential weighted moving average (EWMA) for fault detection. In [9], the failure indicators are obtained by using the simulated current and voltage at the MPP to evaluate the measured data. Then, the indicators are compared with the threshold for fault classification, including short-circuit, open-circuit, and partial shading. However, these methods are difficult to identify various fault types and have low diagnostic accuracy due to the less fault information contained in the operation data. Moreover, the threshold is difficult to determine.

The third category is based on the measured current–voltage (I-V) curves of PV array, which is also the main approach of this paper. The I-V curve contains more information than the operation data at MPP of the PV array. In recent years, more and more PV inverters have integrated the function of scanning I-V curves to make this method more feasible. Moreover, with the development of artificial intelligence (AI), machine learning (ML) has been widely used. More and more researchers use the features extracted from the I-V curves as the input of ML models for FDD. Chen et al. used the optimized kernel extreme learning machine (KELM) to classify faults [10]. The inputs are the normalized features, e.g. short-circuit current, open-circuit voltage, the current and voltage at MPP, series resistance, etc., extracted from the measured I-V curves. Spataru et al. proposed a fault diagnosis method based on PV inverter and analyze the shape of the I-V curve under different faults [11]. Then, the features are extracted and corresponding fuzzy logic rules are established for FDD. In [12], the random forest intelligent algorithm is applied for fault diagnosis of PV arrays by using the I-V curves. In [13], a hybrid algorithm of artificial bee colony algorithm and semi-supervised extreme learning machine is designed for FDD, using five characteristic parameters extracted from the I-V curve. The back propagation (BP) [14] and probabilistic neural network (PNN) [15] are used as diagnostic models and features are manually selected from the I-V curve as inputs. Although the above traditional ML methods have achieved remarkable results in the FDD of PV arrays, these methods still have the following drawbacks: (1) feature extraction relies more on the expert experience of the ML model designer and lacks of automation. (2) The accuracy of FDD depends on the effective feature parameters. Thus, the features of manually selection limits the performance of fault diagnosis methods. (3) the nonlinear characteristics of I-V curves can not be effectively learned by the shallow network structure.

In recent years, with the development of deep learning (DL), the automatic extraction of complex features has been successfully applied to rolling bearing fault diagnosis, medicine applications, image classification, etc. [16–18]. However, the research works on DL for PV faults diagnosis are still insufficient. Chen et al. proposed a deep residual network (ResNet) model for FDD of PV arrays using the output I-V characteristic curves and corresponding environmental parameters [19]. Appiah et al. used the long short-term memory (LSTM) to extract the fault features and the fault features are input into the softmax regression classifier for FDD [20]. Lu et al. converted the original current and voltage signals into two-dimensional sequence diagrams as the input of a convolution neural network (CNN) for FDD [21]. However, the above automatic feature extraction methods are based on supervised methods, which require a large number of labeled data samples to train the networks, and are prone to problems such as overfitting and underfitting [22]. In practice, a large number of labeled fault data samples are difficult to obtain and the long-term fault operation of PV arrays will seriously affect the service life of PV modules. Moreover, these methods do not consider concurrent faults, such as the open-circuit of bypass diode with partial shading, short-circuit with partial shading, etc.

To address the above limitations, a fault diagnosis method of PV array based on stacked auto-encoder (SAE) and clustering is proposed in this paper to mine the inherent characteristics of data samples and realize automatic feature extraction and fault diagnosis. The proposed

method requires only a small number of labeled data to mine data features and establish the diagnostic model. The main contributions of this article can be summarized as follows:

(1) A data mining method based on SAE and clustering is designed and applied for FDD of PV array with original I-V and P-V curves. The proposed method includes feature extraction, feature dimensionality reduction and visualization, clustering, and membership degree calculation.

(2) The residuals between the original measured and the simulated reference I-V and P-V curves are used as the input of SAE to extract effective feature parameters. Next, the extracted feature parameters are reduced and visualized by t-SNE to improve the performance of clustering. Then, the clustering by fast search and find of density peaks (CFSFDP) is used to divide the reduced features into different clusters and obtain the corresponding cluster centers without specifying the number of cluster centers in advance. Finally, the membership degree between the measured data and the cluster centers of the labeled data is calculated to diagnose the single and concurrent faults of the PV array.

(3) The accurate modeling of the improved code-based fast fault simulation model (ICFFSM) is established to simulate various faults for verifying the proposed method. The unknown model parameters of the reverse-biased single-diode model (RSDM) in ICFFSM are extracted by particle swarm optimization (PSO) from the measured I-V curves. Moreover, the functional relationship between the model parameters and irradiance or temperature is fitted in different seasons to predict the unknown model parameters of the RSDM for improving the accuracy of the ICFFSM.

(4) The influence of different faults on the shape of I-V and P-V curves are analyzed based on the simulation model to demonstrate the feasibility of fault diagnosis with I-V and P-V curves. Simulation and experimental data are used to verify the feasibility and performance of the proposed method, and the results indicate that the proposed method has high accuracy and reliability.

The rest of this paper is organized as follows. In Section 2, the improved code-based fast fault simulation model and the model parameters extraction method are proposed. In Section 3, the influence of different faults on the shape of I-V and P-V curves are described. In Section IV, the fault diagnosis method for PV array based on SAE and clustering is detailed. In Section 5, the simulation and experiments are carried out to verify the feasibility and accuracy of the proposed fault diagnosis method. Finally, some significant results are concluded.

2. Improved code-based fast fault simulation model (ICFFSM) and parameters extraction

Previously, the circuit-based simulation software, such as MATLAB/Simulink, PSIM was used to establish the simulation models of PV array by some researchers[23–25]. The circuit-based simulation models have these limitations: poor portability; poor computational efficiency and complex circuit-model; high investment costs; To address the mentioned limitations, the code-based fast fault simulation model (CFFSM) was proposed in [26]. However, it can only simulate three fault types, including short circuit, partial shading, and increased series resistance. The open-circuit of the bypass diode and the corresponding concurrent faults has not been realized. Moreover, the step size is difficult to determine for the fixed-step scanning in CFFSM. Thus, the ICFFSM is proposed to solve the above problems and RSDM is used to model for higher accuracy under non-uniform irradiance in our previous work [27].

In order to obtain an accurate simulation model, the unknown model parameters of the RSDM in ICFFSM are extracted by PSO from the measured I-V curves and the fitting function between the model parameters and irradiance or temperature is established in different seasons to predict the unknown model parameters of the RSDM for improving the accuracy of the ICFFSM. Next, the accuracy of the simulation model is verified by experimental data and the error is analyzed. Finally, the ICFFSM is utilize to obtain the data of various faults under wider environmental conditions, to validate the accuracy and generalization of the FDD.

Fig. 1. The PV experiment platform.

2.1. Parameter extraction of the RSDM for the ICFFSM

The basic equation of the RSDM is as follows [25]:

$$I = I_{ph} - I_s\left[\exp\left(\frac{q(V+R_sI)}{nkT}\right)-1\right] - \frac{V+R_sI}{R_{sh}} - \alpha(V + R_sI)\left(1-\frac{V+R_sI}{V_{bd}}\right)^{-m} \quad (1)$$

where V and I are the output voltage and current, respectively; I_{ph} is the photocurrent;R_s and R_{sh} are the series and shunt resistances, respectively. I_s is the saturation current of the diode;n is the ideality factor. k is the Boltzmann constant ($1.38064852\times10^{-23}\mathrm{JK}^{-1}$), q is the electronic charge ($1.60217662\times10^{-19}C$). T is the temperature of solar cell.α is the correction factor for avalanche breakdown, V_{bd} is the breakdown voltage, and m is the avalanche breakdown exponent.

The classical method of solving the unknown model parameters is the analytical method based on the specification parameters of the PV module provided by the manufacturer [28]. However, the specification parameters may deviate seriously from the actual operation of the PV array due to the inconsistency of PV module products and degradation after the long-time operation, which leads to the reduction of the model accuracy. In this paper, the three parameters, i.e. $\alpha, V_{bd}, m$of RSDM for determining the reverse-biased characteristic are considered as $0.002\Omega^{-1}$, $-21.29V$, and 3 respectively [29–31].Then, there is an apparent mutual relationship between the ideality factor n and the saturation current of the diode I_s, i.e., the increase of ideality factor n will increase the open-circuit voltage of the I-V curve, the increase of the saturation current I_s will reduce the open-circuit voltage of I-V curve [10]. This will lead to multiple solutions for parameters extraction to solve the same I-V curve, which makes the result of parameter extraction

Table 1
Specification of PV module TSM-240 under STC.

Parameters	Value
Maximum power ($P_{mpp,stc}$)	240 W
Voltage at maximum power point ($V_{mpp,stc}$)	29.7 V
Current at maximum power point ($I_{mpp,stc}$)	8.1 A
Open circuit voltage ($V_{oc,stc}$)	37.3 V
Short circuit current ($I_{sc,stc}$)	8.62 A
Temperature coefficent of current (K_i)	0.047 %/°C
Temperature coefficent of voltage (K_v)	−0.32 %/°C

Table 2
Fitting function and correlation coefficient of model parameters.

Parameters	Fitting Function	Spring	Summer	Autumn	Winter
I_{ph}	$I_{ph} = aG_d + b$	$a =$ 0.009601 $b =$ −0.2732	$a =$ 0.009671 $b =$ −0.2805	$a =$ 0.009603 $b =$ −0.2556	$a =$ 0.009362 $b =$ −0.2065
I_s	$I_s = ae^{bT}$	$a =$ 5.527×10^{-12} $b = 0.1598$	$a =$ 1.371×10^{-11} $b = 0.1461$	$a =$ 2.124×10^{-11} $b = 0.138$	$a =$ 9.318×10^{-12} $b = 0.1471$
R_s	$R_s = aG_d^b + c$	$a =$ 1.928×10^5 $b =$ −2.306 $c = 7.632$	$a =$ 2.543×10^6 $b =$ −2.582 $c = 8.056$	$a =$ 3.755×10^4 $b =$ −1.744 $c = 8.256$	$a =$ 4.135×10^4 $b =$ −2.154 $c = 8.239$
R_{sh}	$R_{sh} = aG_d^b + c$	$a =$ -9.64×10^4 $b =$ 0.03463 $c =$ 1.245×10^5	$a =$ 8.299×10^4 $b =$ −0.1354 $c =$ -3.08×10^4	$a =$ 1.899×10^5 $b =$ −0.5289 $c = -3193$	$a = 1.61 \times 10^5$ $b =$ −0.4575 $c = -4783$

fluctuate greatly. In order to ensure the stability of parameter extraction results, one of these two parameters should be determined in advance. Therefore, in this paper, the ideality factor n is fixed as a typical value 1 before parameter extraction [32]. The remaining model parameters $[I_{ph}, I_s, R_s, R_{sh}]$ are extracted by the particle swarm optimizer (PSO) from the measured I-V curves [33].

The measured I-V curves are obtained from the 5.28 kWp PV array, which is installed on the south roof of the College of Mechanical and Electrical Engineering building in Hohai University, Changzhou Campus, as shown in Fig. 1. The PV array consists of 22 multi-crystalline silicon modules connected (TSM-240) in series. The specification of the TSM-240 is given in Table 1. The PV module consists of 3 bypass diodes, and each bypass diode is connected with 20 solar cells in parallel. The I-V curve of the PV array is measured by the inverter GW20KN-DT and stored with the corresponding irradiance and temperature data per 2 min.

The measured data from June 2018 to June 2019 were divided into four seasons for model parameters extraction. The unknown parameters of the model are fitted to with irradiance or temperature in different seasons. The corresponding fitting equations and coefficients in different

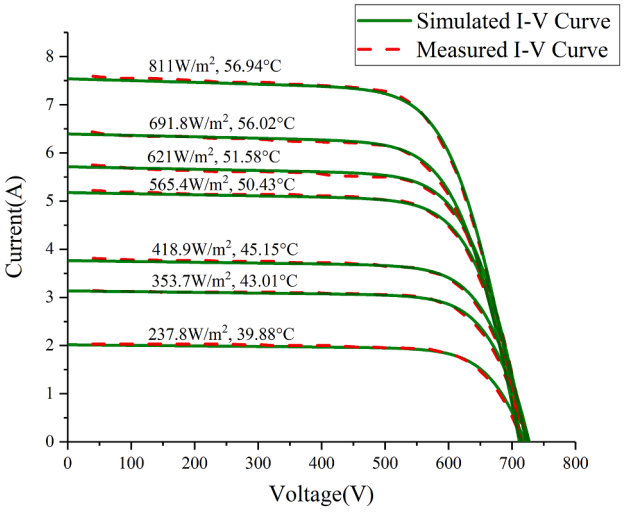


Fig. 3. Comparison of simulated and measured I-V curves under different ambient conditions.

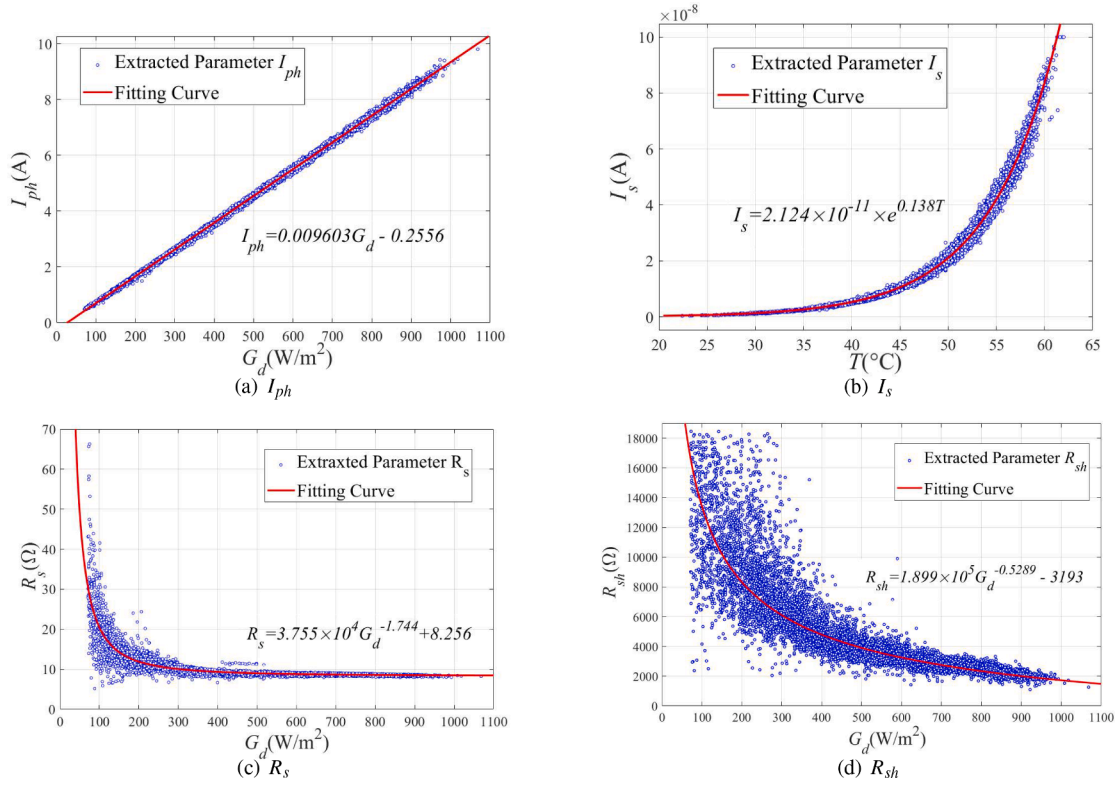


Fig. 2. Distribution of extracted parameters (a)I_{ph}, (b)I_s, (c)R_s, (d)R_{sh} and fitting curves in autumn.

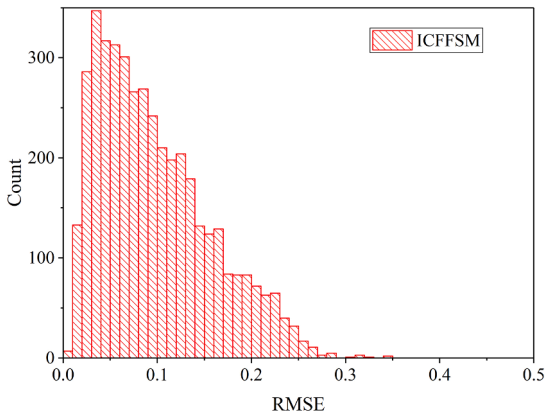


Fig. 4. The distribution of RMSE for ICFFSM under July to September 2019.

seasons are shown in Table 2. The distribution of model parameters relative to the irradiance or temperature, and fitting curves in autumn (Sep. to Nov.) are shown in Fig. 2.

2.2. Experimental verification of the ICFFSM

In order to validate the accuracy of the proposed model, the seven I-V curves under different irradiance and temperature conditions are randomly selected from the measured datasets. The experimental and simulation results are compared in Fig. 3., it is obvious that the simulated curves are highly consistent with the measured curves. The measured I-V curves from July to September 2019 are used to verify the accuracy of the ICFFSM, and the root mean square errors (RMSE) between the simulated I-V curve and the measured I-V curve are calculated to measure the accuracy. The Fig. 4 shows the histogram of the RMSEs for the estimated I-V curves. The proposed modeling method has high accuracy with the mean RMSE 0.0893A. Besides, the accuracy of the proposed method is stable and corresponding RMSE converges at approximate 0.05A. However, due to the degradation phenomenon in the long-term operation of the PV array, it is necessary to periodically update the fitting equation with the measured data, otherwise the performance of the model output will deviate from the actual output state of the PV array.

3. Simulation and analysis of the influence of different faults on I-V curve

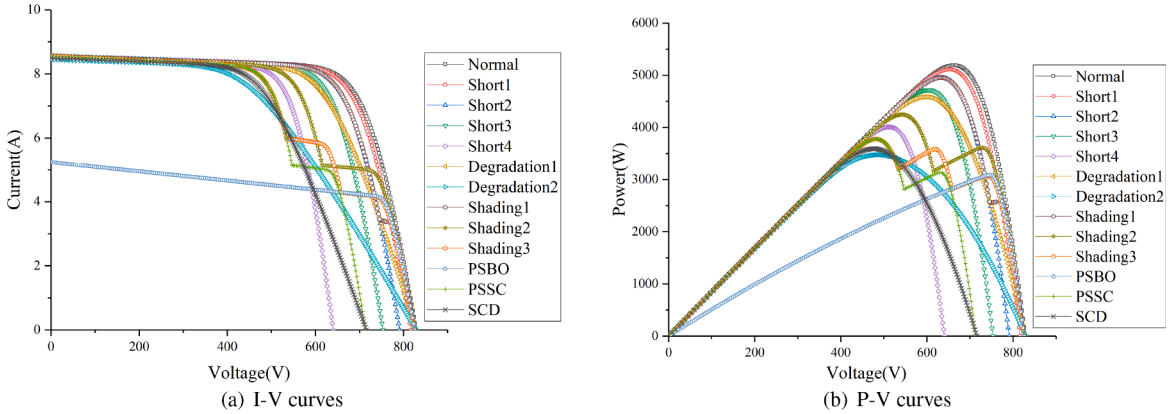
In addition to the three typical faults of short-circuit faults (SCF), degradation faults (DF), partial shading (PS), the concurrent faults such as the partial shading with the bypass diode open-circuit fault (PSBO), partial shading with the bypass diode short-circuit fault (PSSC), short-circuit with the degradation fault (SCD) are also investigated.

The effect of the single or concurrent faults on the I-V and P-V curves are analyzed by the ICFFSM mentioned in Section 2, as depicted in Fig. 4. SCF can be divided into two types. One is the short-circuit of PV modules and the other is the short-circuit of the bypass diodes, which all caused by the accidental connection between two points of the PV array. In this work, the SCF of one bypass diode and different numbers of PV modules (one, two, and five PV modules) are defined as Short1, Short2, Short3, Short4 respectively. As shown in Fig. 5, the short-circuit fault has a serious influence on the I-V and P-V curves. As the degree of short-circuit increases, the open-circuit voltage and maximum power decrease significantly.

DF can be caused by many factors such as the degradation of series-connected cable, mechanical damage and harsh operating conditions, e.g. strong ultraviolet, high temperature, and oxidation lead to discoloration, delamination, and corrosion in the PV modules. These degradation phenomena will result in increased series resistance (R_s) and reduced parallel resistance (R_{sh}). This paper mainly studies the DF of the series resistance (R_s) increase since it commonly occurs and has a greater effect on the efficiency of the PV modules. Two different degrees of DF are investigated, i.e., the series resistance increase of 10Ω, 20Ω are defined as Degradation1 and Degradation2, respectively. As the increase of R_s, the slope of the I-V curve between the MPP and the open-circuit point is changed, and the maximum power is reduced, as illustrated in Fig. 5.

PS refers to the nonuniform of the input irradiance on the plane of PV array caused by the bird droppings, dust, tree, or building shadows. The PS may lead the temperature of the shaded part in the PV array rising and form hot-spots, which will seriously affect the life of PV modules. PS can be divided into two types based on the state of the bypass diode, one is that the bypass diode turns on, and protects the PV module from the damage caused by the hot-spots. Under this condition, the PS causes the distortion of the I-V curves and multiple local maximum power peaks in the P-V curves. Moreover, the different degrees for PS are investigated, e.g., one PV module with 60% irradiance reduction (Defined as Shading1), three PV modules with 40% irradiance reduction (Defined as Shading2), the irradiance reduction of the two shaded PV modules is 50% and the other three is 30% (Defined as Shading3). Fig. 5 reveals that the higher irradiance reduction caused by the shading, the lower current at the inflection point of the I-V curve. The I-V and P-V curves will have several inflection points and local maximum power points respectively when there are multiple shadow blocks with different irradiance reduction.

The other is PSBO (three PV modules with 50% irradiance reduction and corresponding bypass diodes are open-circuit) that is a concurrent fault. In the case of the PSBO, a significant reduction of the short-circuit current and the maximum power of the I-V curve, as is shown in Fig. 4.



(a) I-V curves (b) P-V curves

Fig. 5. Simulation results of the PV array under different fault conditions (a) I-V curves; (b) P-V curves.

Besides, the P-V curve without multiple local maximum power points.

The concurrent fault of the PSSC is also investigated, i.e., three PV modules are shaded with three PV modules are short-circuit. The corresponding I-V curve is shown in Fig. 5(a), it is obvious that the open-circuit voltage is reduced and the I-V curve is distorted. The power at the MPP is reduced and two local peaks occur of the P-V curve, as is shown in Fig. 5(b).

The SCD fault is configured with a series resistance ($R_s = 15$) and short-circuit fault of three PV modules. The slope of the I-V curve between the MPP and the open circuit point is changed and the open-circuit voltage is reduced. The maximum power is reduced and the position of the MPP deviates to the left of the P-V curve, as is shown Fig.5. Therefore, the different types and degrees of the faults have different effects on the I-V and P-V curves, which further reveals the feasibility of feature extraction based on the I-V and P-V curves.

4. The fault diagnosis approach based on SAE and clusterring for PV array

The proposed fault detection and diagnosis method mainly include the following parts: (1) data preprocessing, (2) feature extraction by the stacked autoencoder (SAE), (3) feature dimensionality reduction and visualization by the t-SNE, (4) feature clustering representation and selection of the cluster center points by the CFSFDP, (5) fault diagnosis.

4.1. Data processing

The original I-V and P-V curves are used for FDD to avoid the specialist experience and artificial feature extraction. Then, some simple preprocessing of the original data is needed to improve the accuracy of the FDD. The original I-V and P-V curves of different irradiance and temperature is transformed to STC to eliminate the influence of the meteorological factors. The equations for transforming the raw I-V curve are shown in (2)-(6) [34].First, the short-circuit current and open-circuit voltage under different environments are transformed to STC:

$$I_{sc\_ref} = I_{sc}\frac{G_{stc}}{G}\Big/[1 + K_i(T - T_{stc})] \tag{2}$$

$$V_{oc\_ref} = V_{oc}\Big/[1 + K_g \ln\frac{G}{G_{stc}} + K_v(T - T_{stc})] \tag{3}$$

where G is the measured irradiance. T is the measured temperature. G_{stc} and T_{stc} are irradiance and temperature under STC, i.e.1000 W/m^2 and 25 °C, respectively. K_g is the irradiance correction factor of voltage and is usually 0.06 [34]. V_{oc_ref}, I_{sc_ref} represent the transformed open-circuit voltage and short-circuit current under STC. I_{sc} and V_{oc} are the short-circuit current and open-circuit voltage under operation conditions. Then, the remaining data points on the I-V curve are transformed to the STC by using (4)–(6).

$$I_{ref} = I_{out}\frac{I_{sc\_ref}}{I_{sc}} \tag{4}$$

$$V_{ref} = V_{out} - (V_{oc} - V_{oc\_ref}) + R_{s\_stc}(I_{ref} - I_{out}) \tag{5}$$

$$R_{s\_stc} = \frac{V_{mpp\_stc}(I_{sc\_stc} - I_{mmp\_stc})\ln(1 - \frac{I_{mmp\_stc}}{I_{sc\_stc}}) + I_{mmp\_stc}(V_{oc\_stc} - V_{mmp\_stc})}{I_{sc\_stc}(I_{sc\_stc} - I_{mmp\_stc})\ln(1 - \frac{I_{mmp\_stc}}{I_{sc\_stc}}) + I^2_{mmp\_stc}} \tag{6}$$

I_{out} and V_{out} are the measured current and voltage, respectively. I_{ref}, V_{ref} are the transformed current and voltage of the PV array under STC. R_{s_stc} is the series resistance of PV array under STC, which can be analytically solved via (6).

In order to ensure the validity of the extracted features, it is necessary to normalize the current and voltage of the I-V curve to the range [0 −1].

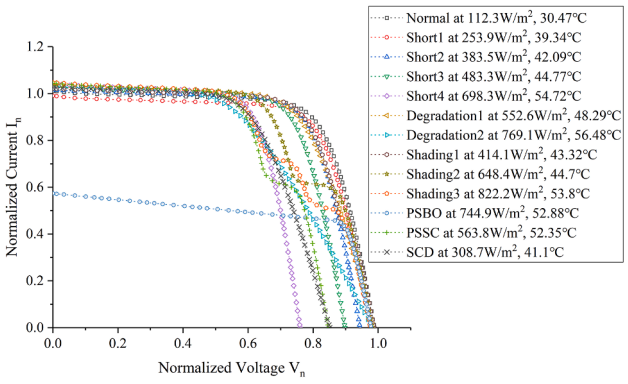


Fig. 6. Normalized I-V curves of different faults under different ambient conditions.

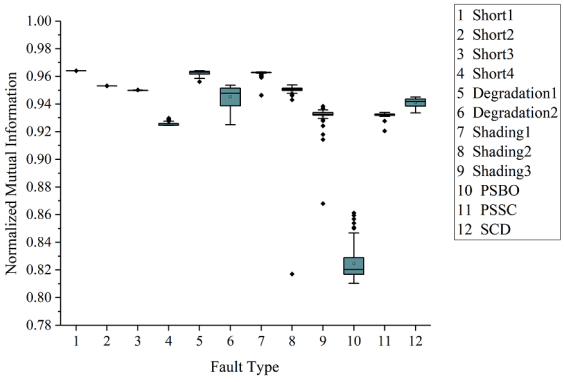


Fig. 7. Normalized mutual information of the fault types.

The dimensionless data makes each variable have the same weight in feature extraction. Moreover, the normalized processing improves the convergence rate of the SAE. The standardized equations are written as follows [35]:

$$I_n = \frac{I_{ref}}{I_{sc\_stc}} \tag{7}$$

$$V_n = \frac{V_{ref}}{V_{oc\_stc}} \tag{8}$$

where I_n, V_n are the normalized values of current and voltage.

The input requirement of SAE for automatic feature extraction is that has the same dimensions. However, the data points on the original I-V curves are non-uniformly distributed. Besides, the number of sampled points on the I-V curves are different under different ambient conditions. Thus, the normalized I-V curves are linearly interpolated to obtain 60 new data points from short-circuit current to open-circuit voltage, to make the points uniformly distributed on the I-V curves. Fig. 6 shows the I-V curves under different ambient conditions are normalized after data preprocessing. The normalized mutual information (NMI) is calculated to measure the similarity between different fault types and the normal operation for 100 I-V curves simulated by ICFFSM [36], and the results are shown in Fig. 7. The NMI of the PSBO is obviously less than other fault types. However, NMI values of the fault types discussed in this paper are more than 80%. Therefore, it is difficult to extract the effective features to distinguish normal and different faults. In order to further highlight the features of faults, the residuals between the measured and reference data are calculated and used as the input of SAE.

4.2. Stacked autoencoder (SAE)

The autoencoder (AE) refers to a three-layer fully connected neural network, which comprises input layer, hidden layer, and output layer for unsupervised learning and automated feature extraction. The input layer and the hidden layer form the encoder network $\boldsymbol{h} = \boldsymbol{f}(\boldsymbol{x})$, and the hidden layer and the output layer constitute the decoder network $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$. The input data of the high dimensional space is nonlinearly mapped to the low dimensional feature space, which is achieved by the encoder:

$$\boldsymbol{h} = f(\boldsymbol{W_E x} + \boldsymbol{b_E}) \tag{9}$$

where $\boldsymbol{x} = [\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}, \ldots, \boldsymbol{x_m}] \in \mathbb{R}^m$ denotes the input data; the $\boldsymbol{h} = [\boldsymbol{h_1}, \boldsymbol{h_2}, \boldsymbol{h_3}, \ldots, \boldsymbol{h_n}] \in \mathbb{R}^n$ represents the hidden layer features and $n < m$;$\boldsymbol{f}$ expresses activation function; $\boldsymbol{W_E} \in \mathbb{R}^{m \times n}$ is the connection weight of the encoder network; $\boldsymbol{b_E}$ is the bias term of the encoder network. Subsequently, the original input data $\boldsymbol{x}$ is reconstructed with the low dimensional feature $\boldsymbol{h}$ by the decoder:

$$\boldsymbol{y} = f(\boldsymbol{W_D h} + \boldsymbol{b_D}) \tag{10}$$

where $\boldsymbol{y} = [\boldsymbol{y_1}, \boldsymbol{y_2}, \boldsymbol{y_3}, \ldots, \boldsymbol{y_m}] \in \mathbb{R}^m$ represents the reconstructed data. The $\boldsymbol{W_D} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{b_D}$ respectively denote connection weight and bias terms of the decoder network. Since the sigmoid function is stable and has been extensively used, the sigmoid function acts as the activation function here. The equation is written as:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{11}$$

The AE aims to minimize the reconstruction error between the input data and the reconstructed data. Next, the representation features in the hidden layer are obtained by minimizing the reconstruction error. The mean square error is applied as the loss function of AE, with the equation as:

$$L = \frac{1}{2N} \sum_{i=1}^{N} ||x^{(i)} - y^{(i)}||_2^2 \tag{12}$$

where N represents the number of the training samples; $\| \|_2$ denotes second norm. The SAE refers to a deep neural network stacked by multiple AE models to learn the deep features of the raw input data. Moreover, the respective model of the SAE has a simple structure and can be trained individually, so the network becomes easier to train and exhibits convergence and accuracy. The deep feature extraction process of SAE is that the raw data is inputted to train the first AE. Subsequently, the features extracted by the first AE are used to train the second AE. This process is repeated till the end of the last AE training. The last AE of feature representation refers to the deep features of the raw data.

4.3. T-distributed stochastic neighbor embedding (t-SNE)

The t-SNE refers to a nonlinear manifold learning method, which keeps the high-dimensional data domain distribution consistent with the low-dimensional data domain distribution [37]. It is capable of effectively realizing dimensionality reduction and embedding the high-dimensional nonlinear data into 2D or 3D to facilitate the data classification. The probability functions are adopted to present the similarity between high-dimensional data. The distribution of data points in high dimensional space is assumed to follow the Gaussian distribution. T-distribution is applied in low-dimensional space, which effectively solves the problem of SNE congestion in low-dimensional data. First, given a high-dimensional data $\boldsymbol{x} = \{\boldsymbol{x_i}\}_{i=I}^N \in \mathbb{R}^D, x_i$ represents the ith high dimensional data sample, N denotes the number of data samples,and D represents the dimension of high-dimensional spatial data. The probability density $P_{j|i}$ between samples (x_i and x_j) is calculated as:

$$P_{j|i} = \frac{\exp(-||x_i - x_j||^2 \big/ 2\sigma_i^2)}{\sum_{k \neq i}^{N} \exp(-||x_i - x_j||^2 \big/ 2\sigma_i^2)} \tag{13}$$

where σ_i^2 represents the Gaussian variance centered on the x_i of the data sample. Besides,the joint distribution of two samples in the high-dimensional space can be expressed as follows:

$$P_{ij} = \frac{P_{j|i} + P_{i|j}}{2N} \tag{14}$$

Next, the joint probability density q_{ij} between samples (y_i and y_j) in low dimensional space ($\boldsymbol{Y} = \{\boldsymbol{y_i}\}_{i=I}^M \in \mathbb{R}^E (M < N)$,$\boldsymbol{y_i}$ represents the ith low dimensional data sample, M is the number of data samples. E represents the dimension of low dimensional data.) is calculated by the t-distribution:

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k \neq l}^{N} (1 + ||y_k - y_l||^2)^{-1}} \tag{15}$$

To align the distribution of high-dimensional data distribution with that of low-dimensional data, the KL divergence between the joint probability distributions P and Q is minimized. The cost function C of the KL divergence is defined as:

$$C = KL(P||Q) = \sum_{i \neq j} \log \frac{P_{ij}}{q_{ij}} \tag{16}$$

The cost function C is optimized with the gradient descent method. Then, the high-dimensional data points are mapped to low-dimensional data points and then fall into different types. The t-SNE is applied to dimensionality reduction and visualization of fault features in the proposed fault diagnosis method.

4.4. Clustering by fast search and find of density peaks (CFSFDP)

The clustering algorithm aims to excavate the inherent characteristics of data samples without any prior knowledge. Data samples are divided into different clusters based on similarity, so the similarity of elements in the same cluster is maximally and the similarity of elements of different clusters is maximally low. After clustering, the identical class of data is aggregated and different classes of data are separated. In [38], the k-means method is used to cluster data samples under different operation conditions for fault diagnosis. However, the algorithm requires a random initial of cluster centers and easily falls into local optimal in the iterative solution. Besides, the Fuzzy C-mean (FCM) is applied for fault diagnosis of PV array [39,40], which clusters the fault samples by calculating the Euclidean distance between samples. However, it is not able to detect the nonspherical clusters and cluster centers will need to determined in advance. Clusters with an arbitrary shape are easily detected by approaches based on the local density of data point. The density-based spatial clustering of applications with noise (DBSCAN) was applied to cluster by selecting core points by setting a density threshold [41]. However, the artificially preset density thresholds will directly affect the performance of clustering.

Fortunately, the CFSFDP is presented in 2014 [42]. The algorithm is not required to determine the number of cluster centers and other parameters in advance and clusters of arbitrary shapes can be divided based on density clustering. Moreover, the CFSFDP is simple and fast in convergence. In this study, the CFSFDP is applied here.

The CFSFDP has the core idea of the selection of cluster center points. The cluster centers are selected based on two distinct characteristics, one is the cluster center has a large local density ρ_i, i.e., the density of cluster center points is higher than that of neighbors. The other is the minimum distance δ_i from other density points.The main steps of the CFSFDP are as follows:

Step1: The local density ρ_i of each sample is defined by the Gaussian

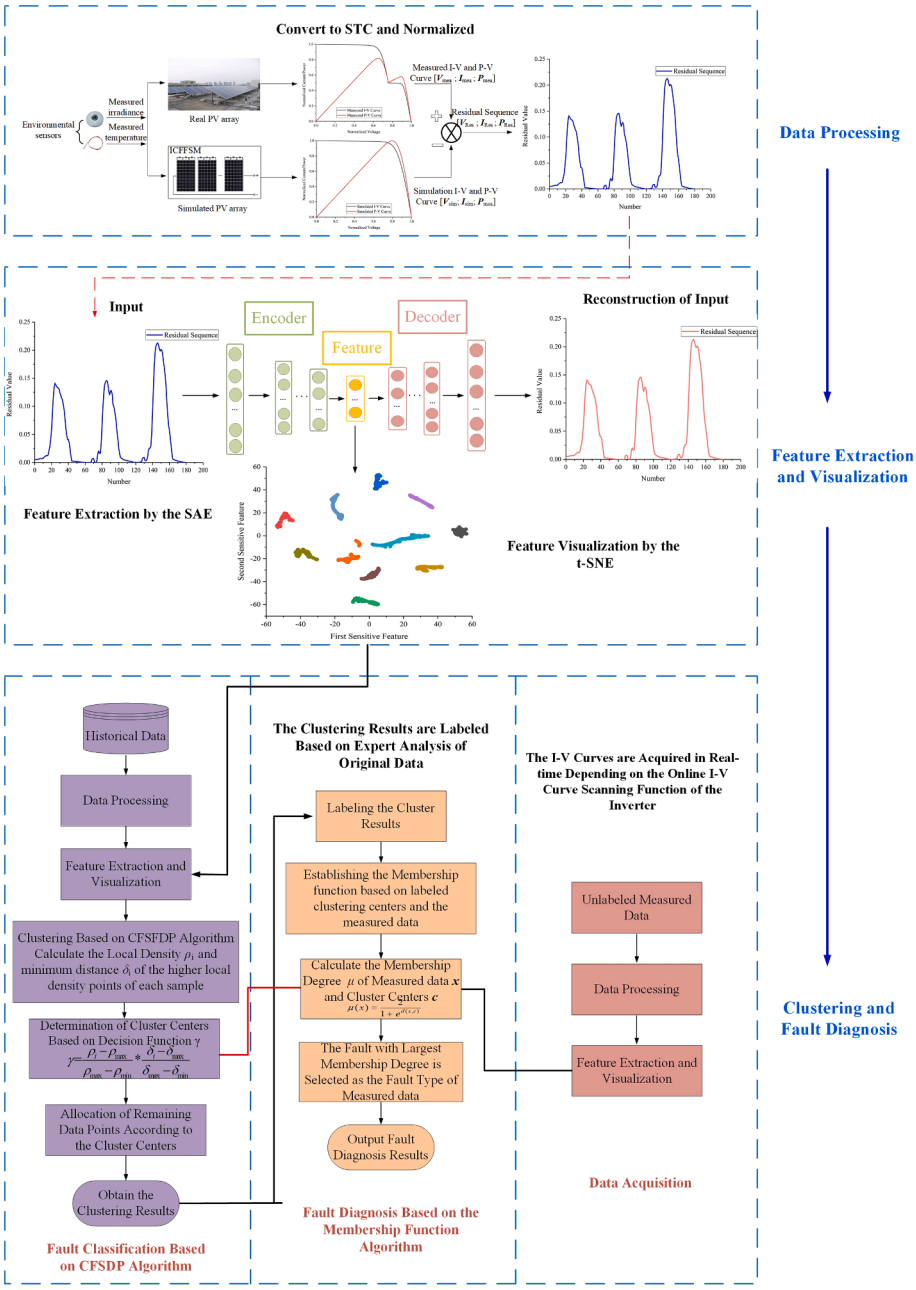


Fig. 8. The framework of the proposed unsupervised approach of PV array fault diagnosis.

kernel function, as expressed:

$$\rho_i = \sum_j \exp\left[-\left(\frac{d_{ij}}{d_c}\right)^2\right] \tag{17}$$

where $d_c(d_c > 0)$ denotes the cutoff distance, which is a vital parameter applied for calculating the local density ρ_i. The local density ρ_i is equated with the number of points in a cutoff distance d_c. The cutoff distance d_c is determined, so the average number of neighbors per point is 2% of the total number of data sets. d_{ij} is the Euclidean distance between i and j of data samples.

Step2:The minimum distance δ_i of the higher local density points is calculated. However, for the point with the highest local density ρ_i, the distance between the data points is taken as the maximum distance of the other data points, as follows:

$$\delta_i = \begin{cases} \min\limits_{j:\rho_j > \rho_i}(d_{ij}), & \exists \rho_j > \rho_i \\ \max\limits_{j}(d_{ij}), & \forall \rho_j \lesssim \rho_i \end{cases} \tag{18}$$

Step3: After the above analysis, the cluster center has high local density ρ and a large distance δ between the point and other data points with high density. Accordingly, the decision function γ is defined to consider local density ρ and distance δ. Moreover, to eliminate the numerical difference between ρ and δ, the ρ and δ are normalized respectively:

$$\gamma = \frac{\rho_i - \rho_{max}}{\rho_{max} - \rho_{min}} * \frac{\delta_i - \delta_{max}}{\delta_{max} - \delta_{min}} \tag{19}$$

Step4: the data points with the large value of the γ are taken as the cluster centers. The corresponding threshold is set to automatically select the larger γ and determine the cluster center points. The threshold definition is expressed as follows:

$$\gamma \geqslant [\mu(\rho^*) + \varepsilon\sigma(\rho^*)]\sigma(\delta^*) \tag{20}$$

where the μ represents the expected value; σ denotes the standard deviation; ρ^* and δ^* are the normalized values of the ρ and δ, respectively. On the whole, ε is usually set to 3 [43].

Step5: After the cluster center is determined, the other remaining data points are divided into clusters with higher local density and nearest cluster center.

4.5. Fault diagnosis method

As revealed from the mentioned, different fault types exert different effects on the I-V curves of the PV array. However, the mutual information is high between each fault type and reference I-V curves. Thus, the residuals of the measured and reference I-V curves is taken as the input of the SAE to extract effective features. The framework for the proposed fault diagnosis of PV array based on the deep learning consists of data preprocessing, feature extraction by the SAE, feature dimensionality reduction and visualization by the t-SNE, feature clustering representation, and selection of the cluster centers by the CFSFDP. Fig. 8 illustrates the proposed fault detection and diagnosis method. The proposed method only requires a small number of labeled samples to mine the inherent characteristic of different faults to build a fault diagnosis model, which does not need considerable prior knowledge and expensive labeled information. Next, the steps of the proposed fault diagnosis method are elucidated below:

Step1: Data acquisition. The small amount of labeled I-V curves of PV array are obtained with the simulation model or the real experimental platform. The irradiance and temperature of real PV arrays are uncontrollable, fault datasets without various ambient conditions. Therefore, more data samples are obtained to verify the accuracy of the fault diagnosis model by the simulation model in Section II.

Step2: Data preprocessing. The I-V curves under different ambient conditions are transformed to the STC to eliminate the influence of the environment. Next, the data is normalized to eliminate the numerical differences of different physical variables (i.e., current, voltage and power). Moreover, the residuals of the current, voltage, and power sequence of the measured and simulated normal data is obtained to highlight the effective features.

Step3: Feature extraction by the SAE. The network structure of the SAE is constructed by determining the number of autoencoder models and regulating the corresponding parameters. The preprocessed residual sequences are as the input of SAE to extract the effective features.

Step4:Feature dimensionality reduction and visualization by the t-SNE. High-dimensional features cover considerable redundant and incoherent information. The performance of the clustering algorithm is reduced as impacted by the direct processing of high-dimensional data. Besides, the similarity of most clustering algorithms is based on the Euclidean distance. However, the Euclidean distance only represents the linear distance, which cannot be directly applied in high-dimensional and nonlinear data. Therefore, the high-dimensional nonlinear features of the PV array are nonlinearly reduced to 2-dimensional vectors by the t-SNE for visualization and further verifies the effectiveness of fault feature extraction.

Step5: Clusters and clustering centers of visualization results of different faults are determined by the CFSFDP.

Step6: Fault diagnosis. The unlabeled real-time data is preprocessed, feature extracted and feature dimensionality reduced to obtain the visualization results. Subsequently, the membership degree of the real-time measured data and labeled clustering centers are calculated for fault diagnosis. The largest membership degree is selected as the fault type of real-time measured data. The degree of membership function is as follows:

$$\mu(x) = \frac{2}{1 + e^{d(x,c)}} \tag{21}$$

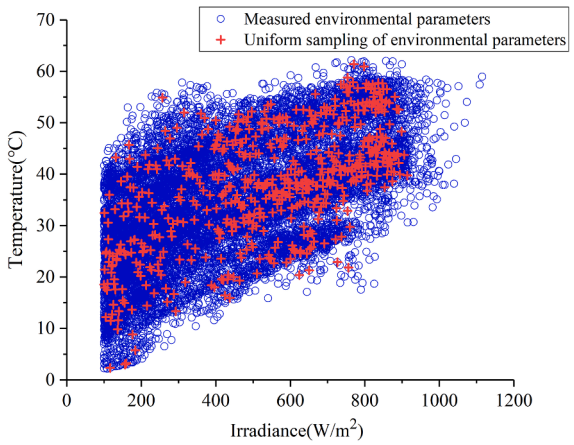


Fig. 9. Uniform sampling of environmental parameters.

$$d(x,c) = \sqrt{\sum_{j=1}^{n} (x_j - c_j)^2} \tag{22}$$

where $\mu(\boldsymbol{x})$ is the membership degree of the real-time measured data $\boldsymbol{x}$ and labeled clustering centers $\boldsymbol{c}$. The membership degree $\mu(\boldsymbol{x})$ is between 0 and 1. $d(\boldsymbol{x},\boldsymbol{c})$ is the distance between real-time measured data $\boldsymbol{x}$ and clustering center $\boldsymbol{c}$.

5. Verification and analysis of the proposed fault diagnosis approach

In order to verify the proposed fault diagnosis method, the three typical faults of short-circuit faults, degradation faults, partial shading, and the concurrent faults such as PSBO, PSSC, SCD are also investigated in this paper, as described in Section III. Firstly, the experimental data are obtained by the outdoor experimental platform or the ICFFSM simulation model. The ICFFSM is applied to obtain the data samples of wider range ambient conditions to verify the applicability of the diagnostic model. Secondly, the residuals of the measured and reference data are obtained by the data preprocessing. Thirdly, the features are extracted and dimension reduced by the method of combining the SAE and t-SNE. Then, the clusters and clustering centers of dimensional reduction results of different faults are determined. Finally, the membership function is established for fault diagnosis, 30% of the data are used to mine the features representation, and the remaining 70% of the data are used to verify the performance and accuracy of the diagnostic model.

5.1. Simulation verification and analysis

1) Data acquisition

As discussed in Section 3, normal operation and three typical faults including SCF (short1, short2, short3, short4), DF (Degradation1, Degradation2), PS (shading1, shading2, shading3), and three concurrent faults (PSBO, PSSC, SCD) are simulated and investigated in this paper. The 300 points of environmental parameters are uniformly sampled from the four different seasons in April, July, October 2019, and January 2020 as the input of the simulation model, as is shown in Fig. 9. Therefore, 300 I-V curves for each fault as well as the normal operation, there are 3900 I-V curves in the simulation dataset. 30% data samples of each fault are randomly selected to mine the fault data sample characteristics and obtain the cluster centers to establish the fault diagnosis model. Then, the remaining 70% data samples are used to evaluate the diagnostic accuracy of the proposed model.

2) Feature extraction and visualization

The more the hidden layers of the neural network, the deeper the

Table 3
Effect Of network layers on feature extraction performance of SAE.

Network layers	Iterations	Reconstruction error	Time (s)
3	300	0.03784	6.6023
4	300	0.02633	9.4946
5	300	***0.00889***	***12.34***
6	300	0.00765	15.3489
7	300	0.01826	17.6548

Table 4
Effect of iterations on feature extraction performance of SAE.

Iterations	Reconstruction error	Time (s)
100	0.02384	4.49
200	0.01460	8.29
300	***0.00889***	***12.035***
400	0.007982	16.5687
500	0.007565	20.3047
600	0.008205	25.42304

network structure will be, and the essential features of raw data can be mined. However, with the increase in the number of network layers, the computing time will be extended rapidly, and the network will fall into overfitting easily, which will cause the performance to decline. Therefore, the performance of the network and computing time are considered to select the appropriate hidden layers. The learning rate is 0.001, the number of neurons in each hidden layer, and the iterations are fixed to 150 and 300, respectively. The reconstruction error and calculation time of the SAE network are obtained, as shown in Table 3. As the number of layers is up-regulated, the error of data reconstruction is reduced significantly under the number of network layers less than five. Under the number of network layers over five, the error of data reconstruction is stable, whereas the calculation time is extened significantly. Thus, the number of hidden layers and network layers of SAE are 3 and 5, respectively.

After the number of network layers is determined, the iteration times of the SAE are tested and then analyzed here. The simulation data is trained under different iteration times, each test is repeated 10 times, and the mean of its run time and average reconstruction error is calculated as shown in Table 4. Under the iterations are less than 300, the reconstruction error of data samples is reduced with the increase in the number of iterations. Under the iterations over 300, the reconstruction error of the data is stable, whereas the calculation time is extended significantly. Thus, the iteration of SAE is set to 300. The decreasing principle is exploited to remove considerable redundant information from the original data to more effectively extract the feature when selecting the number of neurons in each hidden layer. Lastly, the number of neurons in the hidden and output layers of SAE is determined by the

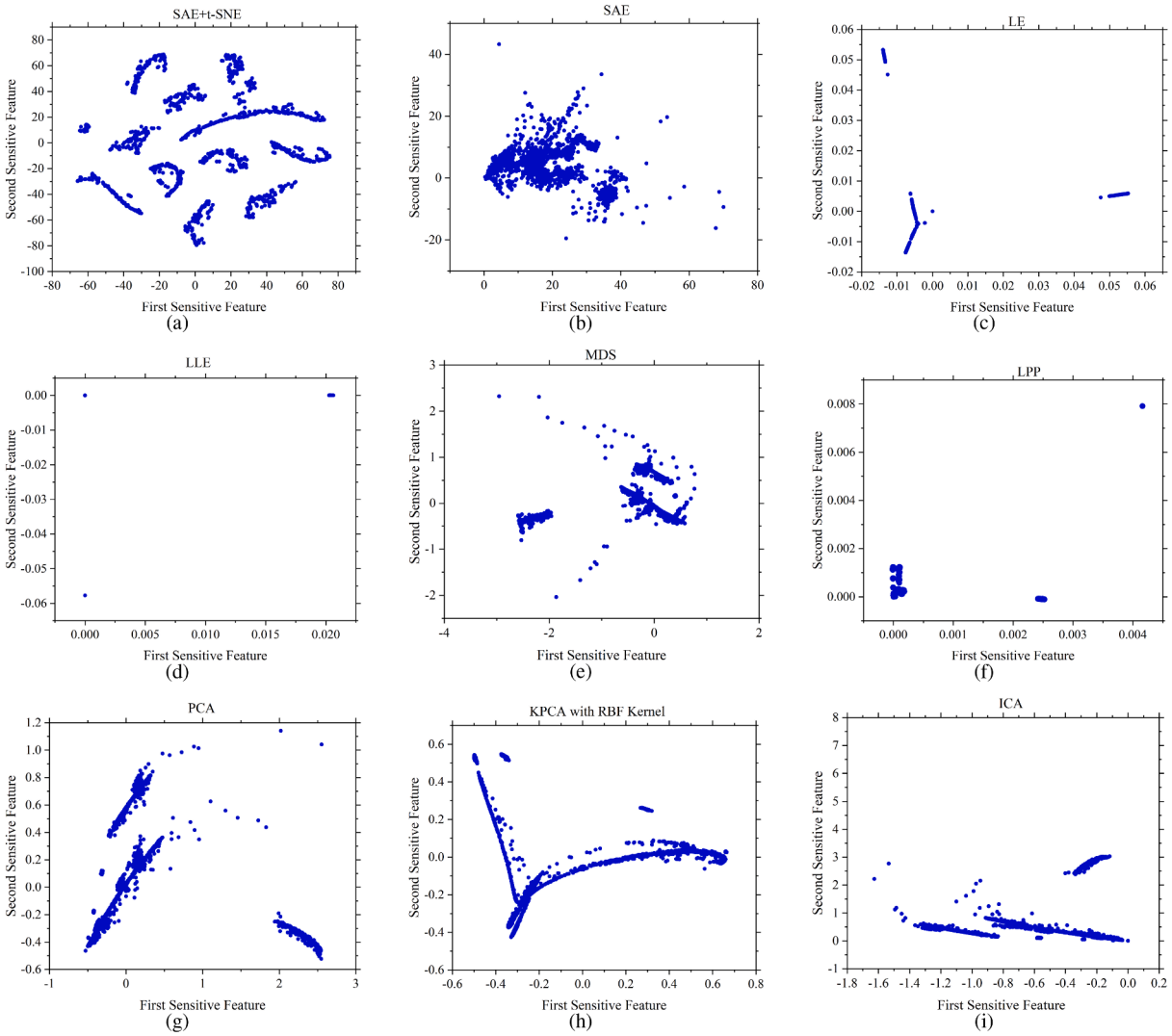


Fig. 10. Visualization results with different dimensionality reduction algorithms.

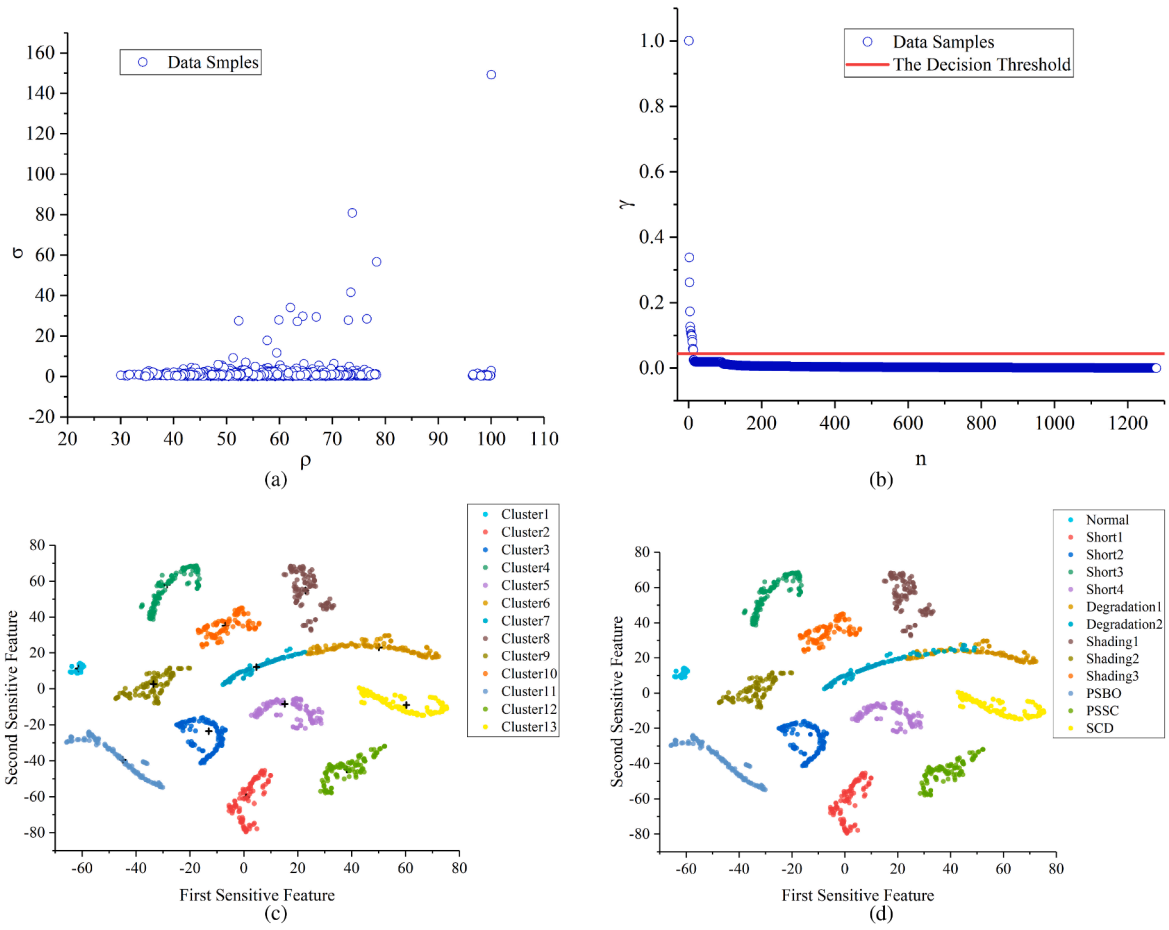


Fig. 11. (a) The decision graph for 2-D features of simulation dataset; (b) The value of γ_i in decreasing order for simulation dataset; (c) The clustering results; (d) The real distribution of 2-dimension features.

grid search method to minimize the reconstruction error. The network structure of encoder is input (180)-hidden1 (120)-hidden2 (70)-hidden3 (40)-output (15) here. The output layer (15) of the encoder is recognized as a valid feature representation of the residual sequences.

However, the high-dimensional data contains covers considerable redundant and irrelevant information, clustering high-dimensional data directly will reduce the performance of the algorithm. Thus, the extracted 15-dimensional features are reduced to 2 dimensions by using the t-SNE algorithm, Fig. 10(a) illustrates the dimensional reduction results. There is a clear separation between different clusters, which ensures the accuracy of clustering and fault diagnosis. The result is compared with other unsupervised methods for extracting valid features of original data, which can be divided into linear and nonlinear methods. The linear methods include Locality Preserving Projections (LPP), Principal Component Analysis (PCA), as well as Independent Component Analysis (ICA). For the nonlinear method, SAE, Laplacian Eigenmaps (LE), Locally Linear Embedding (LLE), Multiple Dimensional Scaling (MDS), and Kernel PCA (KPCA) are included.

The comparison results are shown in Fig. 10, which shows that the proposed feature extraction method of SAE combining t-SNE is superior to other methods, which can make the separation between different clusters clear.

3) Clustering

The 2-dimension features are clustered, and the corresponding clustering centers are obtained by the CFSFDP algorithm. First, the local density ρ_i and distance δ_i of each data point is calculated to obtain the decision graph,as shown in Fig. 11(a). The cluster center points should have a large value of ρ_i and δ_i simultaneously. However, the cluster center points are difficult to determine only from the decision graph, so

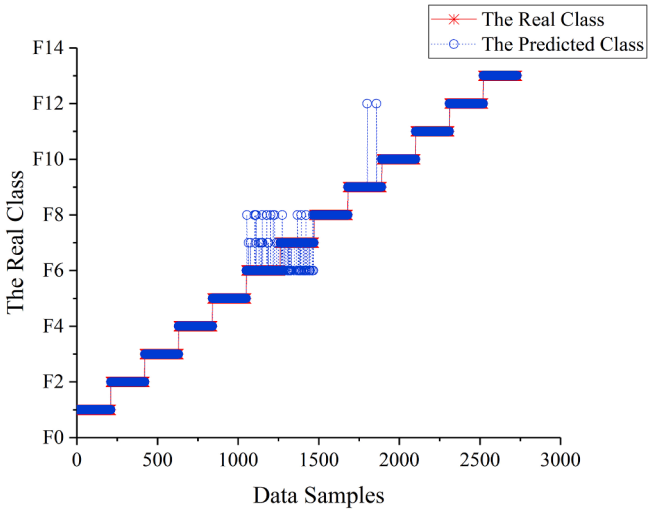


Fig. 12. The comparison of results on real and predicted categories on the simulation dataset.

the values of ρ_i and δ_i are comprehensively evaluated by γ decision function. The values of γ_i are larger, and the probability of cluster centers are greater. The γ_i are arranged in descending, as shown in Fig. 11 (b). As indicated from the diagram, there is a clear boundary between the cluster center points and other data points on the distribution of the γ_i. By setting the threshold, the cluster centers are automatically

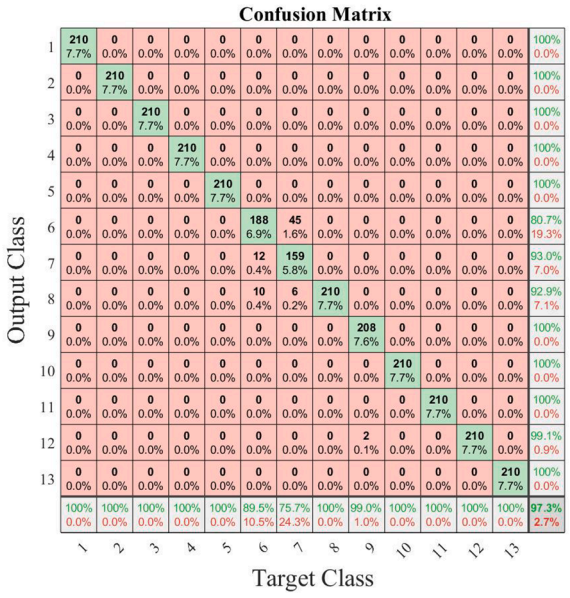


Fig. 13. The confusion matrix of the proposed method on the simulation dataset.

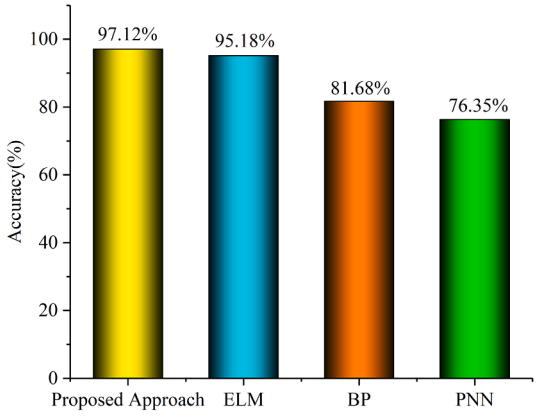


Fig. 14. Testing accuracy of proposed method in comparison with other machine learning methods on simulated dataset.

selected. After the γ_i exceeds the threshold, the point is determined as the clustering center. After the cluster centers are selected, the remaining data points are classified into clusters with higher density and closer distance. The clustering results and actual data sample distribution are shown in Fig. 11(c) and (d), which indicate that there is a clear distinction between different clusters. However, due to the high similarity between the output I-V curves of Degradation1 and Degradation2 at low irradiance, the fault data features are partially overlapped and difficult to distinguish.

4) Fault diagnosis and analysis

Based on the above analysis, 30% of the data samples are used to extract features and obtain cluster centers. The remaining 70% of unlabeled data samples were tested. The comparison between the real and predicted fault categories is shown in Figs. 12–14. It can be seen that the proposed fault diagnosis method has high fault diagnosis accuracy. The other fault types of the diagnosis accuracy are as high as 100% except for F6 (Degradation1), F7 (Degradation2), and F9 (Shading2). In order to further illustrate the performance of this fault diagnosis method, the classification confusion matrix of the test set is drawn in Fig.13 to obtain the classification accuracy of each fault type and the predicted fault types of the incorrect diagnosis. As can be seen from Fig. 11, the fault diagnosis accuracy of F7 is the lowest 75.7% and most of the data are misdiagnosed as F6. Since the I-V curves of F6 and F7 output at low irradiation have high similarity, which leads to the partial overlap of fault features. The overall fault diagnosis accuracy of the test set is 97.3%. To further test the overall performance of the proposed method, three machine learning methods are used for comparison, including the extreme learning machine (ELM), backpropagation (BP) and probabilistic neural network (PNN). Three methods have been successfully applied to fault diagnosis and detection of the PV array, but the limitation is the manual extraction of feature parameters. On the simulated dataset, the three methods are trained and tested using the same way with the proposed method. Finally, the comparison result is shown in Fig. 15, and the proposed method has better accuracy.

5.2. *Experimental verification and analysis*

1) Experimental platform

In order to further verify the performance and feasibility of the proposed method in practice, 5.28KW PV experimental platform is used as shown in Fig.1 to conduct field failure experiments. The different fault simulation methods on the filed PV array are shown in Fig. 15, which are the same as the simulation model mentioned in Section III. Partial shading is simulated by using different light transmittance of plastic to block PV modules. Short-circuit fault of PV modules is simulated by wire short-circuit PV modules. Short-circuit fault of the bypass diode is simulated by a wire short-circuit bypass diode. The degradation fault is simulated by sliding rheostat with high power in series of the PV array. The open-circuit fault of the bypass diode is simulated by disconnecting the bypass diode in the PV module. Other concurrent faults are simulated by combining the above fault types. Fault experiments were carried out from 2020/11/01 to 2021/1/16. Due to the incomplete scanning I-V curve of the inverter under low irradiation and the greater volatility, the irradiance in the experiment was greater than 280 W/m^2, and data for periods under high volatility of weather changes were eliminated. After excluding the abnormal data, a total of 2470 valid I-V curves were collected. 190 data samples were acquired for each fault type. Similarly, 30% of the data samples are randomly selected to mine the features of the data samples and find the clustering centers. Then, the performance of the fault diagnosis algorithm is tested by the remaining 70% of data samples.

2) Fault diagnosis and analysis

Similar to the simulation data, 30% of the data are used for feature extraction and dimensionality reduction by SAE and t-SNE, the results are shown in Fig. 16. The extracted two-dimensional data features are clustered by the CFSFDP clustering algorithm to obtain the clustering centers. The local density ρ_i and distance δ_i of the experimental data samples are calculated as shown in Fig. 17(a), and the descending order of the r values of the decision functions is shown in Fig. 17(b), from which it can be observed that the r-value distribution of the cluster center points and other data points has an obvious boundary. Then, 13 cluster centers are obtained by threshold judgment, which is consistent with the number of simulation fault types. The results of extracted two-dimensional features clustered by the CFSFDP algorithm and real fault samples are shown in Fig. 17(c) and (d), respectively. The clear distinction between different clusters indicates that the clustering algorithm has better performance. In addition to the confusion between Degradation1 and Shading1, some Shading1 data samples are incorrectly identified as Degradation1 clusters after clustering. Finally, the remaining 70% of the data samples were tested for diagnosis. The real fault types are compared with the predicted fault types, as shown in Fig. 18. Due to the experimental data samples have many interference factors, the prediction results are misdiagnosed in various fault types. To analyze the classification results more carefully of the experimental datasets, the confusion matrix is drawn as shown in Fig. 19. From the confusion matrix, it can be concluded that the average accuracy and error rates of this classification are 98.3% and 1.7%, respectively. The

(a) Short-circuit fault of PV mosules (b) Partial shading

(c) Degradation (d) Short-circuit fault of bypass diode

Fig. 15. Faults simulation of the field PV array.

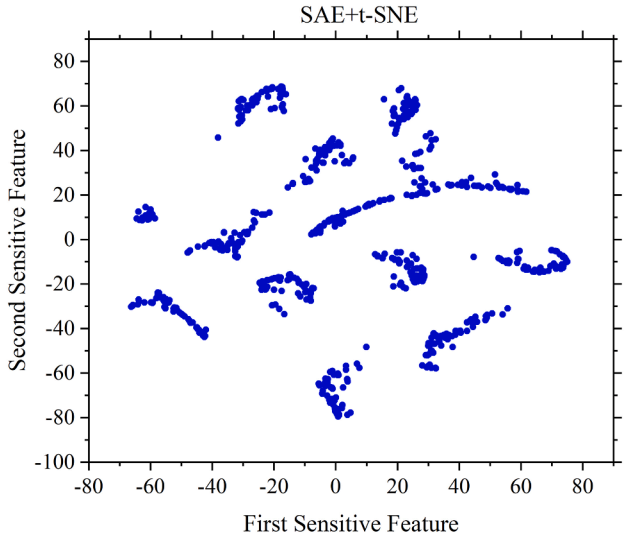


Fig. 16. Fault feature extraction and visualization results on the experimental dataset.

lowest diagnostic accuracy is F6 (Degradation1) 91.7%, and 7.5% of the data samples are misdiagnosed as F8 (Shading1). There are abnormal data points of experimental data which lead to misdiagnosis in many fault types. However, the high irradiance of the experimental data samples eliminates the confusion between F6 and F7 under low irradiation compared with the simulation data. Therefore, the average diagnostic accuracy is higher than that of the simulation dataset. It also shows that this fault diagnosis algorithm is suitable for high irradiance and has high fault diagnosis accuracy. Furthermore, the proposed method and other three different machine learning methods, including ELM, BP, PNN, are trained, tested and compared on the experimental dataset. Compared to other algorithms, the proposed method achieves the high testing accuracy, as shown in Fig. 20.

6. Conclusions

In this paper, the fault diagnosis method based on SAE and clustering with I-V curves is proposed to detect and classify normal and different faults accurately for PV array, including the short-circuit, degradation, partial shading, PSBO, PSSC, and SCD. The accurate simulation model of ICFFSM is established to simulate the I-V and P-V curves under normal and various faults. The unknown model parameters of the RSDM in ICFFSM are extracted by PSO from the measured I-V curves. Besides, the functional relationship between the model parameters and irradiance or temperature is fitted in different seasons to predict the model parameters for improving the accuracy of the ICFFSM. Moreover, the influence of different faults on the shape of I-V and P-V curves are analyzed based on the simulation model to demonstrate the feasibility of fault diagnosis with I-V and P-V curves. The proposed fault diagnosis method is detailed, the effective features of data samples are automatically extracted and visualized by SAE and t-SNE, respectively. After visualization, the features are clustered, and the cluster centers are obtained of different faults. Then, the membership degree of the test data between different clustering centers are calculated to fault classification. The simulation and experimental data are used to verify the performance of the proposed fault diagnosis method. Under the wide ambient conditions of simulation data, the confusion between degradation1 and degradation2 is easy to occur at low irradiation, which leads to low diagnostic accuracy and the average testing accuracy is 97.3%. In the experimental dataset, the average testing accuracy is 98.3%, which is higher than the simulation data because of the high irradiance. However, the measured data will be misdiagnosed in different degrees of various faults due to the volatility. Compared to other machine learning methods, including the ELM, BP, PNN, the proposed method achieves the high testing accuracy on the both simulation and experimental datasets. The simulation and the experimental results demonstrate that the proposed fault diagnosis method has high accuracy. However, the proposed method still requires a small number of labeled data samples to extract features and there is confusion between features of different faults. Therefore, it is necessary to further investigate more effective

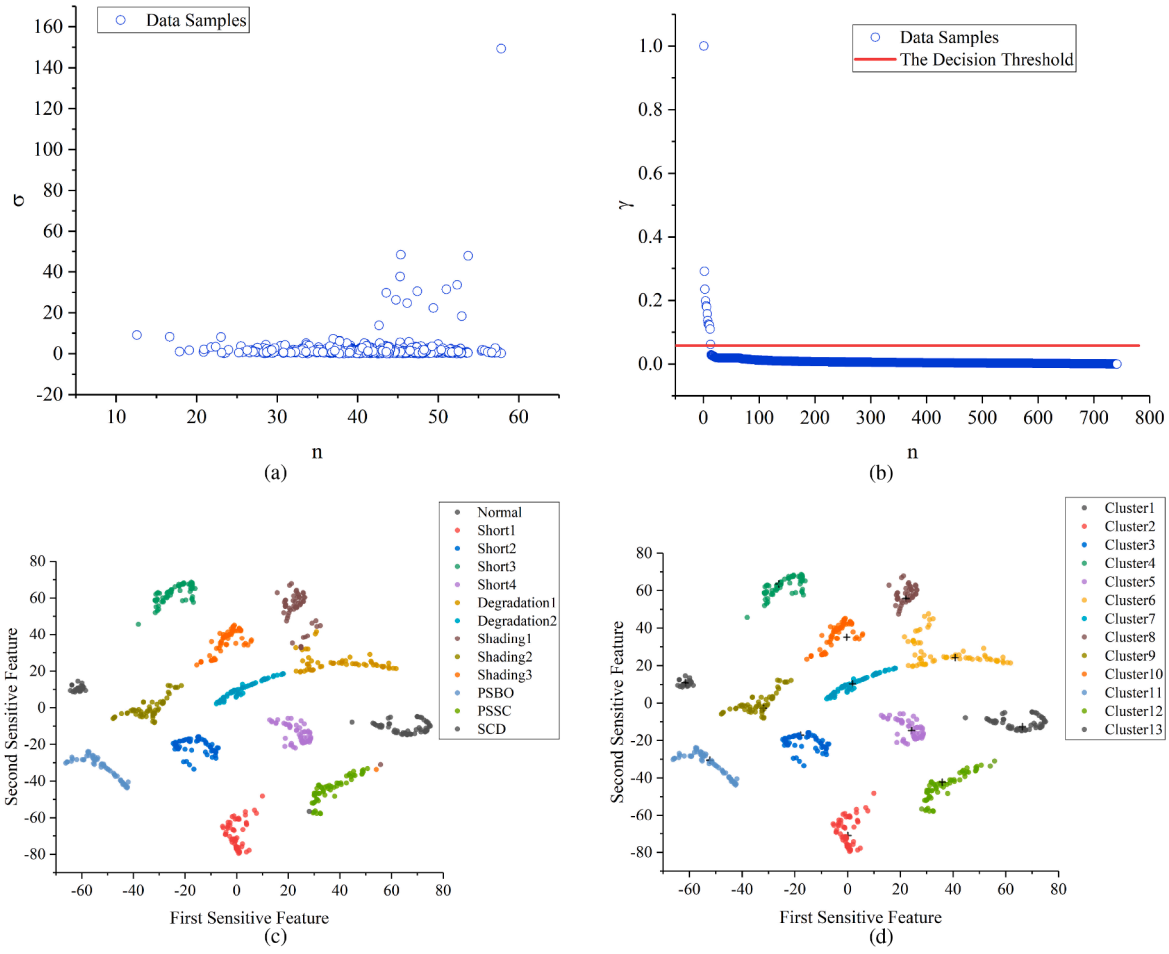


Fig. 17. (a) The decision graph for 2-D features of the experimental dataset; (b) The value of γ_i in decreasing order for the experimental dataset; (c) The clustering results; (d) The real distribution of 2-dimension features.

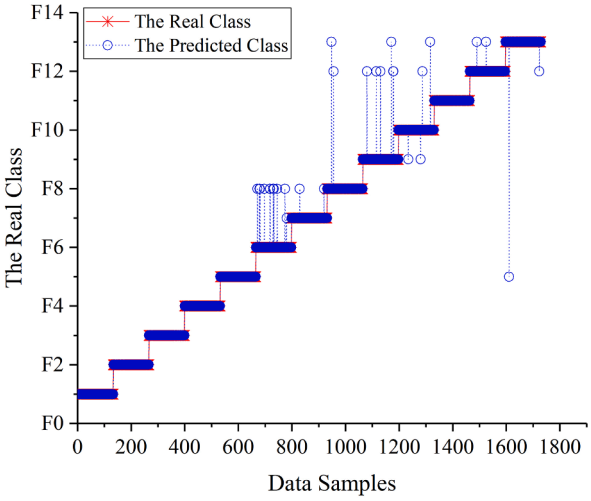


Fig. 18. The comparison of results on real and predicted categories on the experimental dataset.

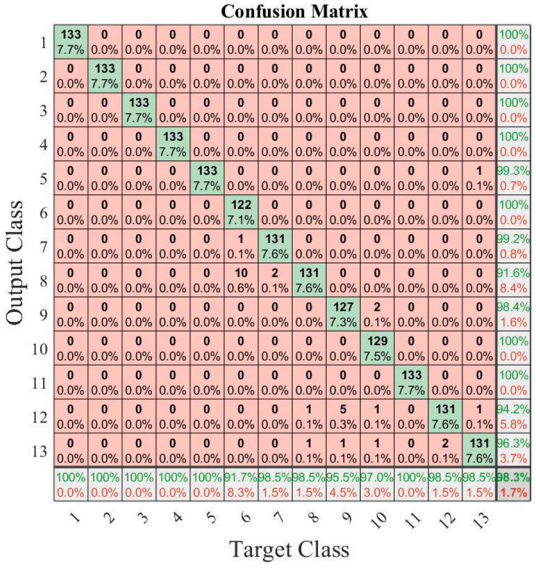


Fig. 19. The confusion matrix of the proposed method on the experimental dataset.

feature extraction technology to improve the diagnosis accuracy. For example, the clustering algorithm is embedded in the feature extraction technique, and the performance of the clustering result is used as the loss function of the feature extraction technique.

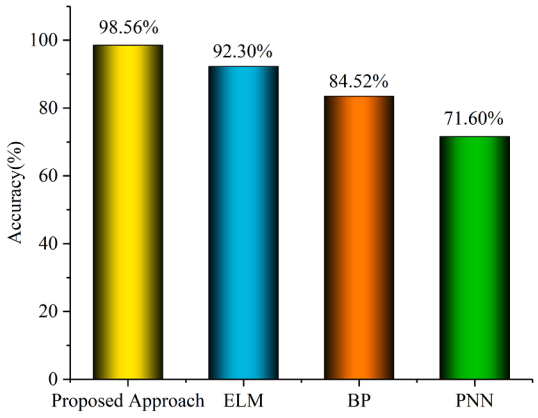


Fig. 20. Testing accuracy of proposed method in comparison with other machine learning methods on experimental dataset.

CRediT authorship contribution statement

Yongjie Liu: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - original draft, Writing - review & editing, Visualization. **Kun Ding:** Resources, Writing - review & editing, Supervision, Project administration. **Jingwei Zhang:** Conceptualization, Methodology, Investigation, Data curation, Writing - original draft, Writing - review & editing. **Yuanliang Li:** Methodology, Writing - original draft, Visualization. **Zenan Yang:** Software, Investigation, Visualization. **Wenming Zheng:** Resources. **Xiang Chen:** Formal analysis.

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 National Natural Science Foundation of China under Grant 51777059, the Qing Lan Project and the Fundamental Research Funds for the Central Universities (B200201051), Changzhou Sci & Tech Program (CJ20200074),Jiangsu Provincial Graduate Scientific Research and Innovation Plan (KYCX21_0464).

References

[1] Renewables 2020 analysis and forecast to 2025 International Energy Agency [online]. Available: https://www.iea.org/reports/renewables-2020/solar-pv#china.
[2] Triki-Lahiani Asma, Abdelghani Afef Bennani-Ben, Slama-Belkhodja Ilhem. Fault detection and monitoring systems for photovoltaic installations: A review. Renew Sustain Energy Rev 2018;82(3):2680–92.
[3] Mellit A, Tina GM, Kalogirou SA. Fault detection and diagnosis methods for photovoltaic systems: A review. Renew Sustain Energy Rev 2018;91:1–17.
[4] Tsanakas JA, Ha L, Buerhop C. Faults and infrared thermographic diagnosis in operating c-Si photovoltaic modules: A review of research and future challenges. Renew Sustain Energy Rev 2016;62(sep.):695–709.
[5] Cubukcu M, Akanalci A. Real-time inspection and determination methods of faults on photovoltaic power systems by thermal imaging in Turkey. Renew Energy 2020;147.
[6] Ding K, Zhang J, Ding H, et al. Fault detection of photovoltaic array based on grubbs criterion and local outlier factor. IET Renew Power Gen 2019;14(4).
[7] Fezai R, Mansouri M, Trabelsi M, et al. Ónline reduced kernel GLRT technique for improved fault detection in photovoltaic systems. Energy 2019;179:1133–54.
[8] Harrou F, Taghezouit B, Sun Y. Improved kNN-based monitoring schemes for detecting faults in PV systems. IEEE J Photovolt 2019;9(3):811–21.
[9] A Y C, B M M, C A C, et al. Simple and efficient approach to detect and diagnose electrical faults and partial shading in photovoltaic systems. Energy Conv Manage 2019;196:330–43.
[10] Chen Z, Wu L, Cheng S, et al. Intelligent fault diagnosis of photovoltaic arrays based on optimized kernel extreme learning machine and I-V characteristics. Appl Energy 2017. S0306261917305214.
[11] Spataru S, Sera D, Kerekes T, et al. Diagnostic method for photovoltaic systems based on light I-V measurements. Sol Energy 2015;119(sep.):29–44.
[12] Chen Z, Han F, Wu L, et al. Random forest based intelligent fault diagnosis for PV arrays using array voltage and string currents. Energy Conv Manage 2018;178 (DEC.):250–64.
[13] Huang JM, Wai RJ, Yang GJ. Design of hybrid artificial bee colony algorithm and semi-supervised extreme learning machine for PV fault diagnoses by considering dust impact. IEEE Trans Power Electr 2020;35(7):7086–99.
[14] Chine W, Mellit A, Lughi V, et al. A novel fault diagnosis technique for photovoltaic systems based on artificial neural networks. Renew Energy 2016;90:501–12.
[15] Garoudja E, Chouder A, Kara K, et al. An enhanced machine learning based approach for failures detection and diagnosis of PV systems. Energy Conv Manage 2017;151(nov.):496–513.
[16] Mao W, Feng W, Liu Y, et al. A new deep auto-encoder method with fusing discriminant information for bearing fault diagnosis. Mech Syst Signal Process 2021;150:107233.
[17] Mishra BK, Thakker D, Mazumdar S, et al. A novel application of deep learning with image cropping: a smart city use case for flood monitoring. J Reliab Intell Environ 2020;6(1):51–61.
[18] Targonski C, Bender MR, Shealy BT, et al. Cellular state transformations using deep learning for precision medicine applications. Patterns 2020. 100087.
[19] Chen Z, Chen Y, Wu L, et al. Deep residual network based fault detection and diagnosis of photovoltaic arrays using current-voltage curves and ambient conditions. Energy Conv Manage 2019;198. 111793-.
[20] Appiah AY, Zhang X, Ayawli BBK, et al. Long short-term memory networks based automatic feature extraction for photovoltaic array fault diagnosis. IEEE Access 2019:1.
[21] B X L A, B P L A, B S C A, et al. Fault diagnosis for photovoltaic array based on convolutional neural network and electrical time series graph. Energy Conv Manage 2019;196:950–65.
[22] Houssein EH, Emam MM, Ali AA, et al. Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review. Expert Syst Appl 2020.
[23] Chen Z, Han F, Wu L, et al. Random forest based intelligent fault diagnosis for PV arrays using array voltage and string currents. Energy Conv Manage 2018;178 (DEC.):250–64.
[24] Chen Z, Chen Y, Wu L, et al. Deep residual network based fault detection and diagnosis of photovoltaic arrays using current-voltage curves and ambient conditions. Energy Conv Manage 2019;198. 111793-.
[25] Huang JM, Wai RJ, Yang GJ. Design of hybrid artificial bee colony algorithm and semi-supervised extreme learning machine for PV fault diagnoses by considering dust impact. IEEE Trans Power Electr 2019;PP(99):1.
[26] Li Y, Ding K, Zhang J, et al. Á fault diagnosis method for photovoltaic arrays based on fault parameters identification. Renew Energy 2019;143:52–63.
[27] Liu Y, Ding K, Zhang J, et al. An improved code-based fault simulation model for PV module. In: 2020 12th IEEE PES asia-pacific power and energy engineering conference (APPEEC). IEEE; 2020.
[28] Ma T, Gu W, Shen L, et al. An improved and comprehensive mathematical model for solar photovoltaic modules under real operating conditions. Sol Energy 2019; 184(MAY):292–304.
[29] Routsolias IA, Batzelis, et al. An explicit PV string model based on the Lambert W function and simplified MPP expressions for operation under partial shading. IEEE Trans Sustain Energy 2014;5(1):301–12.
[30] Mai TD, De Breucker S, Baert K, et al. Reconfigurable emulator for photovoltaic modules under static partial shading conditions. Sol Energy 2017;141(JAN.): 256–65.
[31] Piccoli E, Dama A, Dolara A, et al. Experimental validation of a model for PV systems under partial shading for building integrated applications. Sol Energy 2019;183(MAY):356–70.
[32] Jovanovic, Raka, Barth, et al. PV panel single and double diode models: Optimization of the parameters and temperature dependence. Sol Energy Mater Solar Cells Int J Devot Photovolt Phototherm Photochem Sol Energy Conv; 2016.
[33] Manel M, Anis S, Faouzi MM. Particle swarm optimisation with adaptive mutation strategy for photovoltaic solar cellmodule parameter extraction. Energy Conv Manage 2018;175:151–63.
[34] Ding K, Bian X, Liu H, et al. A MATLAB-Simulink-Based PV module model and its application under conditions of nonuniform irradiance. IEEE Trans Energy Conv 2012;27(4):864–72.
[35] Orioli A, Di Gangi A. A procedure to evaluate the seven parameters of the two-diode model for photovoltaic modules. Renew Energy 2019;139(AUG.):582–99.
[36] Li D, Li Z, Sun K. Development of a novel soft sensor with long short-term memory network and normalized mutual information feature selection. Math Prob Eng 2020;2020.
[37] Maaten LJPVD, Hinton GE. Visualizing high-dimensional data using t-SNE. J Mach Learn Res 2008;9:2579–605.
[38] Liu G, Zhu L, Wu X, et al. Time series clustering and physical implication for photovoltaic array systems with unknown working conditions. Sol Energy 2019; 180:401–11.
[39] Zhao, Qiang, Shao, et al. A new PV array fault diagnosis method using Fuzzy C-mean clustering and fuzzy membership algorithm. Energies; 2018.
[40] Zhu H, Lu L, Yao J, et al. Fault diagnosis approach for photovoltaic arrays based on unsupervised sample clustering and probabilistic neural network model. Sol Energy 2018;176:395–405.

[41] Ester M, Kriegel HP, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. AAAI Press; 1996.
[42] Rodriguez A, Laio A. Clustering by fast search and find of density peaks. Science 2014;344(6191):1492.
[43] Min X, Huang Y, Sheng Y. Automatic determination of clustering centers for lustering by fast search and find of density PeaksJ]. Math Prob Eng 2020;2020(34): 1–11.