


ORIGINAL RESEARCH

FSR-SSL: A fault sample rebalancing framework based on semi-supervised learning for PV fault diagnosis

Qi Liu^{1,2,3} | Xinyi Wang^{1,2,3}  | Bo Yang^{1,2,3}  | Zhaojian Wang^{1,2,3} | Yuxiang Liu^{1,2,3} | Xinping Guan^{1,2,3}

¹Department of Automation, Shanghai Jiao Tong University, Shanghai, China

²Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai, China

³Shanghai Engineering Research Center of Intelligent Control and Management, Shanghai, China

Correspondence

Bo Yang, Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China.
Email: bo.yang@sjtu.edu.cn

Funding information

National Key Research and Development Program of China, Grant/Award Number: 2018YFB1702300; NSF of China, Grant/Award Numbers: 61731012, 62103265, 92167205

Abstract

Photovoltaics face the threat of many potential faults in daily operation, which calls for accurate fault diagnosis to avoid huge economical losses. This paper investigates practical and troublesome scenarios, where the photovoltaics station has a large number of unlabeled samples and only a few are labelled. In the labelled set, the sample sizes of different types are unbalanced. To this end, a new fault sample rebalancing framework based on semi-supervised learning (FSR-SSL) is proposed. Specifically, a dual-threshold selection mechanism is proposed to choose trusted pseudo-label samples from unlabeled data to expand the training set. Moreover, a fault sample rebalancing strategy is designed to further filter the obtained trusted pseudo-label samples, thereby flexibly adding different amounts of pseudo-label data to various types. As the training rounds increase, the fault samples are gradually rebalanced and the model learning bias caused by type imbalance is well overcome. The extensive numerical experiments show that the proposed FSR-SSL method reaches 99% accuracy. Compared with existing methods, the accuracy is increased by up to 33%.

1 | INTRODUCTION

Serious issues such as environmental pollution, energy shortages and sustainable development are attracting more and more attention [1, 2]. For most countries in the world, fossil fuel consumed by power generators is one of the primary reasons for carbon emissions [3]. From research on greenhouse gas emissions (mainly carbon dioxide), more than 40% of carbon emission is generated by the combustion of fossil fuel during power generation [4]. As a renewable energy source, solar energy has been regarded as a promising direction towards low-carbon development, where photovoltaic (PV) is one of the main ways [5]. PV arrays are the fundamental components of PV systems and mainly operate in outdoor conditions. Failures of PV arrays may cause a huge energy loss and even risk of fires [6]. Thus it is essential to accurately detect and diagnose fault types to maintain PV system reliability and sustainable power generation.

Recently, many diagnosis methods are proposed for PV systems. The methods can be mainly classified into three groups including statistical signal and processing methods [7–10],

electrical characteristics-based methods [11–15] and machine learning-based (ML) methods [16–20]. However, statistical signal and processing methods and electrical characteristics-based methods rely heavily on prior knowledge and manual feature extraction, which are difficult to be generalised in distributed PV systems. Thus the ML-based methods that extract fault features automatically are being widely used for fault diagnosis. In ML-based methods, a large number of labelled samples are required for training, which is difficult and expensive to obtain in industrial applications, especially for PV stations due to the complex system structure. In the daily operation process, the PV stations only record electrical data such as current and voltage, and most of the collected samples are unlabelled. Moreover, these huge amounts of samples need to be labelled before training a diagnosis model. Semi-supervised learning (SSL) is a key issue in the field of machine learning, which combines supervised learning and unsupervised learning. It is an effective method to solve the above problems, which can train a diagnosis model with a few labelled samples and large amounts of unlabeled samples [21, 22].

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *IET Renewable Power Generation* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology.

However, most of the existing SSL-based PV fault diagnosis methods directly use the limited labelled samples to train the initial model, and then develop unlabelled data based on the model to make full use of the data to achieve fault diagnosis. They either ignore the distribution of labelled samples, or assume that the number of types in the labelled set is relatively balanced. Although these methods are very important in the fault diagnosis of PV stations, and the accuracy is very high, it is also significant to study scenarios with only a small number of labeled samples and imbalanced types. In fact, these scenarios are more in line with the actual operation of PV stations. The PV station operates normally most of the time, which results in far more samples in the normal state than in the fault state. In addition, the frequency of various faults is different due to the environment and the system itself. Thus the collected samples are seriously type-imbalanced, which causes overfitting of the trained model to the normal state. This means that the model may fail to fully extract the features of different faults. The implicit features of the model trained on type-imbalanced data are summarised [23]. The model acquires high recall but low precision on majority types while acquiring low recall but high precision on minority types. Inspired by this, a new ML-based framework is designed to enable photovoltaic power plants to achieve accurate fault diagnosis when the number of label samples is small and the types are unbalanced.

To design a new fault diagnosis framework for PV stations, the following three challenges need to be addressed: fault feature extraction, unlabelled data utilisation, and fault sample rebalancing. First, the collected samples are I - V characteristic curves of the PV system in various operation states. The data dimensions of these samples are different, and the sampling points are unevenly distributed on the curve. For these reasons, they cannot be directly used for model training, and appropriate data preprocessing methods should be adopted. Second, considering that a large amount of data is unlabelled, it is necessary to design an effective pseudo-label prediction mechanism to expand the training set. Third, due to the unbalanced types of labeled samples in the initial training set, the model's learning degree of various fault types varies greatly. Therefore, it is necessary to propose a rebalancing method for fault samples to avoid over-fitting and under-fitting on different types.

In this paper, we propose a new SSL method that combines the dual-threshold selection mechanism and fault sample rebalancing strategy for PV fault diagnosis. In the case that the initial training set has only a small number of unbalanced labelled samples, the model learning bias caused by type imbalance is overcome by improving the process of adding pseudo labels. This method makes full and effective use of unlabeled data to achieve the rebalancing of types and reaches high accuracy under various data distributions and weather conditions.

Our major contributions are summarised as follows:

- A new fault sample rebalancing semi-supervised learning (FSR-SSL) framework is proposed for PV fault diagnosis. It contains not only data preprocessing methods and a plain CNN model applicable to PV data, but also improved SSL

methods. This framework achieves accurate fault diagnosis with only a few labelled samples, even if the sample sizes of different types are significantly unbalanced and the distribution of labeled data is different from that of unlabelled data.

- An improved semi-supervised learning algorithm is proposed. Specifically, a dual-threshold selection mechanism is proposed to set different confidence thresholds for the types with large and small numbers of samples, which is used to effectively select pseudo-label samples with high accuracy. Moreover, a fault sample rebalancing strategy is designed to selectively add the obtained trusted pseudo-label samples to the training set according to the proportion of different types of labelled data. Thus, the model learning bias caused by type imbalance is well overcome.
- The extensive numerical experiments show that the accuracy of the proposed FSR-SSL method reaches 99% under various experimental data distributions. Compared with the existing methods, the accuracy is increased by up to 33%.

The remainder of this paper is organised as follows. Section 2 reviews the related works. Section 3 describes the system model. Section 4 introduces the FSR-SSL based PV fault diagnosis method. Section 5 presents the numerical simulation results. Finally, Section 6 concludes the paper.

2 | RELATED WORKS

In recent years, different fault diagnosis methods for PV arrays has been proposed. Vergura et al. [24] utilised statistical signals to detect three kinds of common faults in PV arrays. Then the method was improved and applied in the direct-current side of PV arrays [10]. Stauffer et al. [25] introduced an easy method to detect faults by comparing measured power and the value simulated. However, the type of faults cannot be classified. In [26], different parameters (the measured power and voltage) were chosen for comparison to detect and localise three types of faults in PV arrays. Besides, the I - V characteristic analysis is widely used in PV systems. It contains many electrical parameters that can effectively reflect the operating conditions of PV modules. For instance, some common types of faults were investigated based on observation of the I - V characteristic anomalies. However, the fault features are manually extracted in the above methods, which cause additional costs and also reduce the generalisation performance. Different from them, many emerging diagnosis methods are ML-based, which is data-driven and the fault features are automatically extracted relying on massive samples [27–30]. W.Chine et al. [27] developed two different algorithms based on artificial neural networks (ANN) for classifying eight types of faults. In ref. [28], a diagnosis method based on theoretical curves modelling and fuzzy classification system was proposed for detecting short-circuit and hot spots. The minimum detection accuracy achieves 98.8%. Consider the noise of the collected signal, a fault detection algorithm based on multi-resolution signal decomposition and a two-stage support vector machine (SVM) classifier is proposed in ref. [29]. To obtain a higher accuracy in low mismatch levels

and high fault impedance, a method combining the hierarchical classification (HC) platform and machine learning (ML) is proposed for line–line (LL) and line–ground (LG) faults [30]. However, the traditional ML-based methods mentioned above ignore the fact that a large number of samples are unlabelled in actual PV stations, which makes the model based on supervised learning fail to reach high accuracy.

To address this issue, SSL is a promising method, which has achieved excellent performance with a large amount of unlabelled data [31]. It has been used for fault diagnosis in several fields [32–38]. In ref. [32], a diagnosis algorithm combining artificial bee colony algorithm and semi-supervised extreme learning machine was proposed for PV systems. It used a few labelled samples and historical unlabelled samples to train the model. Zhao et al. developed a graph-based SSL model, which used a few labelled samples to achieve model training and visualisation [39]. Yan et al. [33] proposed a semi-supervised approach to detect and diagnose air handling unit faults. The training set contained 8000 normal samples and only 30 samples of each fault type, and the accuracy reached 80%. A scheme based on information fusion and SSL was introduced for diagnosis of gear faults in ref. [34]. The information fusion module was used to integrate multiple sensory streams and the SSL module can improve the diagnostic efficiency. Yu et al. [35] proposed a SSL-based method for intelligent diagnosis of rolling bear. In ref. [36], another diagnosis method for rolling bear was introduced based on data augmentation and metric learning. Jian et al. [37] proposed a method for industrial fault diagnosis based on active learning and SSL. Different from traditional SSL, unlabelled data with larger uncertainty were selected for expert annotation to improve the accuracy. Considering noisy labels in normal working data sets, fault diagnosis with noisy labels was transformed into a SSL procedure. A two-stage SSL method was proposed for fault diagnosis of rotating machinery in ref. [38]. However, the methods mentioned above do not consider the problem of type imbalance in labeled samples.

Although SSL has been extensively studied, work on class rebalancing is still insufficient. To solve the problem of pseudo-label quality degradation, Kim et al. [40] designed a convex optimisation scheme to refine pseudo-labels generated from class-imbalanced biased models to be consistent with the true distribution. However, the construction and solution of this convex optimisation problem are complex. Yang et al. [41] proposed a semi-supervised imbalanced learning framework with the loss function of the classifier re-weighting, and the distribution of samples in the training set is remodelled by meta-learning. However, the internal iterations of meta-learning consume a lot of training time and resources. Zhang et al. [42] proposed a framework for implementing fair SSL by introducing a resampling method in the preprocessing stage, and the impact of mispredicted pseudo-labels on model accuracy was reduced through ensemble learning. However, the resampling method is difficult to take into account the sample size and the class balance at the same time, and it is prone to problems such as underfitting and overfitting. Wei et al. [23] proposed a class-rebalancing self-training scheme (CReST) that selected pseudo-labelled samples of various classes at different frequencies according to the esti-

mated class distribution. A progressive distribution alignment was introduced to further alleviate model bias. However, it only uses the initial labelled data distribution to predict the distribution of unlabeled data and ignores the problem of pseudo-label quality degradation under imbalanced classes.

Since the operating conditions of PV stations are constantly changing, the distributions of labelled samples and unlabeled samples may be quite different. In addition, considering the frequency of different faults is not fixed and the high cost of manual labelling in the actual maintenance process, the imbalance in the labeled samples is exacerbated, resulting in lower data quality of the predicted pseudo-labeled samples. Thus, the above methods are unable to be directly applied and it is important to propose a new SSL method to solve the PV fault diagnosis problem studied in this paper.

3 | SYSTEM MODEL

In this paper, the fault diagnosis problem of a PV station is studied. As shown in Figure 1, the PV station is composed of several arrays. It has data collection, storage and processing capability. Specifically, the I – V characteristic curves, corresponding temperature and irradiance of each array under normal and various fault states can be collected by the I – V tester equipped with an environmental tester. We mainly focus on three common PV array faults in this paper: 1) short-circuit faults, defined as the accidental connection or low impedance between two points in the PV array [43]; 2) degradation faults, defined as the increase in the equivalent series resistance or the decrease in the parallel resistance after certain time of operation [44]; 3) partial shading faults, defined as the reduction of actual effective irradiance of the PV module due to surface dust, clouds, or other objects blocking the sunlight [45].

It is worth noting that this paper considers very practical and troublesome scenarios. The data set D collected by the PV station has only a few labelled samples, and the others are unlabelled. Moreover, there is an imbalance in the number of various types, which is not conducive to the accurate extraction of fault features. In fact, this is very common in the operation of actual PV stations. On the one hand, the cost of collecting labeled samples is high. On the other hand, the PV station operates normally most of the time, which results in far more samples in the normal state than in the fault states. In addition, the frequency of various faults is different due to the environment and the system itself. To effectively analyse the above problems, the labeled samples in the data set D are defined as a labelled set $\mathcal{L} = \{(\mathbf{x}_n, \mathbf{y}_n) : n \in (1, \dots, N)\}$, where $\mathbf{x}_n \in \mathbb{R}^d$ are the I – V characteristic data used for training and $\mathbf{y}_n \in \{1, \dots, I\}$ are regarded as labels. N represents the number of labelled samples. It is worth noting that the types in the labelled set \mathcal{L} are arranged in descending order of cardinality. The number of training samples in \mathcal{L} of label i is denoted as N_i , i.e. $\sum_{i=1}^I N_i = N$ and $N_1 \geq N_2 \geq \dots \geq N_I$. In the meanwhile, the remaining samples constitute an unlabelled set $\mathcal{U} = \{\mathbf{u}_m \in \mathbb{R}^d : m \in (1, \dots, M)\}$, where M is the number of unlabelled samples and $M > N$.

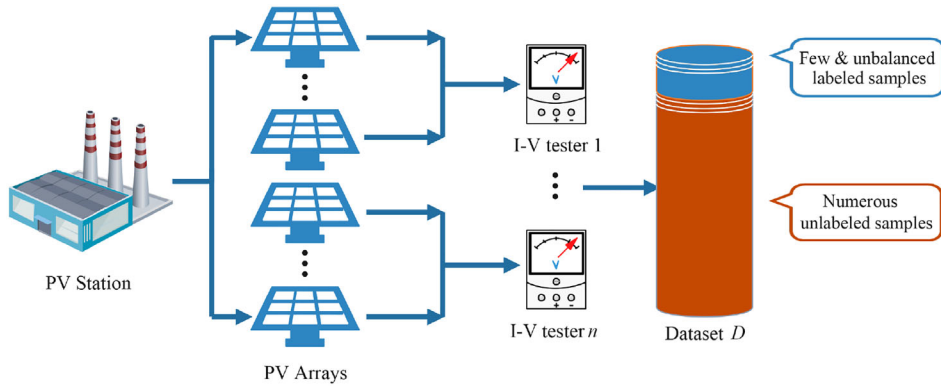


FIGURE 1 The structure of the PV station

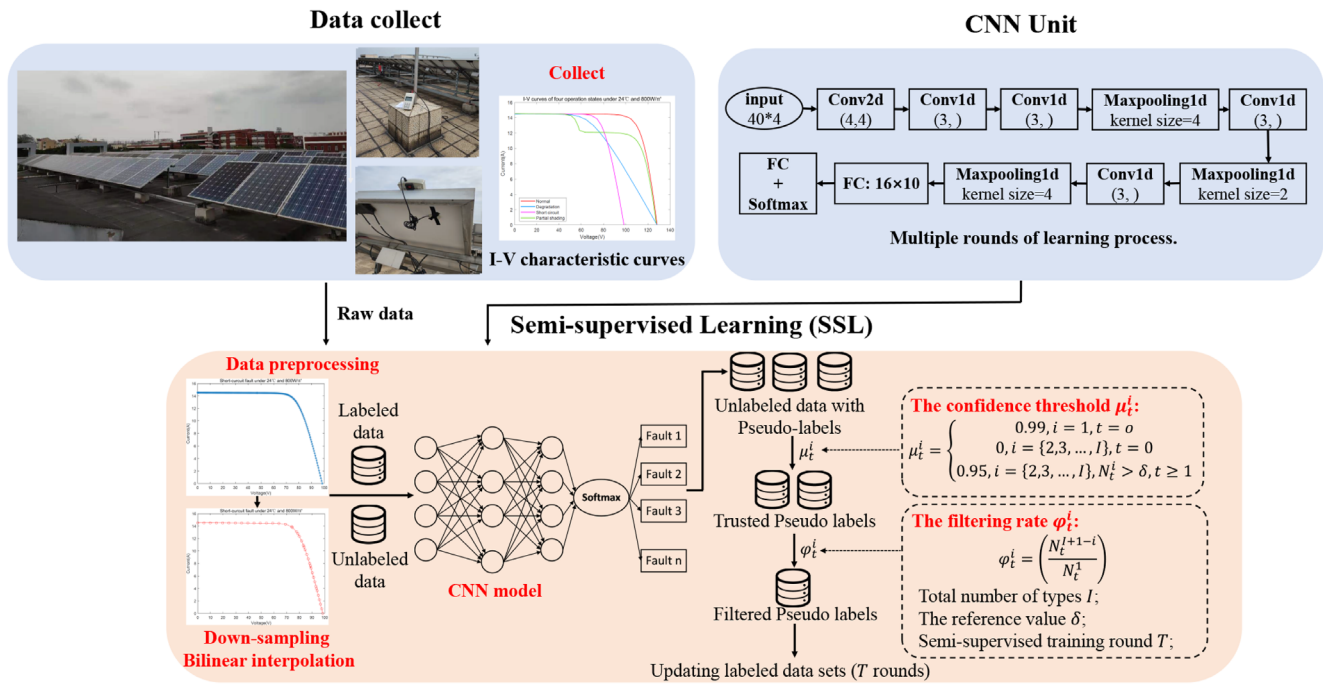


FIGURE 2 The general framework of the proposed method

4 | FSR-SSL FOR FAULT DIAGNOSIS

In this section, we will specifically introduce the implementation of the PV fault diagnosis based on the proposed FSR-SSL method. The main procedure of the proposed FSR-SSL method is illustrated in Figure 2. It is mainly composed of three parts, including preprocessing the original $I-V$ characteristic curves, constructing the CNN to train the fault diagnosis model, and designing a new fault sample rebalancing strategy to improve the accuracy of SSL.

4.1 | Data preprocessing

The performance of PV modules under various fault conditions will be affected, resulting in obvious differences in the $I-V$ characteristic curves [46]. For PV arrays, only the sampling points

where the current and voltage values are greater than or equal to zero in the collected $I-V$ characteristic curves are used. The upper limits of the current and voltage values correspond to the short-circuit current and open-circuit voltage values respectively, which are limited by the actual PV array scale and system operating states. Specifically, the $I-V$ characteristic curves of the normal and three fault states under the irradiance of 800 W/m^2 and the temperature of 24°C are shown in Figure 3.

It is worth noting that the original $I-V$ characteristic curves need to be preprocessed before they are used for fault diagnosis. The specific reasons are as follows. On the one hand, the original $I-V$ characteristic curve has redundant data, which wastes computing resources. On the other hand, the original curve has different numbers of sampling points, which have varying input dimensions and fail to train the CNN model. Moreover, the distribution of sampling points on the curve is uneven, which is not conducive to fault feature extraction. To overcome these

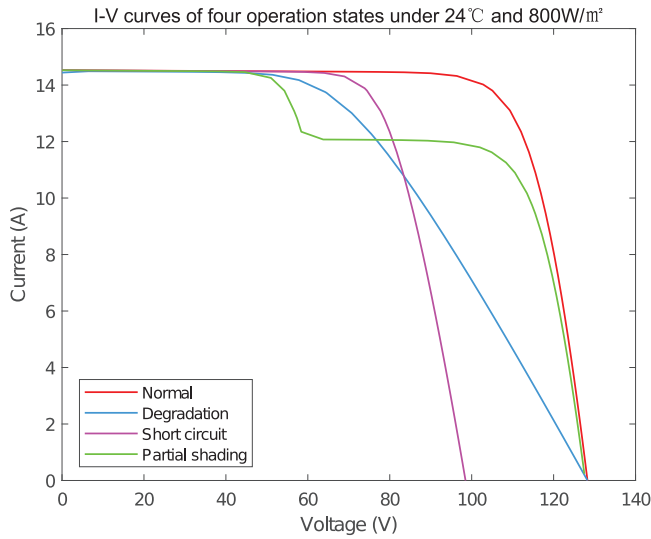


FIGURE 3 I - V characteristic curves under different operation states

challenges, the down-sampling and bilinear interpolation methods are combined to process and enhance the original I - V characteristic curves. Considering calculation efficiency and effective feature information, 20 new data points V_{sample_i} and I_{sample_i} are sampled equidistantly in the range of $[0, V_{\text{oc}}]$ and $[0, I_{\text{sc}}]$ respectively, $i = 1, 2, \dots, 20$. The corresponding voltage value V_n and current I_n , $i = 1, 2, \dots, 20$, are calculated from the original sampling points based on bilinear interpolation. Specifically, the current and voltage of 40 new sampling points are calculated by Equations (1) and (2) respectively. After that, they are sorted in ascending order of voltage.

$$I_n = \frac{(V_{\text{sample}_i} - V_{\text{left}}) \cdot I_{\text{right}} + (V_{\text{right}} - V_{\text{sample}_i}) \cdot I_{\text{left}}}{V_{\text{right}} - V_{\text{left}}}, \quad (1)$$

$$V_n = \frac{(I_{\text{sample}_i} - I_{\text{left}}) \cdot V_{\text{right}} + (I_{\text{right}} - I_{\text{sample}_i}) \cdot V_{\text{left}}}{I_{\text{right}} - I_{\text{left}}}, \quad (2)$$

where V_{left} , I_{left} and V_{right} , I_{right} respectively represent the voltage and current values of the closest left and right samples to V_{sample_i} and I_{sample_i} .

To further visually present the effect of the data preprocessing, the I - V characteristic curves of short-circuit faults before and after data preprocessing under the irradiance of 800 W/m² and the temperature of 24°C are shown in Figure 4. Based on the above data preprocessing method, the original curve is greatly compressed from about 400 data points to 40, which are evenly distributed on the curve. The I - V characteristic curve after resampling can be expressed as a 40×2 I - V vector. Taking into account the effects of temperature and irradiance, they are added to the I - V vector as a 40×2 environmental vector. Finally, the reconstructed 40×4 two-dimensional matrix is used as a sample for training the CNN later.

4.2 | CNN unit

The proposed FSR-SSL method involves multiple rounds of model training. Therefore, it is necessary to establish an effective neural network model to accurately extract fault features from training samples. Since the sample after data preprocessing is a 40×4 feature matrix, a plain CNN proposed in ref. [47] is selected to automatically extract the two-dimensional data features and train the fault diagnosis model in each round. The plain CNN consists of a 2D CNN layer, 1D CNN layers, Max-pooling layers, a fully connected (FC) layer, and the Softmax function. The detailed architecture is listed in Figure 5. Based on the data preprocessing and the CNN unit, the following subsection specifically introduces the fault sample rebalancing strategy.

4.3 | Fault sample rebalancing strategy

The proposed fault sample rebalancing strategy is mainly composed of two parts: self-training method and flexible sample filtering mechanism. The self-training is a widely used method in SSL [48, 49]. It involves multiple rounds of SSL process. In the beginning, it uses the initial labelled data to train a fault diagnosis model built by the CNN mentioned above. Then, this model is used to add pseudo labels for unlabelled data. Next, the pseudo-label data is added to the training set in the next round. After that, it uses the initial labeled data and the added pseudo-label data to train and improve the fault diagnosis model, so as to accurately add pseudo labels to the remaining unlabelled data in subsequent rounds. In this way, the model can converge to a high accuracy after multiple rounds.

However, the self-training method may add false pseudo labels to unlabelled data due to the inability to accurately extract fault features. Especially when the initial labelled samples are few while there is an imbalance in the number of various types. In fact, this is very common in the operation of actual PV stations, because the cost of collecting labelled samples is high, and the frequency of various faults is different and significantly less than the normal state. As revealed in the literature [23], the model trained on type-imbalanced data has two implicit features. First, the type with a small sample size has a low recall rate, but very high precision. Second, the type with a large number of samples has a high recall rate, but low precision. To solve the above problems and realise the balance of training samples, a new flexible sample filtering mechanism is proposed. Specifically, two effective modifications are proposed to the traditional self-training method.

First, in each round of model training, the proposed method only filters out a part of the data for addition instead of adding all pseudo-label samples to the training set. $\hat{\mathcal{F}}_t$ is defined as the pseudo-label data filtered in the t th round of training, and $\hat{\mathcal{F}}_t = \{\hat{\mathcal{F}}_t^1, \hat{\mathcal{F}}_t^2, \dots, \hat{\mathcal{F}}_t^I\}$. Obviously, we have:

$$\hat{\mathcal{F}}_t \subset \hat{\mathcal{P}}_t, \quad (3)$$

$$\mathcal{L}_{t+1} = \mathcal{L}_t \cup \hat{\mathcal{F}}_t, \quad (4)$$

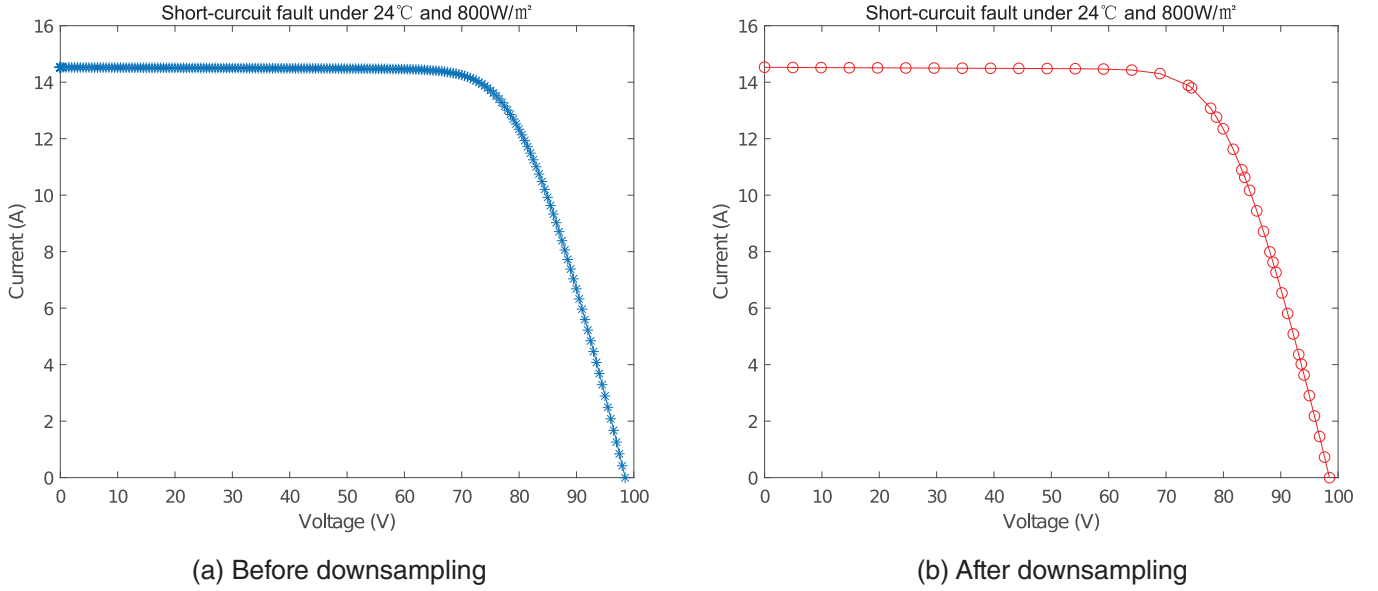


FIGURE 4 I - V characteristic curves of the short-circuit fault

where $\hat{\mathcal{P}}_t$ is the pseudo-label set predicted in round t , and $\hat{\mathcal{P}}_t = \{\hat{\mathcal{P}}_t^1, \hat{\mathcal{P}}_t^2, \dots, \hat{\mathcal{P}}_t^I\}$. $\hat{\mathcal{L}}_t$ denotes the labelled set in round t . The filtering principle for different types of pseudo-label data is that the lower the frequency of type i in the initial labelled samples, the more pseudo-label samples predicted to be type i are retained. Conversely, the higher the frequency of type i' , the less pseudo-label samples are retained. Specifically, the pseudo-label samples predicted to be type i in round t are included into $\hat{\mathcal{F}}_t$ at the rate of

$$\psi_t^i = \left(\frac{N_t^{I+1-i}}{N_t^1} \right), \quad (5)$$

where N_t^i is defined as the sample size of type i in the training set in the t th round. As experiment 6 in Section 5.3, for the type with the least number of samples, i.e. the degradation state, we have $\psi_1^4 = \left(\frac{N_1^{4+1-4}}{N_1^1} \right) = 1$. While for the most majority type, i.e. the normal state, we have $\psi_1^1 = \left(\frac{N_1^{4+1-1}}{N_1^1} \right) = \left(\frac{0.05}{0.9} \right)$. Obviously, the improved method will only retain a few pseudo-label samples of the most majority type at the initial moment. Moreover, the pseudo-label filtering rate ψ_t^i changes dynamically with the training round t , and is adjusted in a timely manner and flexibly based on the distribution of the previous round of the training set.

Second, the proposed method designs a dual-threshold selection mechanism when predicting pseudo labels for unlabelled data. Specifically, the confidence threshold for selecting pseudo label i in round t is defined as μ_t^i . At the beginning, the confidence threshold for predicting a normal state is set to 0.99, while other types are set to 0. Considering that most of the initial labelled data are normal samples, the model is easy to overfit, which leads to a high recall rate. Only if the Softmax

output probability is large enough can it be regarded as an accurate prediction. On the contrary, due to the small number of other types, the recall rate is very low but the accuracy is high. That is, as long as these types are predicted, they are considered correct. In addition, δ is set as the reference value of the sample size. As the training rounds increase, the pseudo-label data is continuously added to the training set. If the sample size of a certain type exceeds this value, the confidence threshold is adjusted to 0.95. This is because the samples of this type are sufficient at this time and no longer meets the prerequisites for low recall rate. Otherwise, overfitting may occur. Based on the above analysis, we have:

$$\mu_t^i = \begin{cases} 0.99, & i = 1, t = 0 \\ 0, & i \in \{2, 3, \dots, I\}, t = 0 \\ 0.95, & i \in \{1, 2, 3, \dots, I\}, N_t^i > \delta, t \geq 1. \end{cases} \quad (6)$$

Through the proposed FSR-SSL methods, the difference between normal and fault samples in the training set can be rebalanced after multiple rounds of self-training process. Based on effective filtering and dual-threshold selection mechanism, pseudo labels are accurately added to expand the training set. The model makes full use of labelled and unlabelled data to extract fault features, thereby greatly improving the accuracy of fault diagnosis. The detailed steps of the proposed FSR-SSL method are shown in Algorithm 1.

5 | SIMULATION STUDIES

In this section, the physical platform and numerical simulations are used to verify the effectiveness of the proposed FSR-SSL method. First, a small number of samples are collected

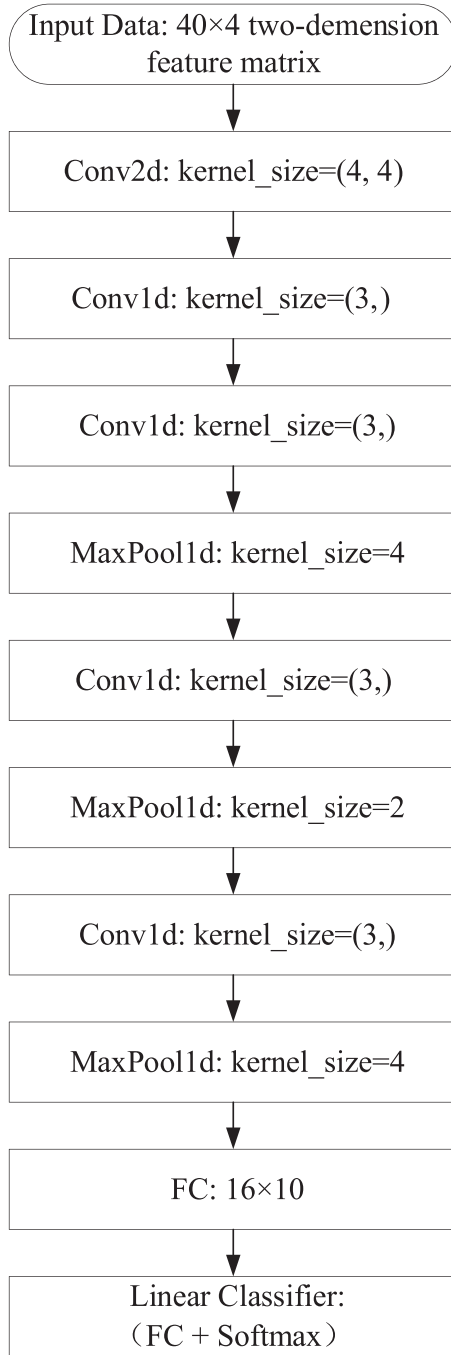


FIGURE 5 The structure of the plain CNN

based on the actual PV arrays to validate the data of numerical simulations. Then, numerical simulations are carried out to prove that the proposed method is suitable for various weather and fault conditions.

5.1 | Actual data collection

The fault diagnosis method proposed in this paper is based on the I - V characteristic curves of PV modules. To ensure the effectiveness of the data collected by the numerical simulations,

ALGORITHM 1 FSR-SSL for PV fault diagnosis

Input:

Initial labelled training set $\hat{\mathcal{L}}_0$, unlabelled data set $\hat{\mathcal{U}}_0$;
 Initial confidence threshold μ_0^i , filtering rate ψ_i^j , and the reference value δ ;
 Initial CNN parameters $w_0, m_0, v_0, l_r, \epsilon, \beta_1$ and β_2 ;
 Semi-supervised training round T
 Total number of types I ;

Output: Accurate fault diagnosis model w_T

```

1:  for  $t = 1$  to  $T$  do
2:      if  $t = 1$  then
3:          Only use the initial labelled set  $\hat{\mathcal{L}}_0$  to train the model;
4:           $w_1 = w_0 - l_r \frac{\hat{m}_t}{\sqrt{v_t + \epsilon}}$ ;
5:      else
6:          Use the updated labeled set  $\hat{\mathcal{L}}_{t-1}$  to train the model;
7:           $w_t = w_{t-1} - l_r \frac{\hat{m}_{t-1}}{\sqrt{v_{t-1} + \epsilon}}$ ;
8:      end if
9:      Use the trained model to predict the current unlabelled set  $\hat{\mathcal{U}}_t$ ;
10:     for  $i = 1$  to  $I$  do
11:         Select the pseudo-label samples  $\hat{\mathcal{P}}_t^i$  based on the confidence threshold  $\mu_t^i$ ;
12:          $\mu_t^i = \begin{cases} 0.99, & i = 1, t = 0 \\ 0, & i \in \{2, 3, \dots, I\}, t = 0 \\ 0.95, & i \in \{1, 2, 3, \dots, I\}, N_t^i > \delta, t \geq 1. \end{cases}$ ;
13:         Filter the pseudo-label samples  $\hat{\mathcal{F}}_t^i = \psi_t^i \cdot \hat{\mathcal{P}}_t^i$ ;
14:          $\psi_t^i = \left( \frac{N_t^{i+1-i}}{N_t^1} \right)$ ;
15:     end for
16:     Construct the filtered pseudo-label set  $\hat{\mathcal{F}}_t = \hat{\mathcal{F}}_t^1 \cup \hat{\mathcal{F}}_t^2 \dots \cup \hat{\mathcal{F}}_t^I$ 
17:     Update the labelled set  $\mathcal{L}_t = \mathcal{L}_{t-1} \cup \hat{\mathcal{F}}_t$ 
18:     Update the unlabelled set  $\mathcal{U}_t = \mathcal{U}_{t-1} - \hat{\mathcal{F}}_t$ 
19:     Update the confidence threshold  $\mu_t^i$ ;
20:     Update the filtering rate  $\psi_t^i$ ;
21: end for
22: return  $w_T$ 
  
```

a part of the I - V characteristic curves under normal and various fault conditions are collected through the actual PV arrays as a reference. The actual PV arrays consists of two parallel-connected PV strings with twenty-two PV modules in series. The data in various situations are collected by the I - V tester (Model: PROVA1011) and an environmental tester with Bluetooth communication function, as shown in Figure 6. The PV module installed in the PV arrays is HT60-156M-C-330, and the detailed parameters are shown in Table 1.

The short-circuit faults, partial shading faults, and degradation faults are simulated respectively on the actual PV arrays, as shown in Figure 7. Specifically, the actual I - V characteristic curves of the above four states are shown in Figure 8.

Then, these samples are used to verify the validity of the following simulation data.

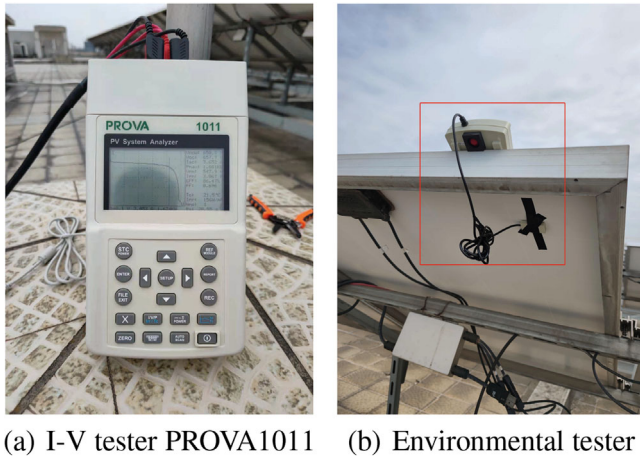


FIGURE 6 The I - V tester PROVA1011 and environmental tester

TABLE 1 Technical data at standard test condition

Module parameter	Value
Maximum power point M_{pp} (W)	99.925
Open circuit voltage V_{oc} (V)	21.5
Short-circuit current I_{sc} (A)	6.03
Maximum power voltage V_{mp} (V)	17.5
Maximum power current I_{mp} (A)	5.71
Fuse current (A)	15
Maximum system voltage (V)	1000

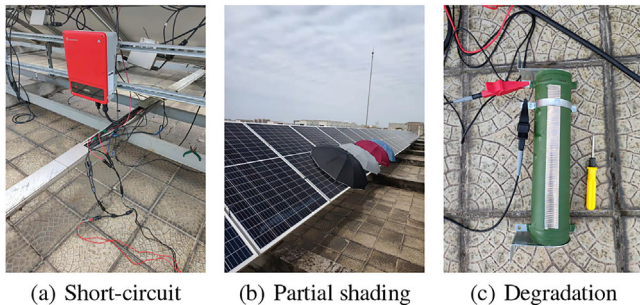


FIGURE 7 Fault simulation in actual PV arrays

5.2 | Numerical data collection

A simulation model of PV module fault is established based on Simulink, as shown in Figure 9. By adjusting the values of the temperature and the irradiation modules, a wider range of samples of normal and three fault states under various weather conditions are collected.

This model consists of three parallel-connected PV strings with six PV modules in series. To save space, the three PV panels in each column contain one, two, and three PV modules, respectively. Specifically, by connecting a resistor R_{short} with a small resistance in parallel with the PV module, the short-circuit

faults are simulated. The partial shading faults are realised by connecting a gain module $Gain_{ps}$ with a value between 0 and 1 in series between the PV panels and the irradiance module. A resistor $R_{degradation}$ is connected in series in the array to simulate the degradation faults. The key parameter settings are shown in Table 3.

By adjusting the values of temperature and irradiation module, the temperature changes from 10 to 70°C with the step length of 2°C. In the meanwhile, the irradiance changes from 50 to 1000 W/m² with the step length of 10 W/m² [47]. Based on the simulation model, the samples in the states of normal, short-circuit, degradation, and partial shading are selected. There are 2976 samples of each state and the total is 11,904.

5.3 | Experiments configuration

In the numerical simulation experiments, we consider the accurate diagnosis of PV stations in different fault situations. To fully prove the effectiveness of the proposed method, six independent experiments are designed. For each experiment, the PV station has a large amount of unlabelled data and a small amount of labelled data, which includes normal and part of the fault states and the number of label samples in different states is quite different. In order to build scenarios of unbalanced label samples, the numerical data is firstly divided into training set and test set at a ratio of 7:3. It is worth noting that each fault type is divided according to this ratio instead of randomly shuffled samples. At this time, different states have the same number of label samples. After that, the different missing ratios are set for the normal and partial fault states respectively, which means that different numbers of labelled samples are randomly discarded in different states in the training set, and these samples are collected as unlabelled data.

The goal of this paper is to utilise these small and unbalanced labelled data and large amounts of unlabelled data to achieve accurate fault diagnosis. The model training rounds T of SSL are set to 20. The missing ratios of each state and the initial sample filtering threshold for SSL in the six independent experiments are listed in Table 2. In each experiment, we set up the PV station to contain different types of fault samples. Taking experiment 1 as an example, it only has normal state samples and short-circuit fault samples. Specifically, these two types have 1872 training samples and 1104 test samples respectively. The sample size of each type in the initial labelled training set is defined as the missing rate multiplied by 1872, and the rest of the training samples are used to construct the unlabelled dataset. The key hyper-parameters settings of CNN are shown in Table 4. The experiments are conducted on HP ZBook Create G7 with Intel (R) Core (TM) i9-10885H CPU, 32 GB RAM, NVIDIA GeForce RTX 2070 Max-Q.

5.4 | Experimental results

To verify the effectiveness and advancement of the proposed FSR-SSL method, the fault diagnosis accuracy of

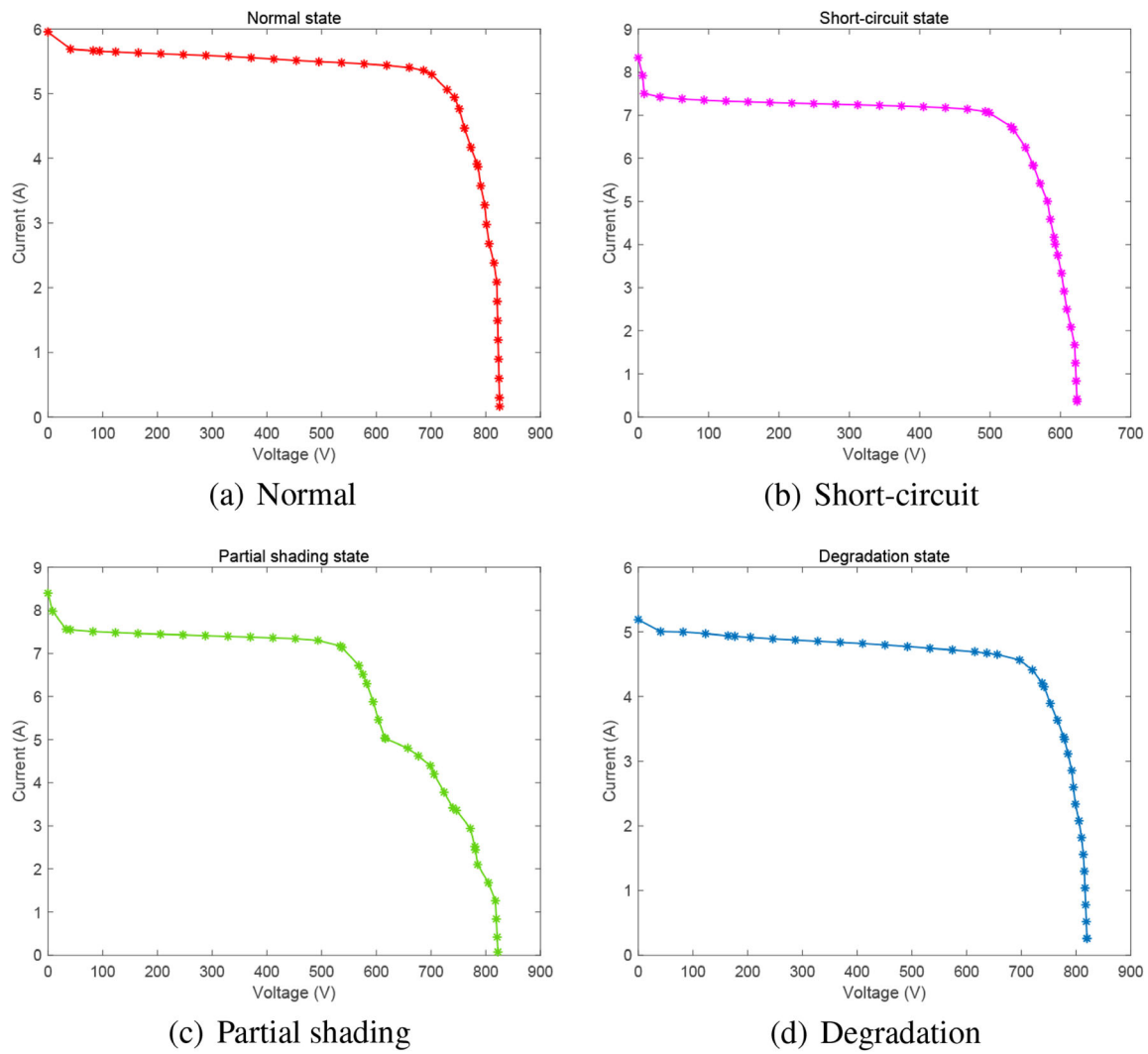


FIGURE 8 The actual I - V characteristic curves under four states

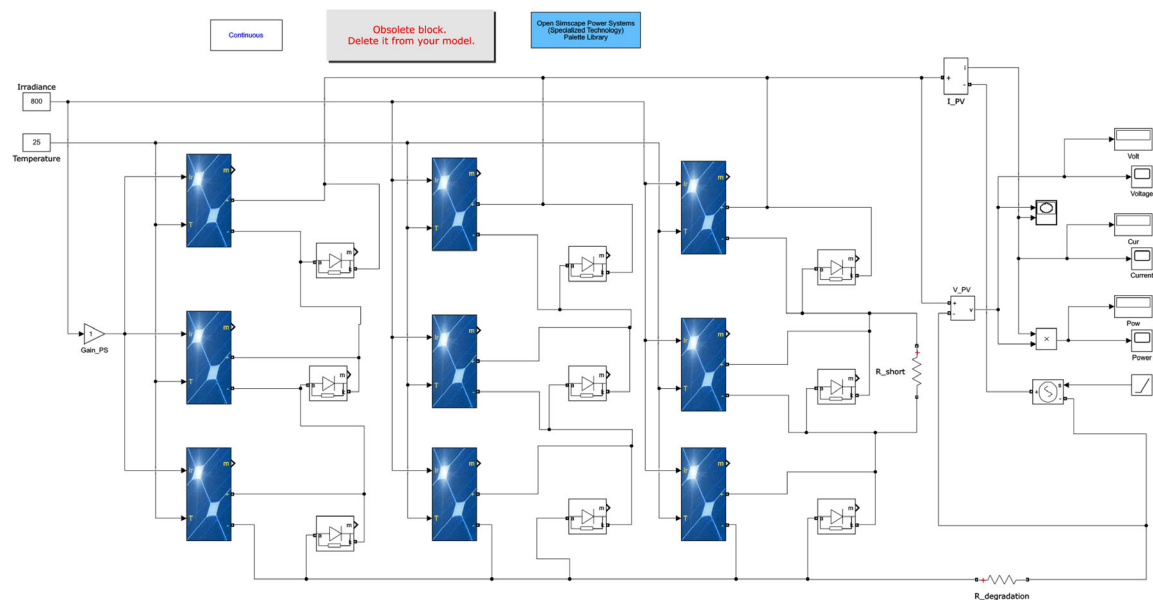


FIGURE 9 The simulation model of PV module faults

TABLE 2 Settings of the missing ratios and the initial sample filtering thresholds for semi-supervised learning

Case	The missing ratio				The filtering threshold			
	Normal	Short-circuit	Degradation	Partial shading	Normal	Short-circuit	Degradation	Partial shading
1	0.1	0.95	\	\	0.99	0	\	\
2	0.1	\	\	0.95	0.99	\	\	0
3	0.1	0.8	0.95	\	0.99	0	0	\
4	0.1	0.8	\	0.95	0.99	0	\	0
5	0.1	\	0.8	0.95	0.99	\	0	0
6	0.1	0.8	0.95	0.7	0.99	0	0	0

Remark: '\' indicates that there is no corresponding fault types in the data of the experiment.

TABLE 3 Settings of the simulation module parameters

Module parameters	Value
Short-circuit resistance R_{short} (ohms)	0.001
Degradation resistance $R_{\text{degradation}}$ (ohms)	3
Partial shading gain Gain_{ps}	0.5
Maximum power point M_{pp} (W)	99.925
Open circuit voltage V_{oc} (V)	21.5
Voltage at maximum power point V_{mp} (V)	17.5
Short-circuit current I_{sc} (A)	6.03
Current at maximum power point I_{mp} (A)	5.71
Light-generated current I_{L} (A)	6.0576
Diode saturation current I_0 (A)	2.0517e-10
Diode ideality factor	0.96445
Shunt resistance R_{sh} (ohms)	551.8793
Series resistance R_{s} (ohms)	0.2392
Cells per module (Ncell)	36
Temperature coefficient of V_{oc} ($\%/^{\circ}\text{C}$)	-0.36
Temperature coefficient of I_{sc} ($\%/^{\circ}\text{C}$)	0.06

TABLE 4 Settings of hyper-parameters

Hyper-parameter	Value	Hyper-parameter	Value
B	128	l_{r}	1e-3
β_1	0.995	β_2	0.999
Epoch	50	ϵ	1e-8

the model is tested under the above six experiments, and it is compared with direct training (DT), conventional semi-supervised learning (SSL), and down-sampling semi-supervised learning (DSSL). The results are shown in Table 5.

It is worth noting that these four methods use the same data set in each experiment to train the fault diagnosis model separately. Among them, DT means that only the labelled samples are used to train the model, not unlabelled data. SSL represents

TABLE 5 The test accuracy of different methods in six experiments

Case/Method	DT	SSL	DSSL	FSR-SSL
1	0.7061	0.5012	0.5004	0.9995
2	0.9928	0.5034	1.0000	1.0000
3	0.6667	0.6667	0.8040	0.9915
4	0.6667	0.6667	0.6667	1.0000
5	0.8288	0.9997	0.8493	1.0000
6	0.8220	0.8836	0.7508	0.9995

the traditional self-learning method, and the pseudo-label filtering threshold is set to 0.9. DSSL means down-sampling the largest number of types in the labelled data, but this may lead to under fitting problem. Obviously, compared with the other three methods, the proposed FSL-SSL method has the highest accuracy rate in all six experiments and reached 99%. Especially in experiments 1, 3, 4 and 6, the accuracy of the proposed FSR-SSL method is increased by 29%, 19%, 33% and 11% respectively compared with the best existing methods. This indicates that the proposed method effectively extracts the fault characteristics from the I - V characteristic curves, and uses the fault sample rebalancing strategy to accurately add pseudo labels to unlabelled samples.

In addition, the accuracy of the proposed method and ref. [23] are compared during each round of the semi-supervised training process in six experiments, where the parameter alpha in reference [23] is equal to 1, and the confidence thresholds are 0, 0.6 and 0.9, respectively. To avoid accidental errors, the result of each experiment is the average of 20 times, as shown in Figure 10. Obviously, the accuracy of our proposed method is significantly better than ref. [23] under various experimental data conditions, which further verifies the effectiveness of the sample rebalancing strategy proposed in this paper.

To further visually analyse the reasons, the differences between the SSL method and the proposed FSR-SSL method in the SSL process are carefully compared. Similarly, the accuracy and pseudo-label addition of each round in the semi-supervised

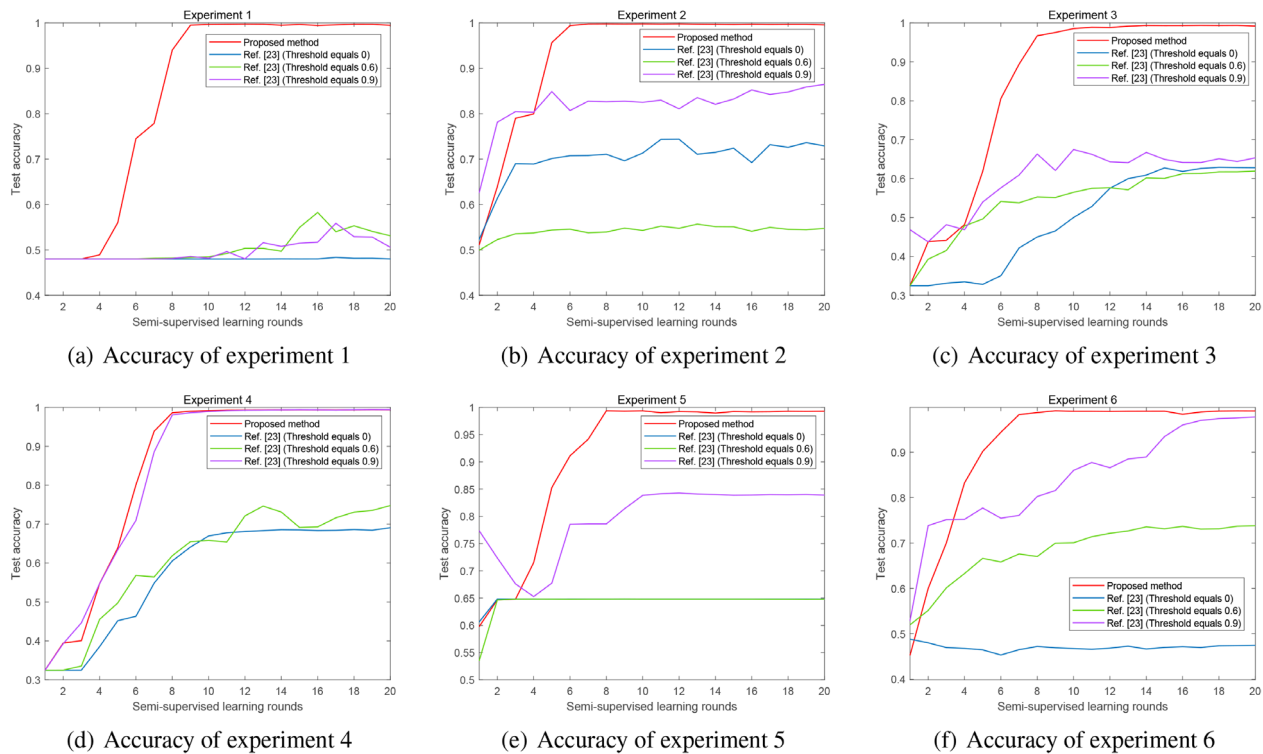


FIGURE 10 Average accuracy of six experiments

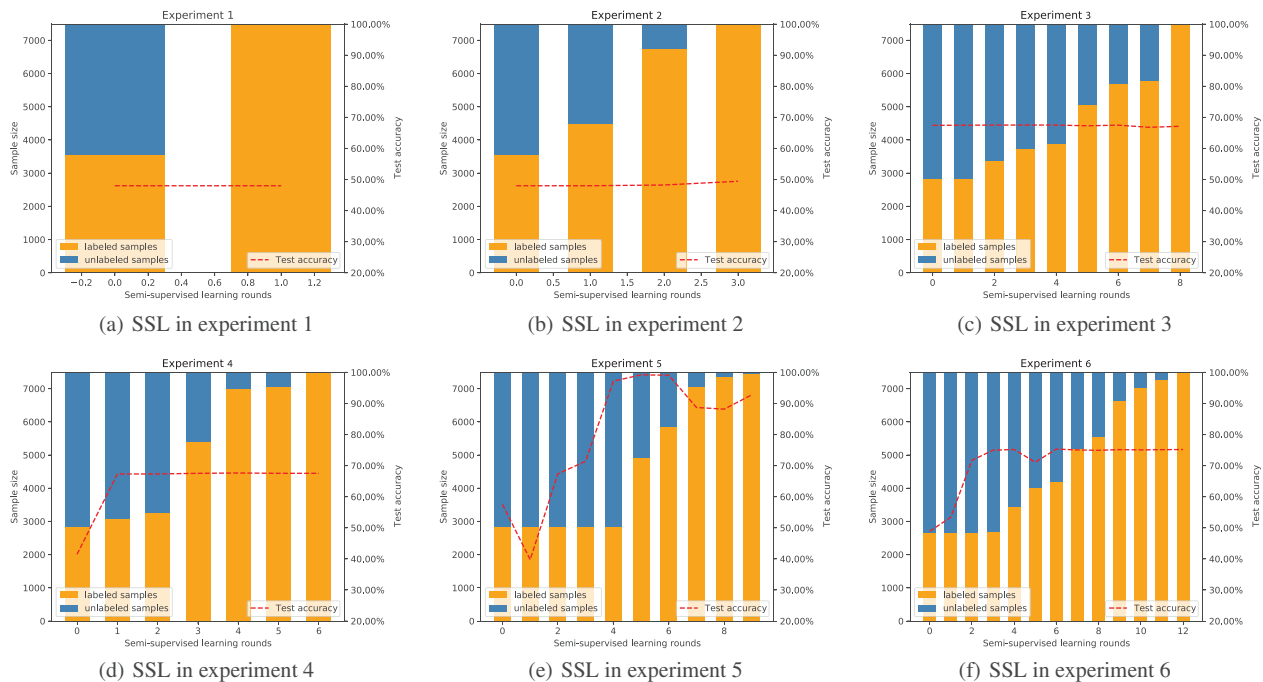


FIGURE 11 SSL training processes in six experiments

training process of SSL method in six experiments are shown in Figure 11.

Obviously, although the utilisation rate of unlabelled data is very high in the six experiments, only experiment

5 has a higher accuracy, and it is very low in other cases. This means that when there are large differences in the number of different types among labelled samples, it is difficult for the traditional semi-supervised method to

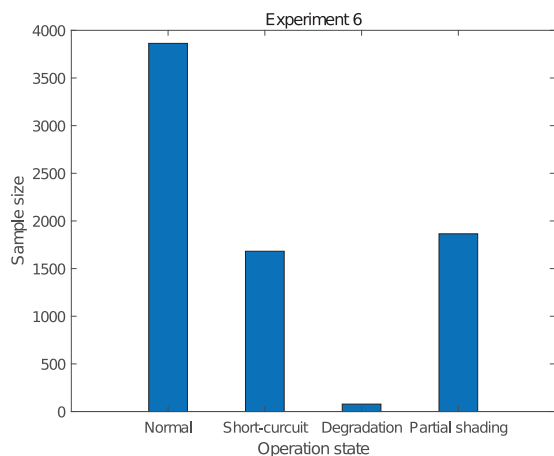


FIGURE 12 The result of SSL pseudo label addition

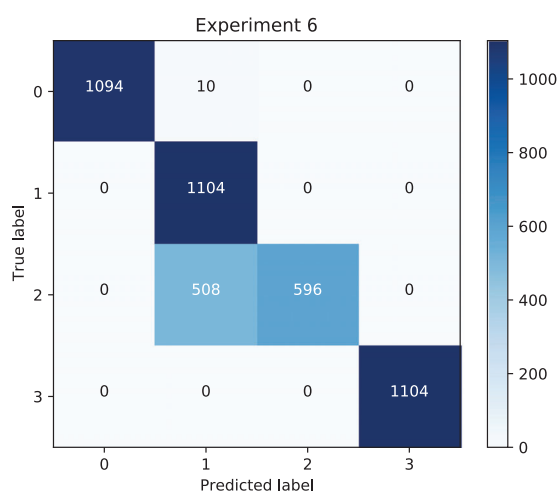


FIGURE 13 The confusion matrix of SSL fault diagnosis

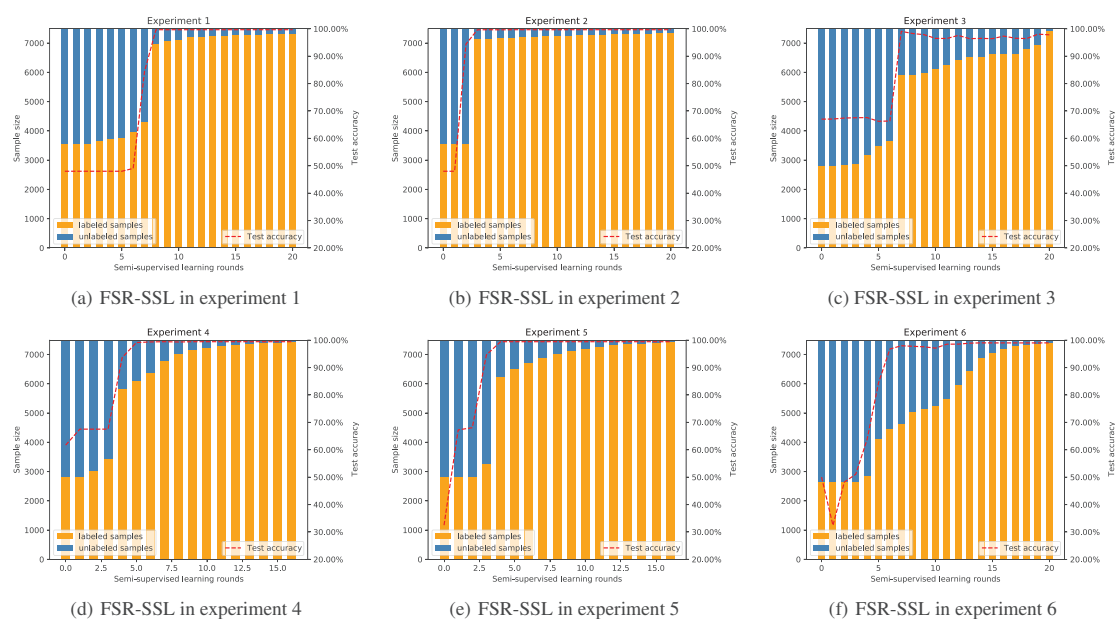


FIGURE 14 FSR-SSL training processes in six experiments

accurately add pseudo labels and cannot achieve accurate fault diagnosis.

Furthermore, the experiment 6 is taken as an example to show the results of pseudo-label addition of the SSL method on the training set and the confusion matrix of the fault diagnosis on the test set, as shown in Figures 12 and 13, respectively. It can be observed that the conventional SSL method will be affected by the imbalance of the initial labelled samples, and the amount of pseudo-label data added for different types varies greatly, as shown in Figure 12.

After SSL, the trained model is difficult to accurately diagnose the fault since the samples of type 3 are significantly less than other types, which can be clearly observed from the confusion matrix in Figure 13.

As a comparison, the accuracy and pseudo-label addition of each round of FSR-SSL in the semi-supervised training process are shown in Figure 14. It can be observed that not only the utilisation of unlabelled data in the six experiments is very high, and all of them can reach the accuracy of 99%. This means that the proposed FSR-SSL method can overcome the problem of unbalanced initial labelled samples, and correctly add pseudo labels to unlabelled data to achieve accurate fault diagnosis. In addition, also take experiment 6 as an example to present the pseudo-label addition result of the FSR-SSL method on the training set and the confusion matrix of the fault diagnosis on the test set, as shown in Figures 15 and 16 respectively. Obviously, the proposed FSR-SSL method uses the fault sample rebalancing strategy to reduce the difference of each type after adding pseudo-label data, as shown in Figure 15. It overcomes the impact of the imbalance of the initial labelled samples and achieve accurate fault diagnosis, as shown in Figure 16. In the meanwhile, the number of self-training rounds,

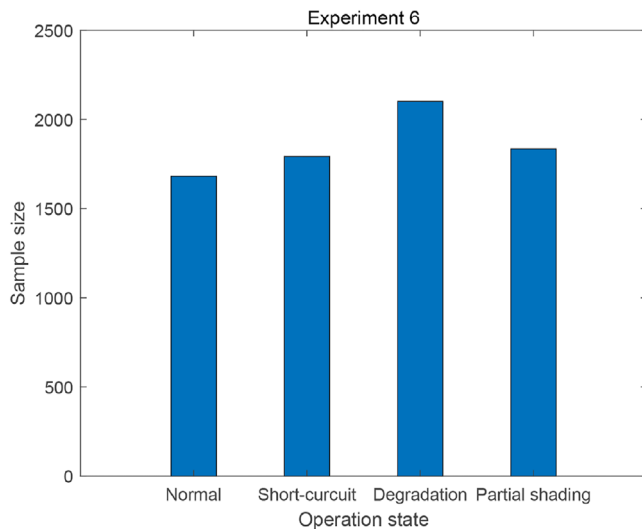


FIGURE 15 The result of FSR-SSL pseudo label addition

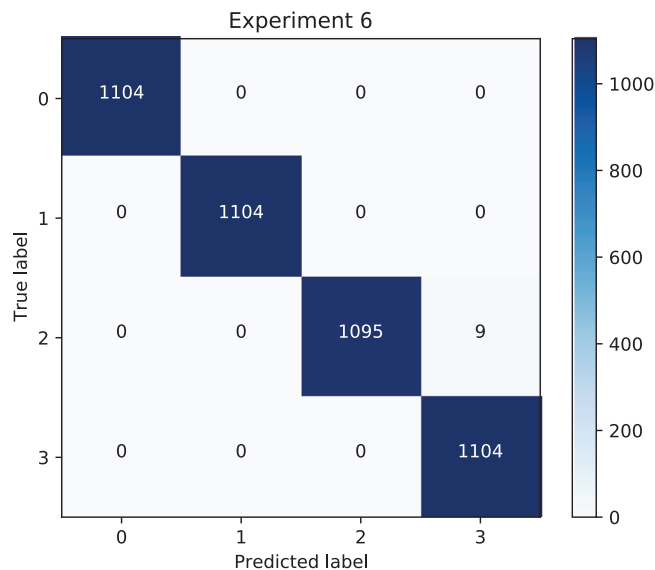


FIGURE 16 The confusion matrix of FSR-SSL fault diagnosis

TABLE 6 The model convergence speed

Result/Case	1	2	3	4	5	6
Round	9	6	10	8	8	7
Time (s)	45.4	39.1	72.7	54.2	51.2	49.5

training and testing time required for the proposed FSR-SSL method to converge under the six experiments in Table 2 are also tested. To avoid accidental errors, the result of each experiment is the average of 20 times, as shown in Table 6.

The above numerical experimental results show that the proposed FSR-SSL method is applicable to various meteorologi-

cal conditions. Through fault sample rebalancing strategy, the problem of unbalanced initial labelled samples is well solved and accurate fault diagnosis is realised.

6 | CONCLUSION

In this paper, a new fault sample rebalancing framework based on semi-supervised learning (FSR-SSL) is proposed for PV fault diagnosis, where three fault types are considered including the degradation, short circuit, and partial shading. This paper studies more practical and troublesome scenarios, where the PV station has only a few labelled samples and the sample size of various types is unbalanced. These challenges affect the extraction of fault features and reduce the accuracy of existing SLL-based methods, especially for degradation faults. To effectively select pseudo-label samples with high accuracy, a dual-threshold selection mechanism is proposed to set different confidence thresholds for various types. Furthermore, a fault sample rebalancing strategy is designed to flexibly add the obtained trusted pseudo-label samples to the training set based on the proportion of different types of labelled data. Thus the model learning bias caused by type imbalance is well overcome. The extensive numerical experiments show that the proposed FSR-SSL method reaches 99% accuracy. Compared with existing methods, the accuracy is increased by up to 33%.

ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China (Grant No. 2018YFB1702300), and in part by the NSF of China (Grants No. 61731012, 62103265, and 92167205).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Xinyi Wang  <https://orcid.org/0000-0002-5351-3446>

Bo Yang  <https://orcid.org/0000-0001-9268-8436>

REFERENCES

- Papadis, E., Tsatsaronis, G.: Challenges in the decarbonization of the energy sector. *Energy* 205, 118025 (2020)
- Zhu, D., Yang, B., Liu, Q., Ma, K., Zhu, S., Ma, C., et al.: Energy trading in microgrids for synergies among electricity, hydrogen and heat networks. *Appl. Energy* 272, 115225 (2020)
- DESA, U.: *Energy statistics yearbook 2014*. New York: United Nations Reproduction Section, New York (2017)
- Wang, Y., Qiu, J., Tao, Y., Zhao, J.: Carbon-oriented operational planning in coupled electricity and emission trading markets. *IEEE Trans. Power Syst.* 35(4), 3145–3157 (2020)
- Saranchimeg, S., Nair, N.K.C.: A novel framework for integration analysis of large-scale photovoltaic plants into weak grids. *Appl. Energy* 282, 116141 (2021)
- Mellit, A., Tina, G.M., Kalogirou, S.A.: Fault detection and diagnosis methods for photovoltaic systems: A review. *Renewable Sustainable Energy Rev.* 91, 1–17 (2018)

7. Takashima, T., Yamaguchi, J., Otani, K., Oozeki, T., Kato, K., Ishida, M.: Experimental studies of fault location in PV module strings. *Sol. Energy Mater. Sol. Cells* 93(6–7), 1079–1082 (2009)
8. Saleh, M.U., Deline, C., Benoit, E.J., Kingston, S.R., Harley, J.B., Furse, C.M., et al.: Detection and localization of damaged photovoltaic cells and modules using spread spectrum time domain reflectometry. *IEEE J. Photovoltaics* 11(1), 195–201 (2020)
9. Davarifar, M., Rabhi, A., El-Hajjaji, A., Dahmane, M.: Real-time model base fault diagnosis of PV panels using statistical signal processing. In: 2013 International Conference on Renewable Energy Research and Applications (ICRERA), pp. 599–604. IEEE, Piscataway, NJ (2013)
10. Harrou, F., Sun, Y., Taghezouit, B., Saidi, A., Hamlati, M.E.: Reliable fault detection and diagnosis of photovoltaic systems based on statistical monitoring approaches. *Renewable Energy* 116, 22–37 (2018)
11. Zhao, Y., Ball, R., Mosesian, J., de Palma, J.F., Lehman, B.: Graph-based semi-supervised learning for fault detection and classification in solar photovoltaic arrays. *IEEE Trans. Power Electron.* 30(5), 2848–2858 (2014)
12. Dhar, S., Patnaik, R.K., Dash, P.: Fault detection and location of photovoltaic based dc microgrid using differential protection strategy. *IEEE Trans. Smart Grid* 9(5), 4303–4312 (2017)
13. Kaplanis, S., Kaplani, E.: Energy performance and degradation over 20 years performance of BP C-Si pv modules. *Simul. Modell. Pract. Theory* 19(4), 1201–1211 (2011)
14. D'Aliento, S., Di Napoli, F., Guerriero, P., d'Alessandro, V.: A modified bypass circuit for improved hot spot reliability of solar panels subject to partial shading. *Solar Energy* 134, 211–218 (2016)
15. Silvestre, S., da Silva, M.A., Chouder, A., Guasch, D., Karatepe, E.: New procedure for fault detection in grid connected PV systems based on the evaluation of current and voltage indicators. *Energy Convers. Manage.* 86, 241–249 (2014)
16. Huang, Z., Wang, Z., Zhang, H.: Multiple open-circuit fault diagnosis based on multistate data processing and subsection fluctuation analysis for photovoltaic inverter. *IEEE Trans. Instrum. Meas.* 67(3), 516–526 (2018)
17. Chine, W., Mellit, A., Lughi, V., Malek, A., Sulligoi, G., Pavan, A.M.: A novel fault diagnosis technique for photovoltaic systems based on artificial neural networks. *Renewable Energy* 90, 501–512 (2016)
18. Chen, Z., Wu, L., Cheng, S., Lin, P., Wu, Y., Lin, W.: Intelligent fault diagnosis of photovoltaic arrays based on optimized kernel extreme learning machine and iv characteristics. *Appl. Energy* 204, 912–931 (2017)
19. Chao, K.H., Chen, P.Y., Wang, M.H., Chen, C.T.: An intelligent fault detection method of a photovoltaic module array using wireless sensor networks. *Int. J. Distrib. Sens. Netw.* 10(5), 540147 (2014)
20. Mohamed, A., Nassar, A.: New algorithm for fault diagnosis of photovoltaic energy systems. *Int. J. Comput. Appl.* 114(9), 26–31 (2015)
21. Razavi.Far, R., Hallaji, E., Farajzadeh.Zanjani, M., Saif, M.: A semi-supervised diagnostic framework based on the surface estimation of faulty distributions. *IEEE Trans. Ind. Inf.* 15(3), 1277–1286 (2018)
22. Razavi.Far, R., Hallaji, E., Farajzadeh.Zanjani, M., Saif, M., Kia, S.H., Henao, H., et al.: Information fusion and semi-supervised deep learning scheme for diagnosing gear faults in induction machine systems. *IEEE Trans. Ind. Electron.* 66(8), 6331–6342 (2018)
23. Wei, C., Sohn, K., Mellina, C., Yuille, A., Yang, F.: Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10852–10861. IEEE, Piscataway, NJ (2021)
24. Garoudja, E., Harrou, F., Sun, Y., Kara, K., Chouder, A., Silvestre, S.: Statistical fault detection in photovoltaic systems. *Solar Energy* 150, 485–499 (2014)
25. Stauffer, Y., Ferrario, D., Onillon, E., Hutter, A.: Power monitoring based photovoltaic installation fault detection. In: 2015 International Conference on Renewable Energy Research and Applications (ICRERA), pp. 199–202. IEEE, Piscataway, NJ (2015)
26. Chen, L., Wang, X.: Adaptive fault localization in photovoltaic systems. *IEEE Trans. Smart Grid* 9(6), 6752–6763 (2017)
27. Chine, W., Mellit, A., Lughi, V., Malek, A., Sulligoi, G., Massi Pavan, A.: A novel fault diagnosis technique for photovoltaic systems based on artificial neural networks. *Renewable Energy* 90, 501–512 (2016)
28. Dhimish, M., Holmes, V., Mehrdadi, B., Dales, M., Mather, P.: Photovoltaic fault detection algorithm based on theoretical curves modelling and fuzzy classification system. *Energy* 140, 276–290 (2017)
29. Yi, Z., Etemadi, A.H.: Line-to-line fault detection for photovoltaic arrays based on multiresolution signal decomposition and two-stage support vector machine. *IEEE Trans. Ind. Electron.* 64(11), 8546–8556 (2017)
30. Eskandari, A., Milimonfared, J., Aghaei, M.: Fault detection and classification for photovoltaic systems based on hierarchical classification and machine learning technique. *IEEE Trans. Ind. Electron.* 68(12), 12750–12759 (2021)
31. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves imagenet classification. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10684–10695. IEEE, Piscataway, NJ (2020)
32. Huang, J.M., Wai, R.J., Yang, G.J.: 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.* 35(7), 7086–7099 (2020)
33. Yan, K., Zhong, C., Ji, Z., Huang, J.: Semi-supervised learning for early detection and diagnosis of various air handling unit faults. *Energy Build.* 181, 75–83 (2018)
34. Razavi.Far, R., Hallaji, E., Farajzadeh.Zanjani, M., Saif, M., Kia, S.H., Henao, H., et al.: Information fusion and semi-supervised deep learning scheme for diagnosing gear faults in induction machine systems. *IEEE Trans. Ind. Electron.* 66(8), 6331–6342 (2019)
35. Yu, K., Ma, H., Lin, T.R., Li, X.: A consistency regularization based semi-supervised learning approach for intelligent fault diagnosis of rolling bearing. *Measurement* 165, 107987 (2020)
36. Yu, K., Lin, T.R., Ma, H., Li, X., Li, X.: A multi-stage semi-supervised learning approach for intelligent fault diagnosis of rolling bearing using data augmentation and metric learning. *Mech. Syst. Signal Process.* 146, 107043 (2021)
37. Jian, C., Yang, K., Ao, Y.: Industrial fault diagnosis based on active learning and semi-supervised learning using small training set. *Eng. Appl. Artif. Intell.* 104, 104365 (2021)
38. Nie, X., Xie, G.: A two-stage semi-supervised learning framework for fault diagnosis of rotating machinery. *IEEE Trans. Instrum. Meas.* 70, 1–12 (2021)
39. Zhao, Y., Ball, R., Mosesian, J., de Palma, J.F., Lehman, B.: Graph-based semi-supervised learning for fault detection and classification in solar photovoltaic arrays. *IEEE Trans. Power Electron.* 30(5), 2848–2858 (2015)
40. Kim, J., Hur, Y., Park, S., Yang, E., Hwang, S.J., Shin, J.: Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning. In: *Advances in Neural Information Processing Systems*. vol. 33. pp. 14567–14579. Curran Associates, Red Hook, NY (2020)
41. Yang, Y., Xu, Z.: Rethinking the value of labels for improving class-imbalanced learning. In: *Advances in Neural Information Processing Systems*. vol. 33. pp. 19290–19301. Curran Associates, Red Hook, NY (2020)
42. Zhang, T., Zhu, T., Li, J., Han, M., Zhou, W., Yu, P.: Fairness in semi-supervised learning: Unlabeled data help to reduce discrimination. *IEEE Trans. Knowl. Data Eng.* 34(4), 1763–1774 (2020)
43. Alam, M.K., Khan, F., Johnson, J., Flicker, J.: A comprehensive review of catastrophic faults in PV arrays: types, detection, and mitigation techniques. *IEEE J. Photovoltaics* 5, 982–997 (2015)
44. Kumar, M., Kumar, A.: Experimental validation of performance and degradation study of canal-top photovoltaic system. *Appl. Energy* 243, 102–118 (2019)
45. Belhaouas, N., Cheikh, M.S.A., Agathoklis, P., Oularbi, M.R., Amrouche, B., Sedraoui, K., et al.: PV array power output maximization under partial shading using new shifted PV array arrangements. *Appl. Energy* 187, 326–337 (2017)
46. Miceli, R., Orioli, A., Di Gangi, A.: A procedure to calculate the I–V characteristics of thin-film photovoltaic modules using an explicit rational form. *Appl. Energy* 155, 613–628 (2015)

47. Chen, Z., Chen, Y., Wu, L., Cheng, S., Lin, P.: Deep residual network based fault detection and diagnosis of photovoltaic arrays using current-voltage curves and ambient conditions. *Energy Convers. Manage.* 198, 111793 (2019)
48. Zou, Y., Yu, Z., Liu, X., Kumar, B.V.K.V., Wang, J.: Confidence regularized self-training. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 5981–5990. IEEE, Piscataway, NJ (2019)
49. Mukherjee, S., Awadallah, A.H.: Uncertainty-aware self-training for text classification with few labels. *arXiv:2006.15315* (2020)

How to cite this article: Liu, Q., Wang, X., Yang, B., Wang, Z., Liu, Y., Guan, X.: FSR-SSL: A fault sample rebalancing framework based on semi-supervised learning for PV fault diagnosis. *IET Renew. Power Gener.* 16, 2667–2681 (2022).

<https://doi.org/10.1049/rpg2.12458>