ENERGY CATALYST
Volume 2 (2026) · ISSN 3103-9952

TECHNICAL ARTICLE

Development of an Integrated DBN-ELM, CNN-SVM and CNN-BiGRU Photovoltaic Array Fault Diagnosis Model Based on Weighted Probability Averaging

Chunmei Guo[1], Weijin Sun[1], Yang Li[1], Yuwen You[1], Zhonglu He[1*]

[1] School of Energy and Safety Engineering, Tianjin Chengjian University, Tianjin 300384, China

RECEIVED 19 March 2026 · REVISED 30 May 2026 · ACCEPTED 14 July 2026 · PUBLISHED 2 August 2026

ABSTRACT

Photovoltaic arrays are continuously exposed to complex environmental conditions over long periods, making them susceptible to seven types of single and multiple failures such as shadow blocking and module aging. Fault diagnosis of photovoltaic arrays is essential to prevent failures that may lead to reduced power generation efficiency and potential safety hazards. This paper proposes a photovoltaic array fault diagnosis model based on weighted probability averaging, integrating DBN-ELM, CNN-SVM, and CNN-BiGRU methods. The model is calculated and experimentally verified. The results demonstrate that the integrated model achieves an overall accuracy, recall, precision, and F1-score of 99.0% across the four evaluation metrics, indicating a highly effective fault recognition capability.

KEYWORDS PV array, Fault diagnosis methods, Weighted probability average integration model

1. Introduction

Due to prolonged exposure to natural environmental conditions, photovoltaic (PV) arrays are prone to various degradation mechanisms, including module aging, open-circuit faults between strings, localized shading of modules, inter-module short circuits, and simultaneous multi-failure occurrences. These issues can lead to reduced power generation efficiency and potential safety risks. Consequently, extensive research has been conducted on photovoltaic fault diagnosis. Artificial intelligence algorithms have emerged as effective tool in this field due to their high computational speed and superior diagnostic accuracy.

In recent years, artificial intelligence technology has gained widespread popularity among industry professionals due to its efficiency, accuracy, and strong adaptability, and has been extensively adopted across various sectors. The process begins with data collection, where sensors are used to gather various operational parameters from photovoltaic systems, which are then preprocessed. Next, an appropriate AI diagnostic model is developed and trained using the segmented dataset, followed by optimization. Finally, the model's performance is evaluated.

Zain et al. (2023) proposed a method using multi-output deep learning algorithms to detect, classify, and locate short-circuit, ground-fault, and open-circuit faults. Compared with existing methods, this approach significantly reduces the number of sensors required per string—by up to 50%—while maintaining a high diagnostic accuracy rate. Lin et al. (2024) proposed a photovoltaic array open-circuit fault diagnosis model based on a one-dimensional VoVNet-SVDD (Variety of View Network with Support Vector Data Description). The model automatically extracts fault features from the input raw current-voltage curve data, then combines these extracted features with environmental parameters to construct an SVDD model, thereby enabling diagnosis. Sridhar Patthi et al. (2024) proposed a multi-layer neural network (MLNN) technique capable of identifying issues of any size or mismatch severity in photovoltaic arrays. Gao et al. (2024) proposed a PV array fault diagnosis

* Corresponding author. Email address: hzlfirst@163.com (Zhonglu He)

doi.org/10.65582/ec.2026.006

method based on Variational Mode Decomposition (VMD) and the Subtractive Average Optimizer (SABO) to optimize the Kernel Extreme Learning Machine (KELM). By utilizing VMD to extract nine feature vectors from fault data and employing the SABO algorithm to optimize the kernel and regularization parameters of the KELM, they established a PV array fault diagnosis model based on VMD-SABO-KELM. Zhong et al. (2024) proposed a lightweight PV array fault diagnosis model based on an improved SqueezeNet. This model significantly reduces computational complexity by replacing traditional convolutions with deep convolutions, while incorporating residual connections to facilitate effective information flow between layers, thereby enhancing diagnostic efficiency and accuracy. Ali et al. (2025) developed a simple and accurate one-dimensional convolutional neural network model that classifies faults based on selected features. Lu et al. (2023) proposed an accurate FDD model for photovoltaic arrays based solely on small datasets, utilizing a Wasserstein-based generative adversarial network to generate more labeled samples to improve the performance of the CNN-based classifier. Liu et al. (2021) proposed a PV array fault diagnosis method based on stacked autoencoders and clustering algorithms, which can automatically extract features and mine data sample characteristics for fault diagnosis using a small number of labeled data samples. Fu et al. (2025), to address the issue of data imbalance in photovoltaic arrays affecting the performance of fault diagnosis models, proposed integrating an improved slime mold algorithm with a CatBoost model based on a polyloss function to achieve accurate composite fault diagnosis. Zahra Yahyaoui et al. (2024) proposed a novel Bayesian optimization-based diagnostic framework using interval-gated recurrent units to reduce the computational and storage costs associated with sensor uncertainty. Shiue-Der Lu et al. (2024) were able to rapidly and accurately measure and locate short-circuit and open-circuit faults in bypass diodes by integrating voltmeters and support vector machines into solar photovoltaic array modules. Mansour Hajji et al. (2021) developed a fault diagnosis method based on principal component analysis (PCA) to extract and select the most relevant multivariate features, and applied supervised machine learning classifiers to address the challenge of selecting more relevant and sensitive features.

The NCA-CNN model proposed by Umit Cigdem et al. (2025) achieves a fault recognition rate of up to 99% for both aging and series resistance faults. The 1D-CNN model proposed by Belqasem (Aljafari *et al.* 2024) achieves a 98.15% accuracy in identifying shadow faults in open- circuit conditions. Aziz, F et al. (2020) proposed a fault diagnosis method based on a 2D-CNN. The results show that the proposed fine-tuned pre-trained CNN model achieves a fault detection accuracy of 73.53% and 70.45% in noisy scenes, outperforming existing machine learning and deep learning methods.

In summary, four typical types of AI-detected photovoltaic (PV) faults are identified for single fault scenarios, including module surface shading, aging, open circuits, and short circuits. The diagnostic accuracy ranges from 93% to 99%. However, although the aforementioned models demonstrate relatively high accuracy in diagnosing several types of single faults, the diagnosis of multiple failures still requires further improvement.

In this paper, an ensemble learning model is proposed based on weighted probabilistic averaging. This integrated model effectively combines the strengths of DBN-ELM, CNN-SVM, and CNN-BiGRU. Seven types of faults, encompassing both single and multiple faults, were identified. Corresponding training and testing datasets were then constructed, and the performance of the model was computed and evaluated.

2. Construction of PV array fault dataset

As the foundation of fault diagnosis research, the construction and application of photovoltaic system fault datasets have a decisive impact on the development and optimization of diagnostic models. Voutsinas et al. (2022) used a single-diode model to calculate the photocurrent generated by given irradiance and temperature settings, creating values for a dataset to be used in the neural network training process. Karmacharya et al.(2018) modeled and simulated an ungrounded photovoltaic system in a real-time digital simulator (RTDS). Signals from the model outputs were collected to construct the dataset. Wang et al. (2019) set up a fault data acquisition system and, based on this, used MATLAB to establish a 4×3 PV array model. They simulated the PV array's normal, open-circuit, short-circuit, inter-string, and multi-fault states, collecting 779 data samples across these five states. Based on the fundamental principles of the various components of off-grid photovoltaic systems, fault simulations are performed using a simulation model in this chapter, and fault labels are established and a fault dataset is constructed from the results of the simulation analysis.

To conduct an in-depth analysis of the output performance of a photovoltaic array under various fault conditions, this paper uses MATLAB/Simulink to build a 3×3 photovoltaic array model with a total output power of 2 kW. The study examines the output characteristics of four single faults—shading on the module surface, short circuits within module strings, open circuits between strings, and array aging—as well as three types of concurrent faults. Figure 1 shows a schematic diagram of a photovoltaic array. The detailed parameters of each photovoltaic module under standard test conditions are shown in Table 1. The schematic illustrates the behavior of several types of single faults.

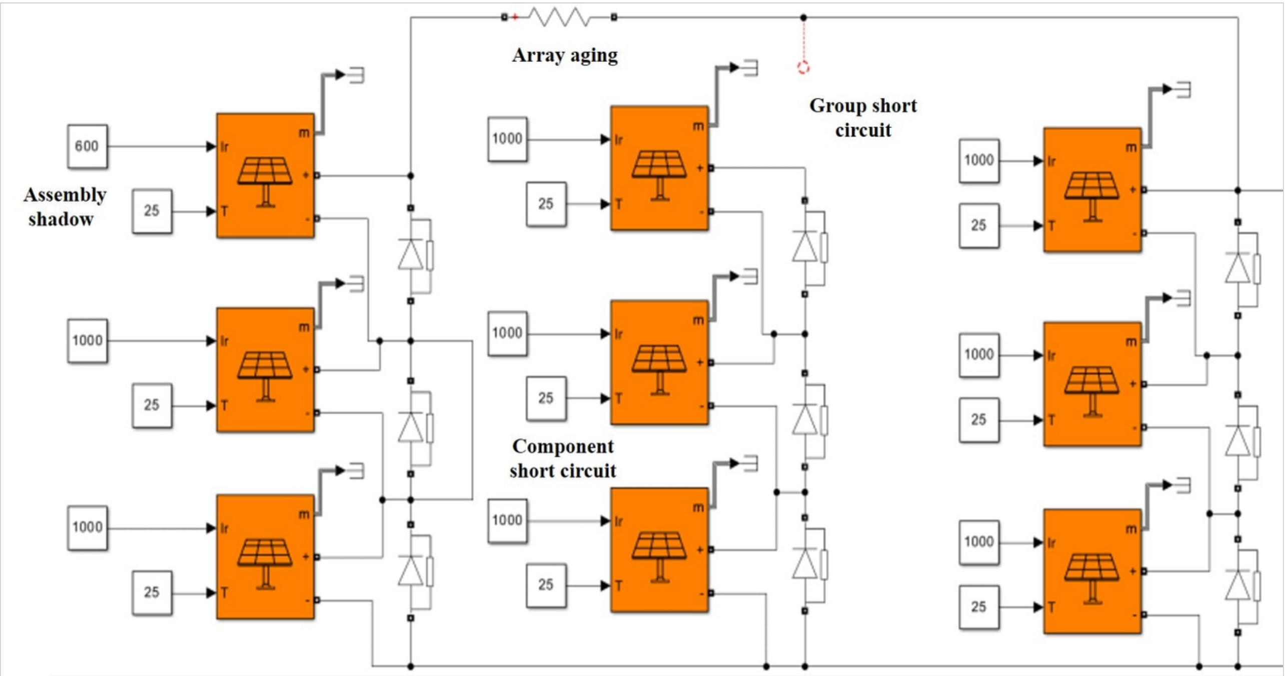

Figure 1. Simulation modelling of PV array failures.

Table 1. Parameters under Standard Test Conditions for PV modules

Parameters	Value
Maximum power P_{mppt}/W	218.87
Maximum - power voltage V_{mppt}/V	29.30
Maximum - power current I_{mppt}/A	7.47
Open circuit voltage V_{oc}/V	36.60
Short circuit current I_{sc}/A	7.97

By establishing a simulation model of the output characteristics of photovoltaic modules, multi-condition simulations were conducted by varying irradiance (200–1000 W/m^2) and temperature (15–55 °C) to reveal the dynamic relationship between the modules' electrical parameters and environmental factors. Simulations show that when the module temperature is held constant at 25 °C, a decrease in irradiance from 1000 W/m^2 to 200 W/m^2 causes the short-circuit current (Isc) to decrease approximately linearly from 41 A to 8 A (a reduction of 80.5%), resulting in a drop in maximum power point from 3700 W to 600 W; The open-circuit voltage (Voc) decreased from 110 V to 103 V, a reduction of approximately 6.3%. As shown in the figure 2 to figure 5.

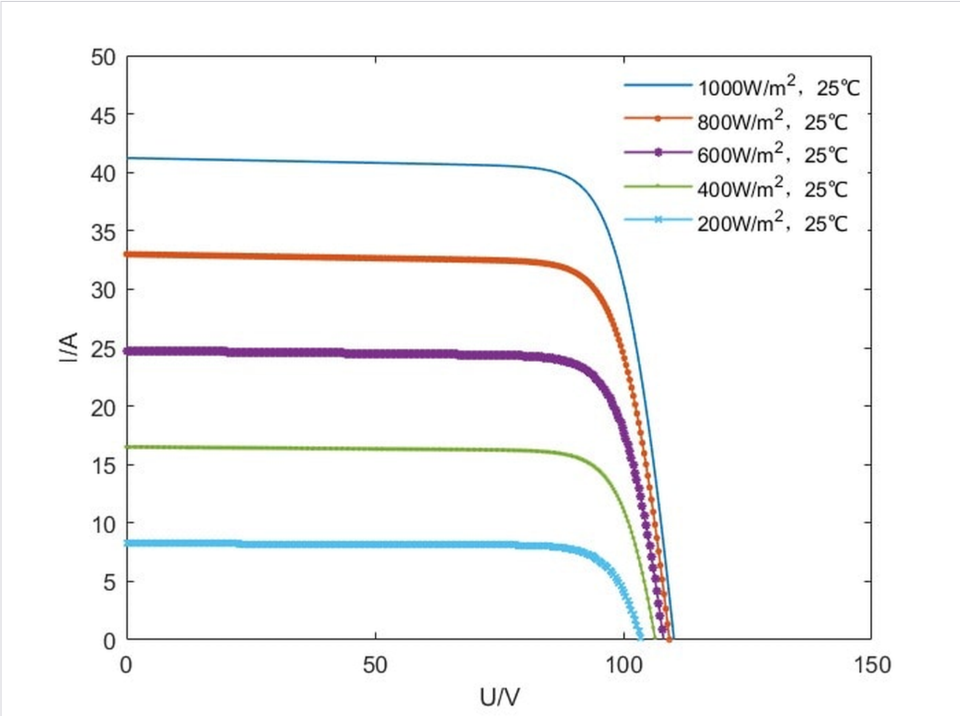

Figure 2. I-V Curves of a Photovoltaic Array as a Function of Irradiance.

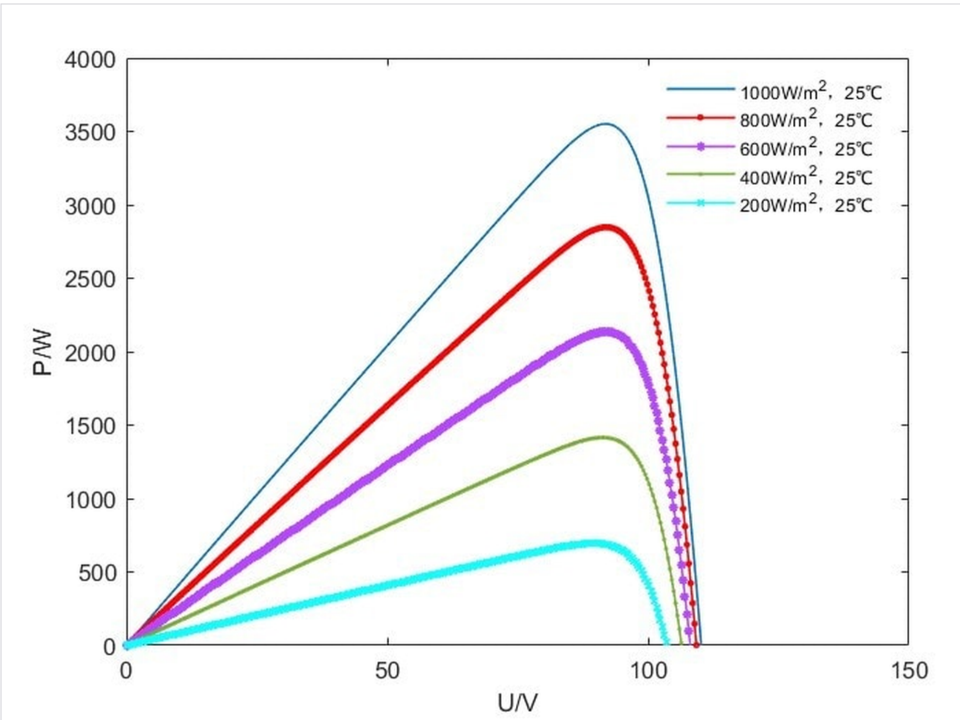

Figure 3. P-V Curves of a Photovoltaic Array as a Function of Irradiance.

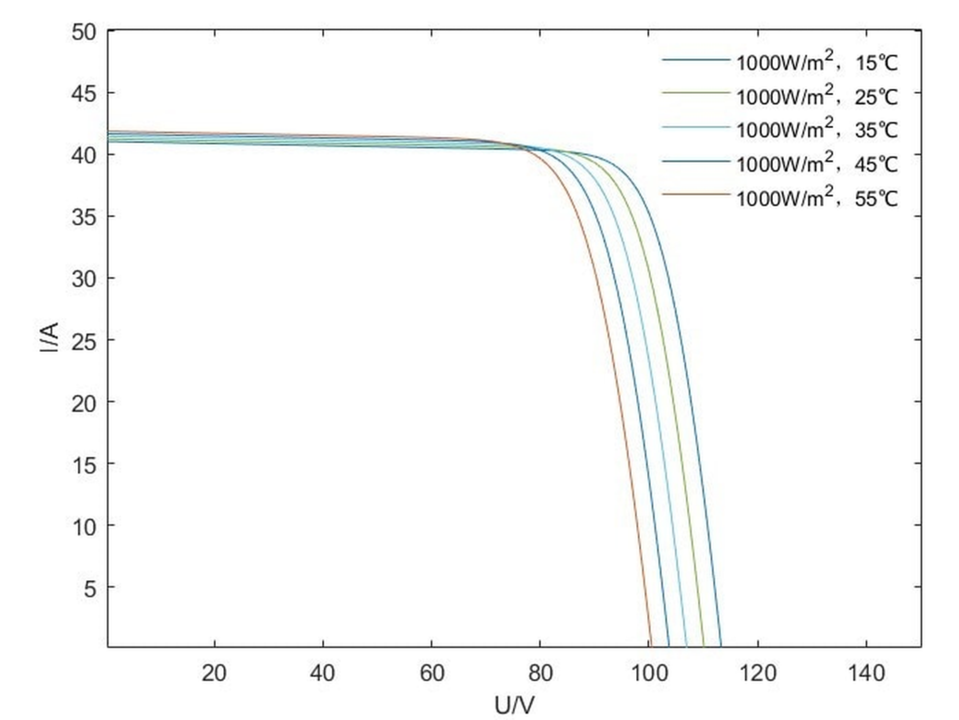

Figure 4. I-V Curves of a Photovoltaic Array as Module Temperature Varies.

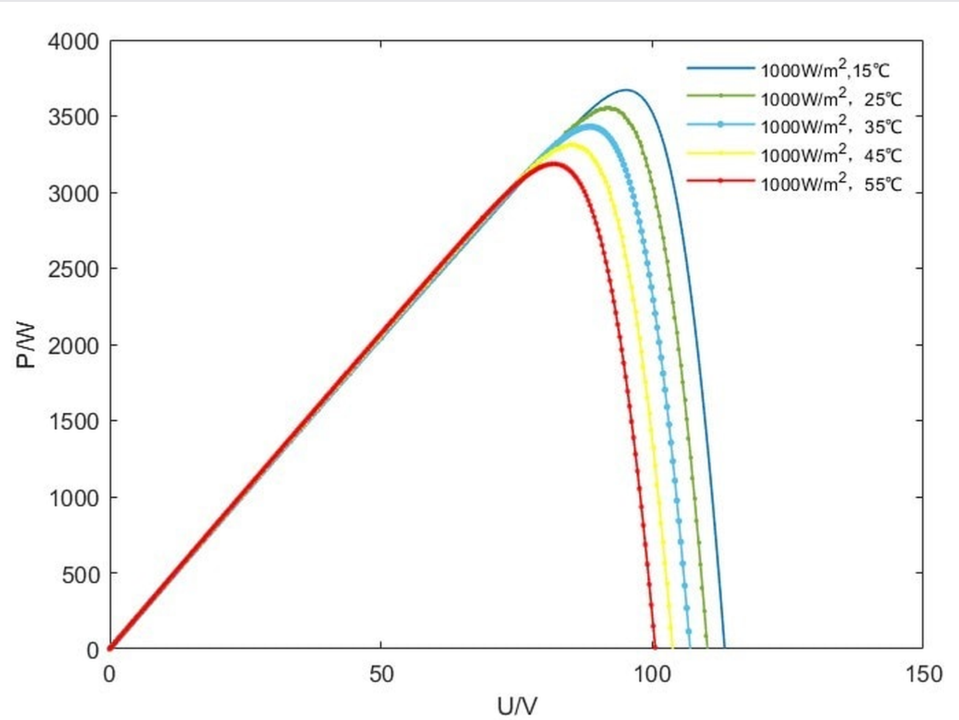

Figure 5. I-V Curves of a Photovoltaic Array as Module Temperature Varies.

Under conditions of constant irradiance (1000 W/m^2), simulation data shows that as the module temperature rises from 15°C to 55°C, the open-circuit voltage (Voc) exhibits a significant negative temperature coefficient, decreasing from 115 V to 101 V at a rate of -0.35%/°C, with a total reduction of 12.2%. In contrast, the short-circuit current (Isc) shows only slight changes. The combined effect of these factors causes the maximum power point to decrease from 3700 W to 3100 W at a rate of -0.41%/°C, representing a reduction of approximately 16.2%.

These regular patterns provide important evidence for fault diagnosis in photovoltaic arrays. Under normal operating conditions, changes in output parameters should strictly follow the aforementioned radiation-temperature coupling laws, and any deviation from these patterns may indicate a potential fault.

Simulations were conducted to obtain the I-V and P-V characteristic curves for a single photovoltaic module under surface shading conditions. These simulations were performed under STC (Standard Test Conditions) with an irradiance of 1000 W/m^2 and a module temperature of 25 °C. Analysis shows that when the module is shaded, the array output current decreases in a “step-like” manner with voltage, and the output power first increases, then decreases, and then increases again, exhibiting multiple peaks; the open-circuit voltage and short-circuit current do not change significantly, while the maximum output power decreases by approximately 1000 W as shown in the figure 6 and 7.

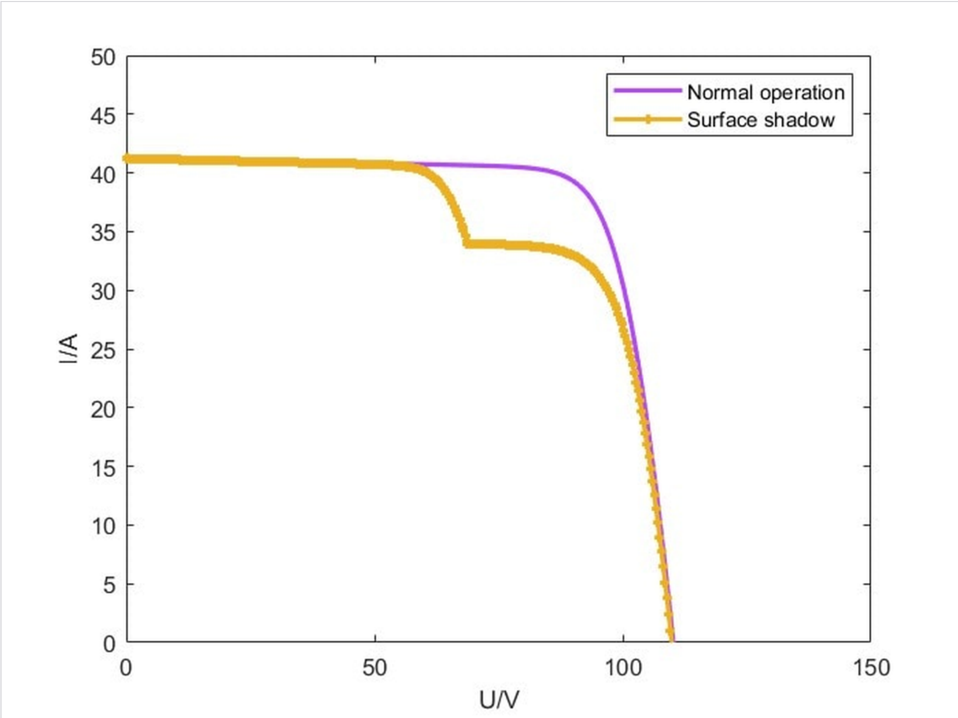

Figure 6. I-V curve when shadows fall on the module surface.

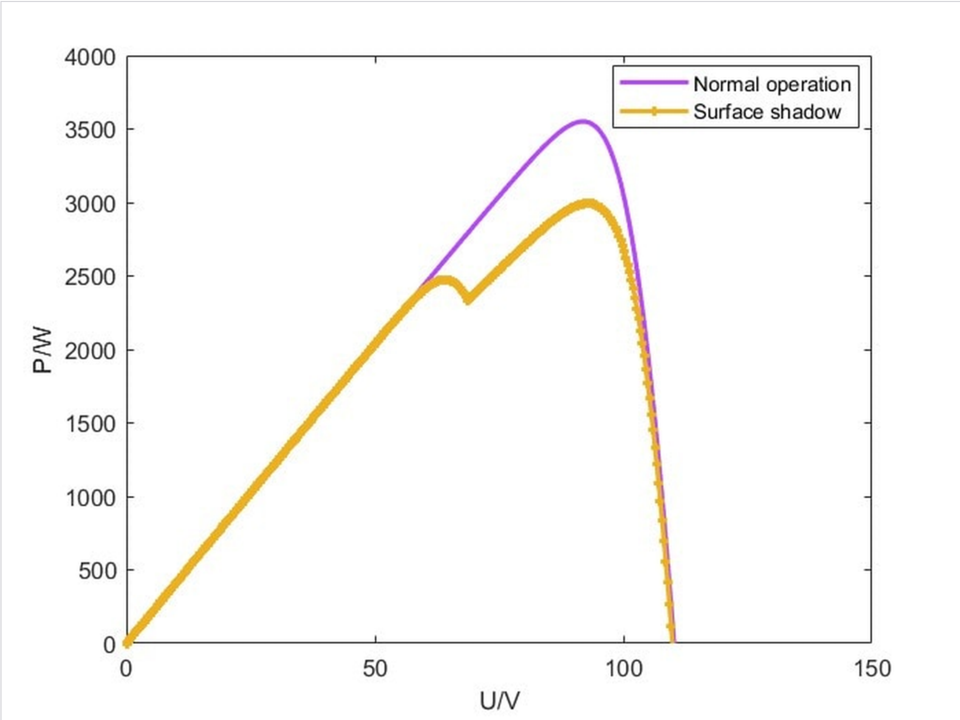

Figure 7. P-V curve when shadows fall on the module surface.

Analysis of the I-V and P-V characteristic curves of a photovoltaic array under STC conditions during a short-circuit fault, as obtained through simulation, reveals that when a shorting wire is used to short-circuit a single PV module in the array, the open-circuit voltage of the array drops significantly compared to normal operating conditions—by approximately 30 V—resulting in a reduction in maximum power output of about 1,100 W. This reduction is directly proportional to the number of affected modules as shown in the figure 8 and 9.

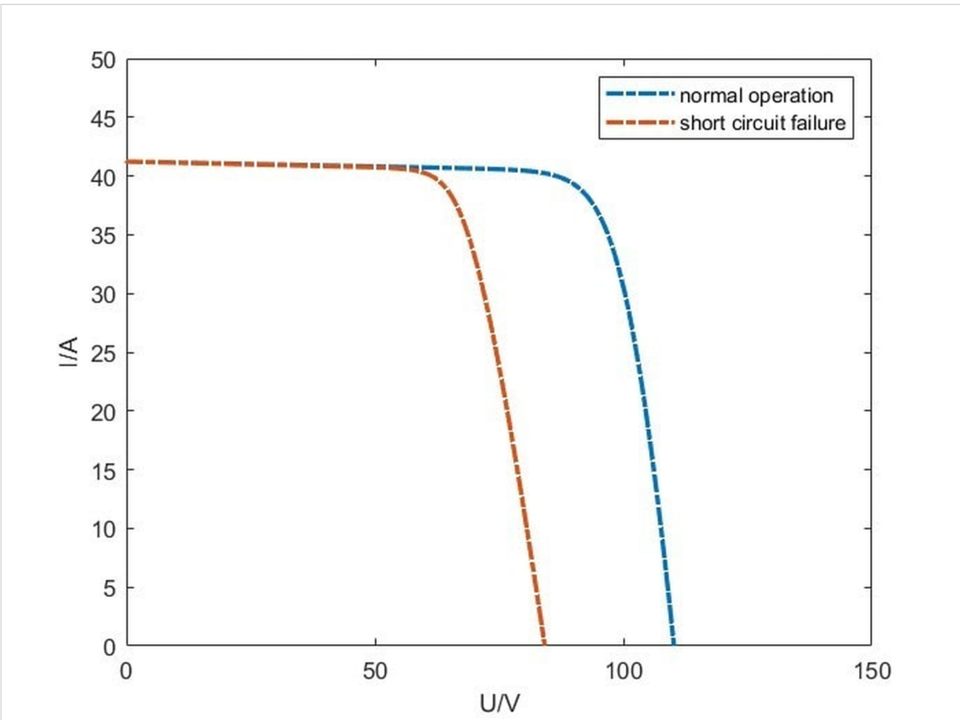

Figure 8. I-V Curve During a Short Circuit Between Cells.

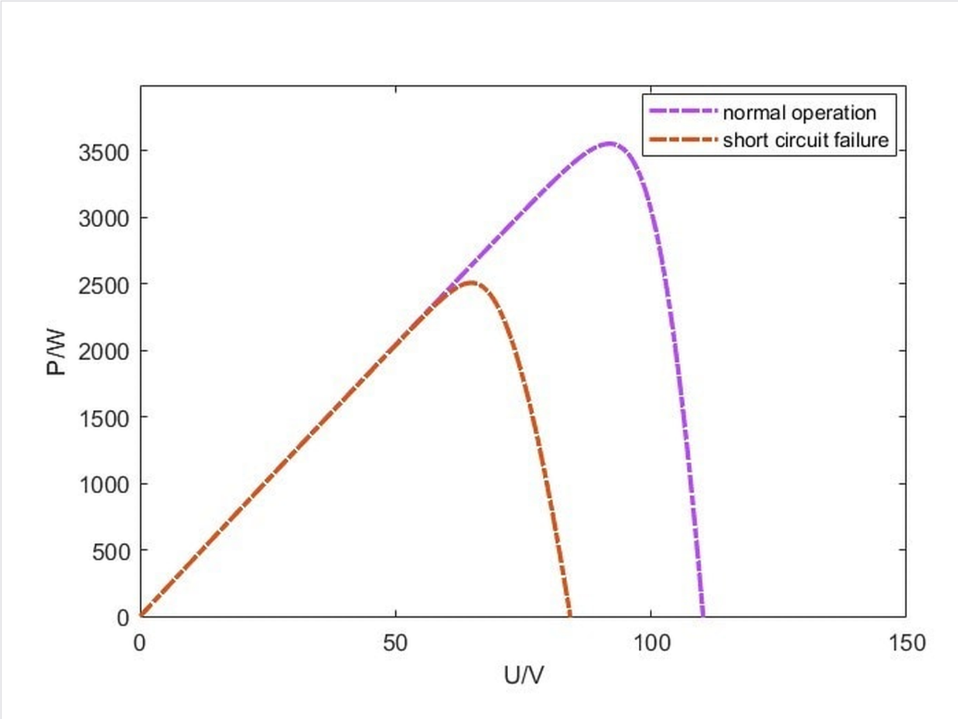

Figure 9. P-V Curve During a Short Circuit Between Cells.

Analysis of the I-V and P-V characteristic curves of the photovoltaic array under STC conditions following a simulation of an open-circuit fault reveals that when a single PV string is disconnected from the array, the array's open-circuit voltage remains largely unchanged, while the short-circuit current decreases significantly, leading to a drop in output power. Furthermore, the maximum output power decreases by approximately 1,200 W compared to normal operating conditions, as shown in figure 10 and 11.

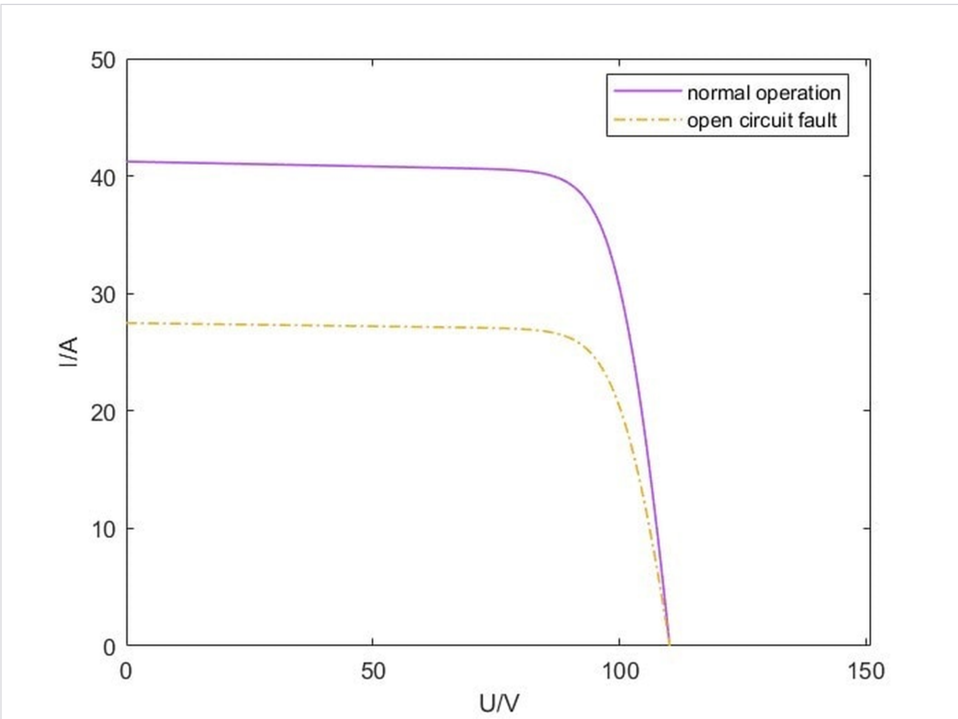

Figure 10. I-V curve for an open-circuit string.

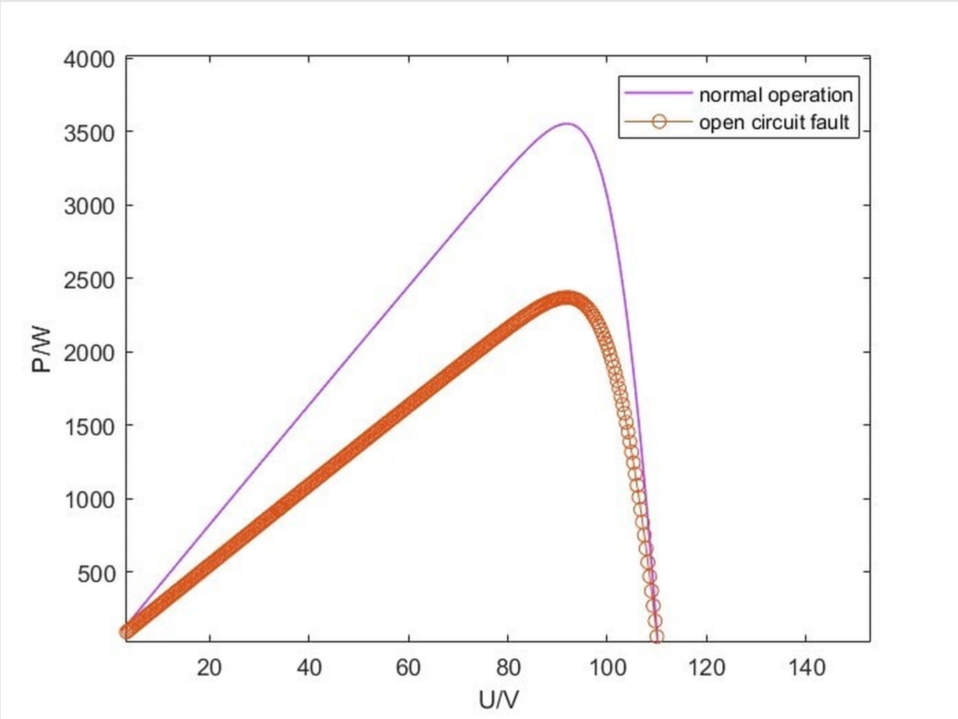

Figure 11. P-V curve for an open-circuit string.

Analysis of the I-V and P-V characteristic curves of the photovoltaic array, obtained through simulation, when aging-related failures occur under STC conditions reveals that, prior to the maximum power point, the output current decreases more rapidly as the output voltage increases compared to normal operating conditions. This results in a reduction in maximum power output relative to normal conditions, with the extent of this reduction being closely related to the degree of array aging, as shown in figure 12 and 13.

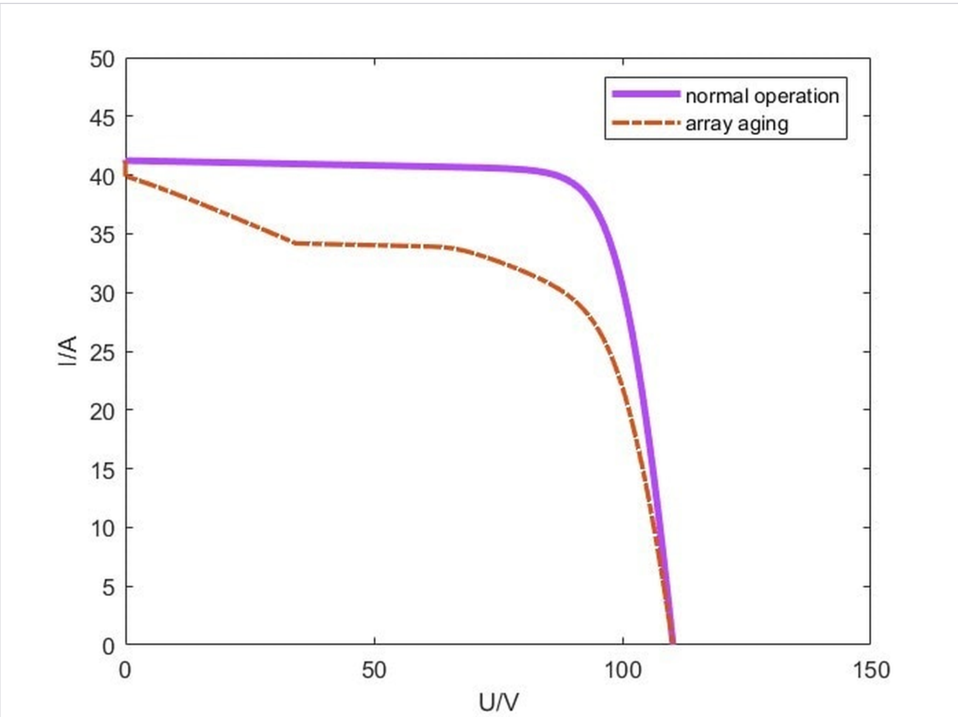

Figure 12. I-V Curves During Array Aging.

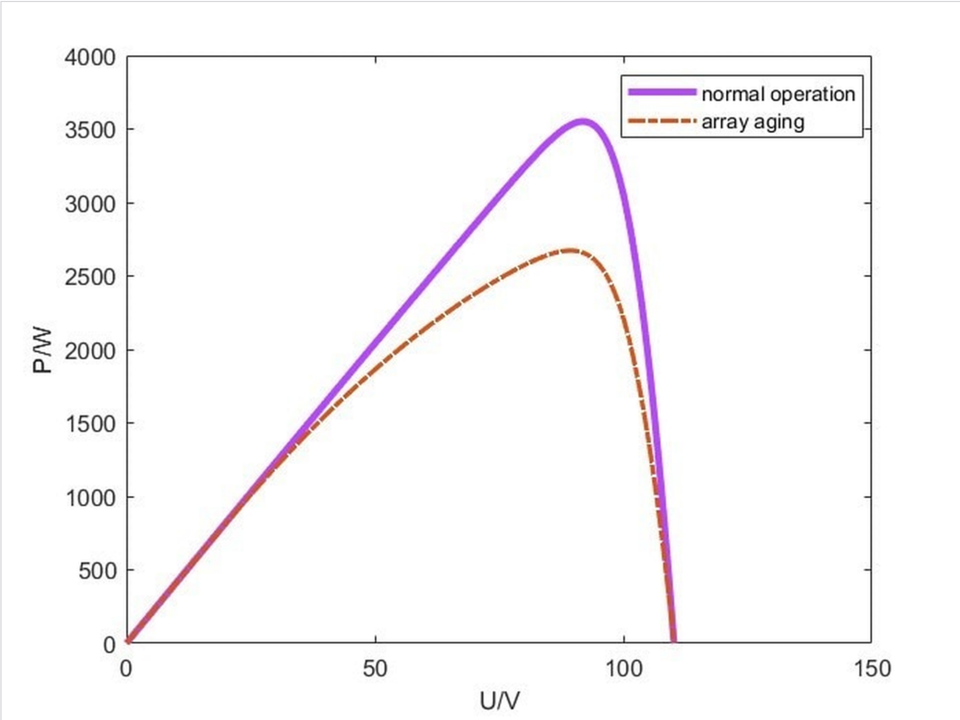

Figure 13. P-V Curves During Array Aging.

This study addresses the issue of array fault diagnosis in off-grid photovoltaic systems by constructing an 8-label classification system comprising seven typical fault modes and one normal operating state, as listed in Table 2. Specifically, Label 1 represents the normal operating state of a 3×3 photovoltaic array and serves as the baseline reference for subsequent fault identification. Label 2 simulates shading on the module surface by randomly selecting 1 to 3 modules in the array and reducing their irradiance by 30% to 90% to characterize varying degrees of shading effects. Label 3 corresponds to aging-related faults in the PV array, achieved by connecting a 5–10 Ω resistor in series to represent performance degradation caused by increased series resistance after long-term operation. Label 4 simulates an open-circuit fault between strings by removing an entire row of PV strings, replicating open-circuit scenarios such as connector detachment or cable breakage. Label 5 describes a short-circuit fault between modules, using ultra-low-resistance resistors of 1×10^{-6} Ω to randomly short-circuit several modules, reflecting short-circuit phenomena caused by hidden cracks in solar cells or connection failures during actual operation. Furthermore, Labels 6–8 focus on more complex concurrent faults: Label 6 combines an inter-string open-circuit fault with shading on the module surface, simultaneously removing one string and applying 30%–90% irradiance attenuation to the remaining modules; Label 7 represents the combination of array aging and shading, superimposing random radiation attenuation on the array in addition to series-connected 5–10 Ω aging resistors; Label 8 combines inter-module short circuits with shading, applying both ultra-low-resistance short circuits and radiation density attenuation simultaneously to the array, thereby comprehensively covering common composite failure scenarios encountered in actual operation. The simulation model covers a solar irradiance range of 100–1000 W/m^2 and a module temperature range of 10–70°C. To enhance the randomness and representativeness of the samples, the intervals between changes in irradiance and module temperature were not fixed, thereby fully simulating the output characteristics under various operating conditions. For each operating scenario, 500 samples were collected, ultimately constructing an array failure dataset comprising 4,000 samples. The dataset was divided into a training set and a test set in a 7:3 ratio, with the training set containing 2,800 samples and the test set containing 1,200 samples (as shown in Table 2). Additionally, Table 3 lists the sample values for various array failures under Standard Test Conditions (STC), providing a baseline reference for subsequent diagnostic analysis.

Table 2. Fault classification labels.

Operating Conditions	Labels	Training set (group)	Test set (group)
Normal operation	1	350	150
Shadows on the component surface	2	350	150
Array aging	3	350	150
Open circuit between strings	4	350	150
Short circuit between components	5	350	150
Path & Shadows	6	350	150
Aging & Shadows	7	350	150
Short Circuit & Shadow	8	350	150

Table 3. Sample fault data of PV arrays under STC.

Operating conditions	G/（W/m²）	T/(°C)	V_{mppt}/(V)	I_{mppt}/(A)	V_{oc}/(V)	I_{sc}/(A)
Normal operation	1000.00	25.00	91.84	38.67	110.19	41.28
Component surface shadow	1000.00	25.00	94.74	25.86	109.58	41.27
Array aging	1000.00	25.00	89.66	29.81	110.19	41.21
Open circuit between strings	1000.00	25.00	91.84	25.78	110.19	27.51
Short circuit between components	1000.00	25.00	64.98	38.57	84.20	41.28
Open circuit and shadow	1000.00	25.00	93.65	19.35	109.52	27.52
Aging and shadow	1000.00	25.00	89.30	29.68	110.13	41.1
Short circuit and shadow	1000.00	25.00	67.16	33.02	83.94	41.27

3. Fault diagnosis modeling

To identify a high-performance fault diagnosis model for photovoltaic arrays, this study developed and compared five diagnostic models based on different principles. Convolutional neural network (CNN) models possess strong feature extraction capabilities and can effectively extract key information from photovoltaic array fault data; the CNN-LSTM model combines the feature extraction strengths of CNN with the temporal data processing advantages of LSTM, enabling better handling of time-dependent fault sequences; CNN-SVM uses SVM to classify features extracted by CNN, further improving diagnostic accuracy; CNN-BiGRU leverages bidirectional gated recurrent networks to enhance the learning and understanding of temporal features in fault data; DBN-ELM combines the deep feature learning capabilities of deep belief networks with the rapid convergence characteristics of extreme learning machines to improve the model's diagnostic efficiency.

Subsequently, these five diagnostic models were systematically trained using a training dataset comprising 2,800 sets of fault samples. Next, each model was applied to the fault diagnosis of off-grid photovoltaic arrays using a test dataset containing 1,200 fault samples. Table 4 details the overall accuracy and accuracy rates for each fault category. The overall accuracy rates for these five models were 84.3%, 86.4%, 90.8%, 92.3%, and 94.7%, respectively. Among them, the DBN-ELM, CNN-SVM, and CNN-BiGRU models achieved relatively high overall accuracy rates, demonstrating excellent fault diagnosis performance. However, there are shortcomings in identifying single-point failures caused by array aging and concurrent failures involving both component shadowing and array aging.

Table 4. Diagnostic accuracy of individual models.

Model	Overall accuracy / %	Accuracy of each fault classification label/ %							
		1	2	3	4	5	6	7	8
CNN	84.3	89.3	80.7	84.7	100.0	98.0	81.3	40.7	100.0
CNN-LSTM	86.4	97.3	86.0	51.3	98.0	100.0	93.3	68.0	97.3
DBN-ELM	90.8	95.3	82.0	72.7	97.3	100.0	98.7	80.0	100.0
CNN-BiGRU	92.3	100.0	86.7	89.3	100.0	100.0	97.3	66.0	98.7
CNN-SVM	94.7	100.0	94.0	82.0	95.3	100.0	100.0	91.3	94.7

To effectively address this issue, this paper proposes the use of a weighted probabilistic averaging ensemble algorithm to perform ensemble optimization on the DBN-ELM, CNN-SVM, and CNN-BiGRU models, thereby constructing a novel diagnostic model.

3.1. Data Preprocessing Layer

The data preprocessing layer is used to standardize the raw data so that it meets the model's input requirements. This study primarily employs normalization to eliminate differences in data units and numerical ranges, thereby improving the stability of model training. Based on the operating characteristics of the photovoltaic array's maximum power point voltage and current, the maximum power voltage and maximum power current are normalized and mapped to a unified range to ensure model convergence and accuracy. Normalize the maximum power voltage and maximum power current; the expressions are shown in Equations 1 and 2.

$$U_{norm} = \frac{U_{mppt}}{U_{oc}} \tag{1}$$

In the equation U_{norm} is the normalized voltage (V); U_{mppt} is the maximum power output voltage of the array (V); U_{oc} is the open-circuit voltage of the array (V).

$$I_{norm} = \frac{I_{mppt}}{I_{sc}} \tag{2}$$

In the equation I_{norm} is the normalized current (A); I_{mppt} is the maximum power output current of the array (A); I_{sc} and is the short-circuit current of the array (A).

3.2. Base Model Layer

3.2.1. DBN-ELM Model

Deep Belief Networks (DBN), as an efficient deep learning model, excel at complex feature extraction and data representation, while Extreme Learning Machines (ELM), as single-hidden-layer feedforward neural networks, are renowned for their efficient training and inference speeds, as well as their rapid convergence and ease of use.

The DBN-ELM model developed in this paper comprises three hidden layers: the first two layers follow the DBN architecture, using a restricted Boltzmann machine to adaptively extract deep features from fault data; the third layer is replaced with an ELM hidden layer, which serves as the classifier. After the features extracted by the DBN are fed into the ELM, the model leverages the ELM's ability to rapidly compute output weights to efficiently generate sample classification probabilities, thereby balancing feature learning capabilities with diagnostic efficiency. The mathematical expression is shown in equation 3.

$$\sum_{i=1}^{N} \beta_i g\left(W_i H_{n-1} + b_i\right) = o_j,\ j = i,\ \ldots,\ m \tag{3}$$

In the equation: $g(x)$ is the activation function; N is the number of nodes in the n th hidden layer; m is the number of nodes in the $n-1$ th hidden layer; β_i is the output weight from the n th hidden layer to the output layer; W_i is the weight from the $n-1$ th hidden layer to the n th hidden layer; H_{n-1} is the output of the $n-1$ th hidden layer; b_i is the bias from the $n-1$ th hidden layer to the n th hidden layer.

3.2.2. CNN-SVM Model

Convolutional neural networks (CNN) consist of convolutional layers and pooling layers. The convolutional layer is the core component of a CNN. It performs convolution operations by sliding a convolutional kernel over the input data the mathematical expressions are given in Equations 4 through 6.

$$O_{i,j,k} = \sigma\left(\sum_{c=1}^{C}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} K_{k,c,m,n} I_{c,s\cdot i+m-p,s\cdot j+n-p} + b_k\right) \tag{4}$$

In the equation: $O_{i,j,k}$ denotes the value of the k th channel of the output feature at position (i,j); $\sigma(x)$ denotes the nonlinear activation function; C denotes the number of input channels; M, N denotes the size of the convolution kernel; $K_{k,c,m,n}$ denotes the value of the k th convolution kernel at position (m,n) in the c th input channel; $I_{c,x,y}$ denotes the value of the input at position (x,y) in the c th channel; s denotes the stride; p denotes the padding; b_k denotes the bias term of the k th convolution kernel.

$$P_{i_1,j_1,k} = \max_{0\le m_1\le M_1, 0\le n_1\le N_1} O_{k,s_1\cdot i_1+m_1,s_1\cdot j_1+n_1} \tag{5}$$

In the equation: $P_{i,j,k}$ represents the value of the k th channel at position (i_1,j_1) in the pooled output; O represents the input feature; M_1, N_1 represents the pooling window size; s_1 represents the pooling stride.

$$y = \sigma(W \cdot vec(O) + b) \tag{6}$$

In the equation: y is the output vector; W is the weight matrix; $vec(O)$ is the processed input feature vector; b is the bias vector.

Support Vector Machines (SVM) is a commonly used classification method in machine learning, SVMs use a kernel function to map the data to a high-dimensional space, where they seek a hyperplane that maximizes the margin. This approach adapts to complex data distributions and improves classification accuracy; the mathematical expressions are given in Equations 7 through 11.

$$\min_{w,b,\varepsilon} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \varepsilon_i \tag{7}$$

$$y_i\left(w^T x_i + b\right) \geq 1 - \varepsilon_i,\ \varepsilon_i \geq 0,\ \forall\ i = 1,\ 2,\ \dots,\ N \tag{8}$$

In the equation: w is the weight vector; b is the bias term; ε_i is the regularization term for the i th sample; C is the penalty parameter; x_i is the feature vector for the i th sample; y_i is the label for the i th sample.

$$\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j K\left(x_i,\ x_j\right) \tag{9}$$

$$0 \leq \alpha_i \leq C,\ \sum_{i=1}^{N} \alpha_i y_i = 0 \tag{10}$$

In the equation: α_i is the Lagrange multiplier for the i th sample; $K(x_i, x_j)$ is the sum function.

$$f(x) = sign\left(\sum_{i=1}^{N} \alpha_i y_i K\left(x_i,\ x\right) + b\right) \tag{11}$$

In the equation: f(x) is the predicted class of sample x; $sign(\bullet)$ is a symbolic function that outputs the classification result.

The CNN-SVM model developed in this paper employs an alternating structure of two layers of convolutional and pooling operations to extract features, followed by an SVM for classification. Through convolutional and pooling operations, the CNN adaptively extracts highly representative feature vectors from the raw fault data; the SVM then constructs an optimal classification hyperplane based on these feature vectors to accurately identify fault types and output classification probabilities.

3.2.3. CNN-BiGRU Model

The Gate-Recurrent Unit (GRU) is an improved type of recurrent neural network unit that controls information propagation via an update gate and a reset gate. their mathematical expressions are shown in Equations 12 through 15.

$$z_t = \sigma\left(W_z x_t + U_z h_{t-1} + b_z\right) \tag{12}$$

In the equation: z_t is the update gate vector, which controls the extent of the state update; W_z, U_z is the weight matrix of the update gate; b_z is the bias vector of the update gate; x_t is the input vector for the current time step; h_{t-1} is the hidden state from the previous time step; $\sigma(x)$ is the activation function.

$$r_t = \sigma\left(W_r x_t + U_r h_{t-1} + b_r\right) \tag{13}$$

In the equation: r_t is the reset gate vector; W_r, U_r is the reset gate weight matrix; b_r is the reset gate bias vector.

$$\widetilde{h}_t = tanh\left(W_h x_t + U_h\left(r_t \odot h_{t-1}\right) + b_h\right) \tag{14}$$

In the equation: $\widetilde{h}_t$ is the candidate hidden state vector; W_h, U_h is the weight matrix for the candidate states; b_h is the bias vector for the candidate states.

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \widetilde{h}_t \tag{15}$$

In the equation: h_t is the hidden state output at the current time step.

The Bidirectional Gated Recurrent Unit (BiGRU) adds a backpropagation path to the standard GRU. This bidirectional structure is well-suited for handling such complex temporal dependencies. Its mathematical expressions are shown in Equations 16 to 18.

$$\vec{h}_t = (1 - \vec{z}_t) \odot \vec{h}_{t-1} + \vec{z}_t \odot \overrightarrow{\widetilde{h}}_t \tag{16}$$

$$\overleftarrow{h}_t = (1 - \overleftarrow{z}_t) \odot \overleftarrow{h}_{t+1} + \overleftarrow{z}_t \odot \overleftarrow{\widetilde{h}}_t \tag{17}$$

$$h_t = [\vec{h}_t; \overleftarrow{h}_t] \tag{18}$$

In the equation: The parameters marked with $\rightarrow$, $\leftarrow$ are those for the forward GRU and the backward GRU, respectively; h_t is the merged bidirectional hidden state.

The CNN-BiGRU model developed in this paper consists of an alternating structure of two convolutional layers and pooling layers. The convolutional layers perform deep extraction of spatial features from array data by using convolutional kernels of different sizes. Based on the features in the CNN's output, the number of channels in the BiGRU is determined to connect the CNN and BiGRU. The BiGRU layer then further captures the temporal and dynamic features of the data, helping the model better identify correlations among the data. The output layer uses the Softmax activation function to calculate the probability of each category corresponding to a given data set. Finally, the output layer outputs the model's probability assessment of the sample type.

During the integration process, the three base models are first trained and optimized to ensure that each model can independently produce accurate classification probabilities; subsequently, through multiple rounds of testing, the optimal weighting scheme is determined by evaluating metrics such as overall accuracy and recall.

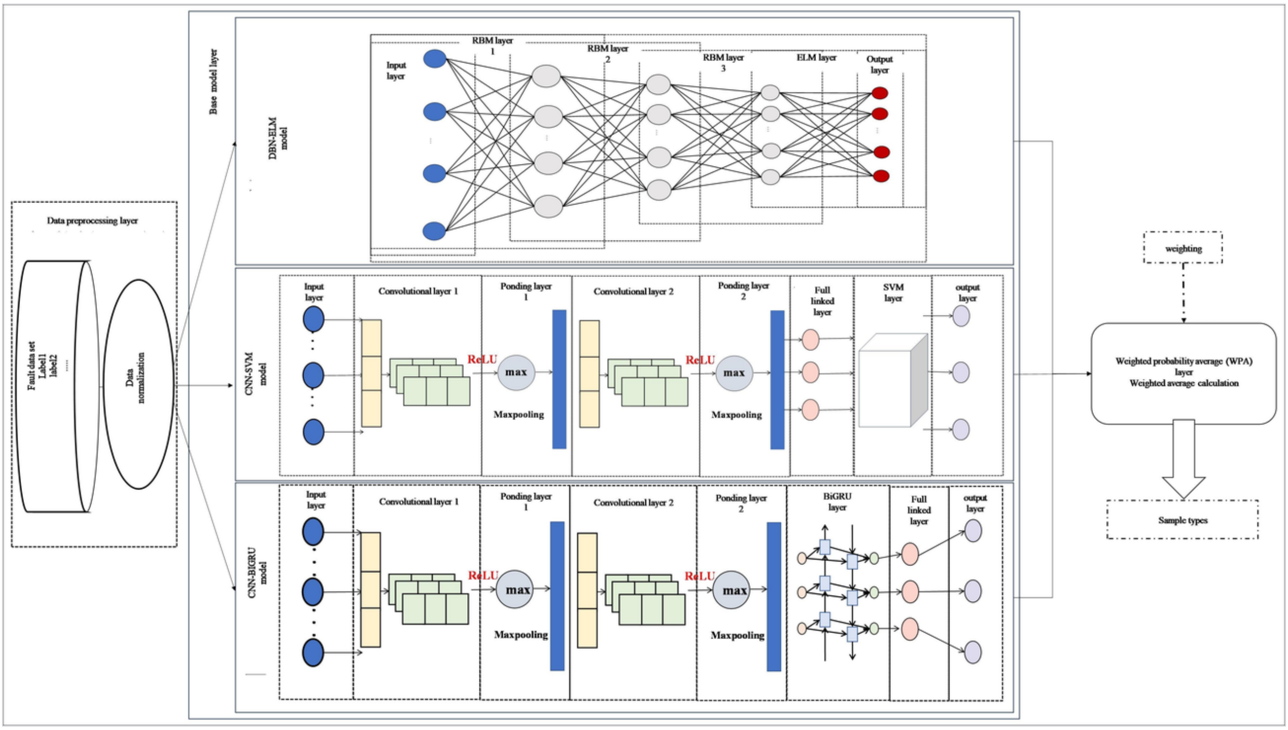

Figure 14. Schematic diagram of the integration model.

3.3. Weighted Probability Averaging Layer

The schematic diagram of the weighted average probability-based integration model proposed in this paper is presented in Figure 14. The data preprocessing layer processes raw data to convert it into a format that satisfies the requirements of the task. This model primarily employs normalization for the maximum power voltage and maximum power current. The base model layer consists of the integrated diagnostic models: DBN-ELM, CNN-SVM, and CNN-BiGRU. The weighted probability average calculation layer employs the weighted probability average ensemble learning algorithm to compute the weighted average of the sample type prediction probabilities generated by each base model in the base model layer. Specifically, this approach synthesizes the performance of each individual model across different samples, assigns an appropriate output weight accordingly, and identifies the type with the highest computed probability as the final prediction result. The detailed calculation is presented in Equation 19.

$$P_n(x) = \sum_{k=1}^{K} \omega_k P_{kn}(x) \tag{19}$$

In the equation: K is the total number of base models; k is the model index, x is the sample number index, and n n denotes the operating conditions index; ω_k is the weight assigned to the base model of index k ; $P_{kn}(x)$ is the probability that the base model classifies sample k as type n ; and $P_n(x)$ is the probability that the integrated model classifies sample x as type n.The final judgment result is calculated as shown in Equation 20.

$$E(x) = max\ P_n(x) \tag{20}$$

During the model ensemble phase, the three base models are first thoroughly trained and optimized. Through learning from a large dataset, we ensure that each model can independently and accurately output the probability of a sample's class. Subsequently, through multiple rounds of comparative testing, different weighting schemes are validated on the test set. The optimal weights are determined by evaluating performance metrics such as overall accuracy and recall.

The category probabilities output by each base model are then weighted and averaged according to the optimal weights, and the category with the highest weighted probability is selected as the final diagnostic result for the sample, thereby achieving accurate identification of photovoltaic faults.

3.4. Integrated Model Diagnostic Process

The model diagnosis process is shown in Figure 15.

1. Divide the collected dataset of output data from the off-grid PV system array into training and testing datasets in a 7:3 ratio;
2. Normalize the training and testing datasets, scaling the data to the range [0,1];
3. Set the initial model network parameters and construct the three basic diagnostic models;
4. Train the three composite diagnostic models using the normalized training set, optimize the parameters, and complete feature extraction from the training data;
5. After parameter optimization, calculate the loss value; when the loss value is minimized, save the current parameters, which represent the optimal state for each model;
6. Ensemble the three composite diagnostic models in their optimal states using the weighted probability averaging algorithm to build the ensemble learning diagnostic model, then set the weight allocations;
7. Finally, validate the model's diagnostic performance on the normalized test set and output the performance evaluation results.

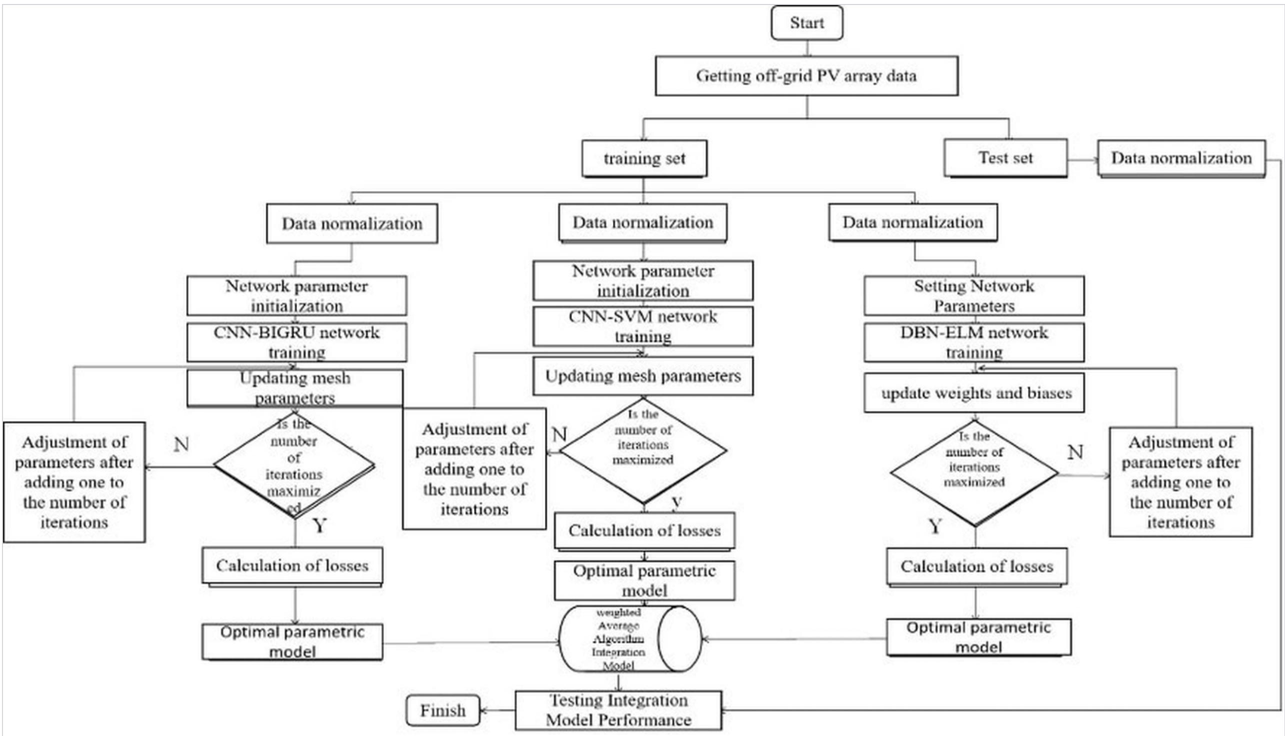

Figure 15. Diagnostic process.

4. Model Testing and Evaluation

4.1. Evaluation Criteria

To evaluate the performance of the fault diagnosis model for the PV array, this paper employs four widely recognized diagnostic evaluation metrics: accuracy, recall, precision, and the F1-Score. Accuracy is the proportion of correctly classified samples to the total number of samples, reflecting the overall classification accuracy of the model; Recall is the proportion of successfully predicted positive samples to all true positive samples, reflecting the model's ability to identify fault samples; Precision is the proportion of true positive samples to predict positive samples; F1-Score is the harmonic average of Recall and Precision, reflecting the reliability of the model's prediction results. The formula for calculating each indicator is shown in equations 21 to 24.

$$A = \frac{T_P + T_N}{T_P + F_N + F_P + T_N} \tag{21}$$

$$R = \frac{T_P}{T_P + F_N} \tag{22}$$

$$P = \frac{T_P}{T_P + F_P} \tag{23}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{24}$$

In the equation: A is the accuracy rate; R is the recall rate; P is the precision; F_1 is the F1-Score; T_P refers to the number of samples in which faulty samples are correctly identified as faulty; T_N refers to the number of samples in which normal samples are correctly identified as normal; F_P refers to the number of samples in which faulty samples are incorrectly classified as normal; F_N refers to the number of samples in which normal samples are incorrectly classified as faulty.

The performance of the integrated model is significantly influenced by key parameters, primarily including the dataset split ratio, the hyperparameters of each base model network, and the weight allocation among models. This study trained the model using the 2,800 training samples described earlier. Through multiple rounds of iterative experiments and parameter tuning, the optimal values for each parameter were determined: the dataset was split into a 7:3 training-to-test ratio to ensure both sufficient model learning and the validity of generalization capability assessments; the hyperparameters of the base models were optimized via grid search and cross-validation; and the ensemble weights were dynamically adjusted based on test set performance. The final optimized values for each parameter are detailed in Table 5.

Table 5. Diagnostic Model Parameters.

Model Structure		Parameter	Description
Data Preprocessing Layer		Training set : Test set ratio	7:3
		Data type	8
		Data length	4000
Base Model Layer	DBN-ELM Model	DBN hidden layer 1	Number of neurons: 20
		DBN hidden layer 2	Number of neurons: 20
		ELM hidden layer	Number of neurons: 41
	CNN-BiGRU Model	Convolutional layer 1	Number of kernels: 16; Kernel size: 2×1; Input channels per kernel: 1; Stride: 1
		Pooling layer 1	Size: 2×1; Stride: 1
		Convolutional layer 2	Number of kernels: 32; Kernel size: 2×1; Input channels per kernel: 16; Stride: 1
		Pooling layer 2	Size: 2×1; Stride: 1
		BiGRU layer	Forward GRU hidden units: 4; Backward GRU hidden units: 4
	CNN-SVM Model	Convolutional layer 1	Number of kernels: 32; Kernel size: 3×1; Input channels per kernel: 1; Stride: 1
		Pooling layer 1	Size: 2×1; Stride: 1
		Convolutional layer 2	Number of kernels: 32; Kernel size: 2×1; Input channels per kernel: 32; Stride: 1
		Pooling layer 2	Size: 2×1; Stride: 1
		SVM layer	Kernel function: Sigmoid function
Weighted Probability		Weight distribution ratio	CNN-SVM : CNN-BiGRU : DBN-ELM = 5 : 4 : 1

After several iterations of training and experimental adjustments, the debugged integrated diagnostic model was employed to diagnose fault samples in the designated test dataset. The four evaluation metrics achieved an accuracy of 99 %, as shown in Figure 16. The confusion matrix for the diagnostic model shows that only 0.67% of the samples were misclassified when distinguishing between a single array aging failure (Label 3) and a concurrent failure involving both component shading and array aging (Label 7).

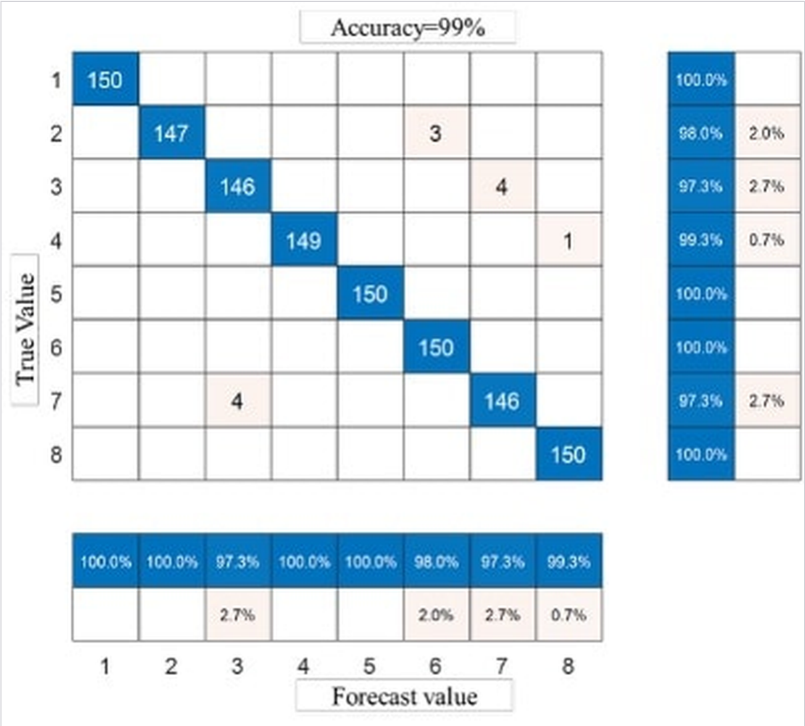

Figure 16. Confusion matrix for the integrated model.

Comparative performance tests were also conducted between the integrated model and several other models, and the results demonstrated that the integrated model significantly outperformed the others, as illustrated in Figure 17.

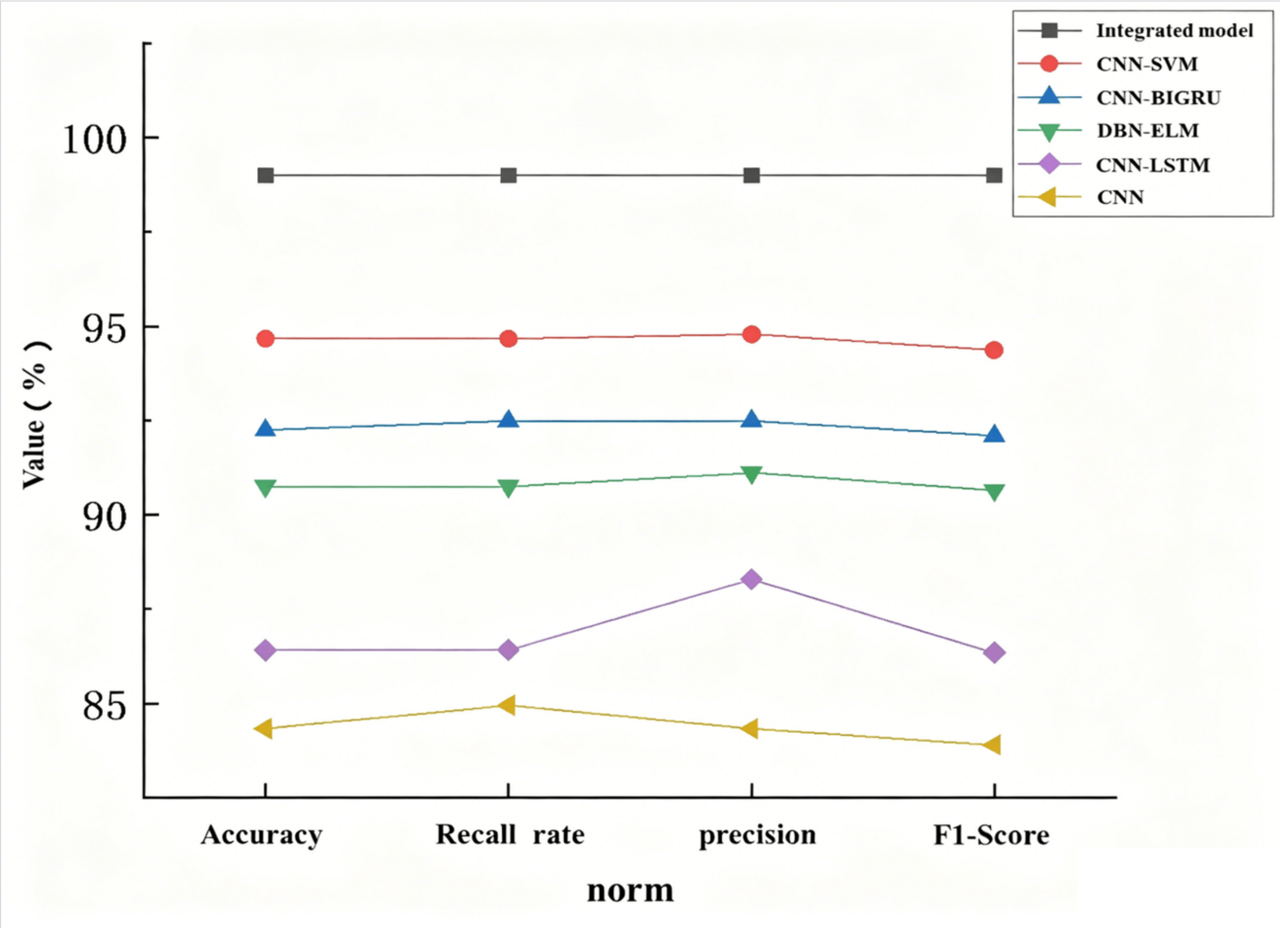

Figure 17. Comparison of evaluation indicators.

A systematic comparative analysis was conducted between the integrated model and the five fault diagnosis models mentioned. During the experiments, the overall evaluation metrics for each model and the sub-metrics for each fault category were calculated. The results show that, under the same dataset and testing conditions, the overall diagnostic accuracy rates for the five models—CNN, CNN-LSTM, DBN-ELM, CNN-BiGRU, and CNN-SVM—were 84.3 %, 86.4 %, 90.8 %, 92.3 %, and 94.7%, respectively. The performance of the ensemble model is significantly better than that of a single model.

4.2. Experimental verification

The integrated diagnostic model developed in this paper was applied to a small-scale, standalone off-grid photovoltaic experimental system to comprehensively evaluate the model's performance under actual operating conditions. The performance evaluation adopted a modular design approach, integrating photovoltaic array power generation, fault simulation, and data acquisition into a single system. Data collection spanned a period of three months, and the collected data was used to conduct a comprehensive assessment of the integrated diagnostic model's performance.

By applying the integrated diagnostic model to the small-scale, standalone off-grid photovoltaic experimental system that has been constructed as shown in Figure 18, this study aims to comprehensively evaluate the model's performance under actual operating conditions. The performance evaluation adopts a modular design approach, integrating photovoltaic array power generation, fault simulation, and data acquisition into a single system. Data collection spanned a period of three months, and the collected data was used to conduct a comprehensive assessment of the integrated diagnostic model's performance.

Figure 18. Photovoltaic array.

The fault simulation device comprises four single-fault modules and three concurrent-fault modules. The four single-fault configurations are shown in figure 19, and their specific construction and operating principles are as follows: Surface shading simulation is achieved by partially obscuring the photovoltaic module with a cardboard sheet having 0% light transmittance; the shaded area can be flexibly adjusted within the range of 30% to 90%. Module short-circuit simulation is achieved by directly shorting the module's positive and negative terminals using a double-ended shorting lead; String open-circuit simulation is achieved by disconnecting the wiring of a single PV string; array aging simulation is performed by connecting a sliding resistor in series within the array and adjusting its resistance to simulate varying degrees of aging and The three concurrent fault simulations are based on the integration of the aforementioned single-fault modules. By simultaneously triggering multiple fault types, they simulate complex fault scenarios encountered during actual operation.

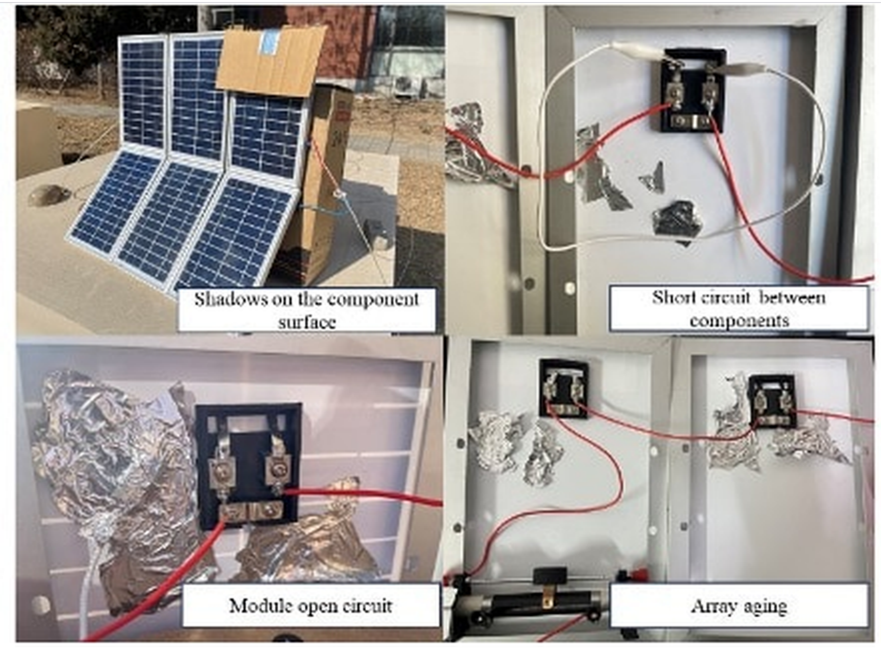

Figure 19. Array Failure Settings.

During data collection, with the photovoltaic array left unchanged under both sunny and cloudy conditions, modules representing seven different fault states were introduced, and the data acquisition system recorded the readings. Data was collected at 15-minute intervals. A portion of the collected data is shown in the Table 6.

Table 6. Application Results

	Accuracy rate /%	Recall rate/%	Accuracy /%	F1score/%
overall	96.5	96.7	98.7	97.7
label1	100.0	100.0	100.0	100.0
label2	98.3	98.3	100.0	99.2
label3	93.3	93.3	100.0	96.6
label4	100.0	100.0	100.0	100.0
label5	100.0	100.0	100.0	100.0
label6	100.0	100.0	98.4	99.2
label7	80.0	80.0	92.3	85.7
label8	100.0	100.0	100.0	100.0

The integrated fault diagnosis model, which leverages multi-source data fusion, maintains high accuracy and strong robustness in real-world environments, achieving an overall diagnostic accuracy rate of 96.5%—a decrease of only 2.5 percentage points compared to simulation environments. The model demonstrates excellent adaptability to complex operating conditions and equipment anomalies, providing reliable technical support for the operation and maintenance of off-grid photovoltaic systems.

5. Conclusion

This paper addresses the fault diagnosis problem of photovoltaic arrays by constructing an 8-label classification system and analyzing corresponding output characteristics. A fault diagnosis method is proposed for photovoltaic arrays based on a weighted probabilistic averaging ensemble model. This method achieves a high recognition rate for photovoltaic array faults as validated by experiments. The main conclusions are as follows:

1. A fault dataset containing 4000 samples was constructed, with the training set and test set divided in a 7:3 ratio. Specifically, the training set consisted of 2800 samples, while the test set contained 1200 samples. For each of the eight operating conditions, there were 350 samples in the training set and 150 samples in the test set.
2. A photovoltaic array fault diagnosis model based on the weighted probability average ensemble learning algorithm is constructed. The model integrates the advantages of CNN-SVM, CNN-BiGRU, and DBN-ELM three base models, and by weighting and summing the judgment probabilities of the outputs of these three base models, the fault type with the highest weighted probability sum is finally determined as the diagnosis result.
3. Accuracy, recall, precision, and F1-Score were selected as effective evaluation metrics for the model. The proposed diagnostic model achieved a score of 99.0% on all four metrics, demonstrating a high level of accuracy in identifying both single and multiple faults.
4. The diagnostic model proposed in this paper achieves a high diagnostic accuracy rate of 96.5% in actual off-grid photovoltaic systems, which is only 2.5 percentage points lower than its performance on simulated data, fully validating its feasibility and practicality in real-world applications.

Declarations

Author contributions. Chunmei Guo: Conceptualisation: Developed the overarching themes and objectives of the manuscript. **Weijin Sun**: Writing – Original Draft: Wrote the initial manuscript draft, including key sections. **Yang Li:** Data Curation: Organized and synthesised data from selected studies for inclusion in the manuscript. **Yuwen You**: Methodology: Designed the review framework and criteria for literature selection. **Zhonglu He**: Writing – Review & Editing: Assisted in editing and refining the final manuscript, ensuring accuracy and completeness

Conflicts of 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.

Data availability. The raw experimental data, MATLAB simulation codes and parametric calculation files supporting the findings of this study are available from the corresponding author upon reasonable request.

Ethics. Ethical approval is not applicable as this work contains no human or animal related experiments.

Editorial independence. All research work in this paper was completed independently by the authors. No editors or editorial staff intervened in the design, experiment, simulation and writing of this study.

Acknowledgements. This research was supported by National Key R&D Program of China (2023YFC3806100).

References

Ali, Y.M., Ding, L., Qin, S. (2025) 'An efficient approach for diagnosing faults in photovoltaic array using 1D-CNN and feature selection Techniques', *International Journal of Electrical Power & Energy Systems*, 166, pp. 110526. Available at: https://doi.org/10.1016/j.ijepes.2025.110526

Aljafari, B., Satpathy, P.R., Thanikanti, S.B., Nwulu, N. (2024) 'Supervised classification and fault detection in grid-connected PV systems using 1D-CNN: Simulation and real-time validation', *Energy Reports*, 12, pp. 2156-2178. Available at: https://doi.org/10.1016/j.egyr.2024.08.008

Aziz, F., Ul Haq, A., Ahmad, S., Mahmoud, Y., Jalal, M., Ali, U. (2020) 'A Novel Convolutional Neural Network-Based Approach for Fault Classification in Photovoltaic Arrays', *IEEE Access*, 8, pp. 41889-41904. Available at: https://doi.org/10.1109/ACCESS.2020.2977116

Fu, H., Liu, H., Xie, S., Liu, S., Han, H., Ma, J. (2025) 'Multi-coupling fault detection and diagnosis of photovoltaic arrays with improved slime mould algorithm and PolyCatBoost', *Process Safety and Environmental Protection*, 194, pp. 523-541. Available at: https://doi.org/10.1016/j.psep.2024.11.135

Gao, X. and Qian, Y (2024) 'Fault diagnosis of photovoltaic arrays based on VMD-SABO-KELM', *Information Technology and Informatization*.

Hajji, M., Harkat, M.F., Kouadri, A., Abodayeh, K., Mansouri, M., Nounou, H., Nounou, M. (2021) 'Multivariate feature extraction based supervised machine learning for fault detection and diagnosis in photovoltaic systems', *European Journal of Control*, 59, pp. 313-321. Available at: https://doi.org/10.1016/j.ejcon.2020.03.004

Karmacharya, I., Gokaraju, R. (2018) 'Fault Location in Ungrounded Photovoltaic System Using Wavelets and ANN', *2019 IEEE Power & Energy Society General Meeting (PESGM)*, pp. 1-1. Available at: https://doi.org/10.1109/PESGM40551.2019.8973821

Lin, P., Guo, F., Lu, X., Zheng, Q., Cheng, S., Lin, Y., Chen, Z., Wu, L., Qian, Z. (2024) 'A compound fault diagnosis model for photovoltaic array based on 1D VoVNet-SVDD by considering unknown faults', *Solar Energy*, 267, pp. 112155. Available at: https://doi.org/10.1016/j.solener.2023.112155

Liu, Y., Ding, K., Zhang, J., Li, Y., Yang, Z., Zheng, W., Chen, X. (2021) 'Fault diagnosis approach for photovoltaic array based on the stacked auto-encoder and clustering with I-V curves', *Energy Conversion and Management*, 245, pp. 114603. Available at: https://doi.org/10.1016/j.enconman.2021.114603

Lu, Xiaoyang; Lin, Yaohai; Lin, Peijie; He, Xiangjian; Fang, Gengfa; Cheng, Shuying; Chen, Zhicong; Wu, Lijun (2023) 'Efficient fault diagnosis approach for solar photovoltaic array using a convolutional neural network in combination of generative adversarial network under small dataset', *Solar Energy*, 253, pp. 360-374. Available at: https://doi.org/10.1016/j.solener.2022.12.037

Lu, S.D., Liu, H.D., Wang, M.H., Wu, C.C. (2024) 'A novel strategy for multitype fault diagnosis in photovoltaic systems using multiple regression analysis and support vector machines', *Energy Reports*, 12, pp. 2824-2844. Available at: https://doi.org/10.1016/j.egyr.2024.08.074

Mustafa, Zain; Awad, Ahmed S.A.; Azzouz, Maher; Azab, Ahmed (2023) 'Fault identification for photovoltaic systems using a multi-output deep learning approach', *Expert Systems with Applications*, 211, pp. 118551. Available at: https://doi.org/10.1016/j.eswa.2022.118551

Patthi, S., Murali Krishna, V., Reddy, L., Arandhakar, S. (2024) 'Photovoltaic string fault optimization using multi-layer neural network technique', *Results in Engineering*, 22, pp. 102299. Available at: https://doi.org/10.1016/j.rineng.2024.102299

Turhal, U.C., Onal, Y., Turhal, K. (2025) 'Enhanced Fault Detection and Diagnosis in Photovoltaic Arrays Using a Hybrid NCA-CNN Model', *Computer Modeling in Engineering & Sciences*, 143(2), pp. 2307-2332. Available at: https://doi.org/10.32604/cmes.2025.064269

Voutsinas, S., Karolidis, D., Voyiatzis, I., Samarakou, M. (2022) 'Development of a multi-output feed-forward neural network for fault detection in Photovoltaic Systems', *Energy Reports*, 8, pp. 33-42. Available at: https://doi.org/10.1016/j.egyr.2022.06.107

Wang, X.x., Dong, L., Liu, S.y., Hao, Y., Wang, B. (2019) 'A Fault Classification Method of Photovoltaic Array Based on Probabilistic Neural Network', *2019 Chinese Control And Decision Conference (CCDC)*, pp. 5260-5265. Available at: https://doi.org/10.1109/CCDC.2019.8832338

Yahyaoui, Z., Hajji, M., Mansouri, M., Kouadri, A., Bouzrara, K., Nounou, H. (2024) 'Enhancing Fault Diagnosis of Uncertain Grid-Connected Photovoltaic Systems using Deep GRU-based Bayesian optimization', *IFAC-PapersOnLine*, 58(4), pp. 449-454. Available at: https://doi.org/10.1016/j.ifacol.2024.07.259

Zhong, S., Chen, Z., Wu, L (2024) 'Fault diagnosis of photovoltaic arrays based on an improved SqueezeNet', *Power Electronics Technology*.

HOW TO CITE THIS ARTICLE

Guo, C., Sun, W., Li, Y., You, Y. and He, Z. (2026). Development of an Integrated DBN-ELM, CNN-SVM and CNN-BiGRU Photovoltaic Array Fault Diagnosis Model Based on Weighted Probability Averaging. Energy Catalyst, 2, 106-131. https://doi.org/10.65582/ec.2026.006