Measurement Science and Technology



PAPER

RECEIVED
4 January 2026

REVISED
7 May 2026

ACCEPTED FOR PUBLICATION
21 May 2026

PUBLISHED
8 June 2026

# Few-shot learning photovoltaic fault diagnosis based on Tri-MSTCN

**Yanbo Jian**[1], **Aimin An**[1,2], **Shuyou Yu**[1,3,*] **and Tao Chang**[1]

[1] School of Automation and Electrical Engineering, Lanzhou University of Technology, Lanzhou, People's Republic of China
[2] Key Laboratory of Gansu Advanced Control for Industrial Processes, Lanzhou, People's Republic of China
[3] Department of Control Science and Engineering, Jilin University, Changchun 130012, People's Republic of China
[*] Author to whom any correspondence should be addressed.

**E-mail:** yushuyou@126.com and 2631361590@qq.com

**Keywords:** photovoltaic array, few-shot learning, triplet Siamese network, fault diagnosis

## Abstract

To address the scarcity of fault samples in photovoltaic (PV) arrays and the tendency of conventional data-driven models to overfit under small-sample conditions, this paper proposes a triplet-Siamese fault-diagnosis framework, termed triplet-Siamese-multi-scale temporal convolutional network (Tri-MSTCN). First, an operating model of the PV array is established to collect time-series samples. A triplet sampling strategy is then adopted to construct anchor–positive–negative tuples, which enlarges the effective training set and increases sample diversity. The proposed network employs a multi-scale temporal convolutional network (TCN) to extract features at multiple temporal resolutions, integrates a channel-attention module to emphasize informative channels, and introduces a gated residual fusion mechanism to adaptively combine the reweighted features with the original input, thereby enhancing feature representation. The experimental results show that from 45 samples to 135 samples, the accuracy of the Tri-MSTCN model significantly improved, increasing from 96.67% to 98.89%. In contrast, it is clearly higher than other models, such as GRU, CBG, and TCN. As the sample size increases, the accuracy of Tri-MSTCN increases significantly, further validating the model's effectiveness and generalization ability in PV array fault diagnosis.

## 1. Introduction

Photovoltaic (PV) power generation, as a key component of clean energy systems, has become an important driver of the global energy transition. With the increasing penetration of renewable energy, PV arrays deployed in the field are inevitably exposed to environmental variability, component aging, and external disturbances, which may induce various faults [1]. These faults not only reduce the energy yield of PV systems but may also accelerate equipment degradation and, in severe cases, threaten grid stability. Therefore, developing efficient and accurate fault-diagnosis techniques for PV arrays is of great significance for improving system reliability and supporting sustainable energy deployment.

To address the challenges posed by limited fault samples, several existing studies have primarily focused on transfer learning, data augmentation, meta-learning, and metric-learning-based few-shot diagnosis frameworks. For instance, literature [2] proposed a meta-transfer learning framework embedded with prior knowledge, where diagnostic prior knowledge was integrated with a meta-learning module to improve model generalization under small-sample conditions. In the context of rotating machinery fault diagnosis, literature [3] developed a small-sample transfer learning framework based on a unified one-dimensional convolutional network, systematically comparing feature transfer, fine-tuning, and meta-learning methods under various transfer scenarios. Moreover, literature [4] introduced an interpretable multi-domain meta-transfer learning framework that incorporated a multi-task transfer structure and an activation-mapping fusion mechanism, enhancing robustness and interpretability under variable operating conditions. Additionally, literature [5] proposed a constructive incremental learning and

integrated domain adaptation method, which improved adaptability and diagnostic accuracy in small-sample cross-domain scenarios by combining wavelet packet decomposition, incremental domain adaptation, and a majority-voting mechanism. Similarly, literature [6] introduced an information-fusion-based meta-transfer diagnosis method, where sparse principal component analysis (PCA) and multi-sensor information fusion were employed to improve the generalization performance of small-sample fault diagnosis under different operating conditions.

In addition to transfer learning, data augmentation has been widely adopted to alleviate the scarcity and imbalance of fault samples by expanding the training set through signal transformations and generative models. Basic augmentation strategies, such as noise injection and time stretching, have been verified to be effective in literature [7], where these techniques were applied to raw vibration signals combined with deep networks for fault diagnosis. Generative models have further advanced this direction. Specifically, literature [8] and literature [12] employed improved generative adversarial networks to generate fault signals and enhance diagnostic accuracy under small-sample conditions, while literature [9] improved the quality of generated samples by using a weighted modified conditional variational autoencoder, achieving better performance under class-imbalanced conditions. More recently, diffusion-based augmentation methods were introduced in literature [10] and literature [11], which improved the quality, diversity, and stability of generated samples by enhancing vibration signal generation or simulating the noise diffusion process. Furthermore, literature [13] and literature [14] incorporated continuous wavelet transform, attention mechanisms, contrastive learning, and Fourier CNN into the augmentation framework, making the generated samples closer to the original data distribution and further improving the recognition accuracy and generalization capability of diagnostic models.

From another perspective, meta-learning aims to improve rapid adaptation to new tasks by learning transferable knowledge from multiple related tasks. Representative optimization-based meta-learning methods, including MAML, Meta-SGD, Reptile, and ANIL, have laid the methodological foundation for fast adaptation by learning transferable initialization parameters, update directions, and lightweight adaptation strategies [15–18]. Building on these advances, literature [19] formulated small-sample fault diagnosis as an $N$-way $K$-shot task and performed meta-training based on MAML, enabling rapid adaptation under complex operating conditions. Furthermore, literature [20] proposed a GMAML-based method for cross-domain few-shot bearing fault diagnosis driven by heterogeneous signals, where a channel-interaction attention feature encoder and an inner-loop weight-guidance mechanism were designed to enhance the extraction of shared knowledge across tasks.

In addition to transfer learning, data augmentation, and meta-learning, metric-learning-based few-shot frameworks have shown strong potential in small-sample fault diagnosis, particularly those based on Siamese networks [21]. However, PV-array fault diagnosis remains challenging because PV electrical signals are multivariate time series, which are strongly influenced by irradiance and temperature. Existing few-shot models often struggle to capture multi-scale temporal dependencies or are prone to overfitting under scarce data conditions. To address this issue, this paper proposes Triplet-Siamese multi-scale temporal convolutional network (Tri-MSTCN), a triplet Siamese framework that combines triplet sampling, multi-scale temporal convolution, channel attention, and gated residual fusion for robust PV-array fault diagnosis under small-sample conditions.

Based on the above design, the main contributions of this work are summarized as follows:

1. **A parallel multi-branch multi-scale temporal convolutional network (TCN) encoder for PV time series.** A parallel multi-branch temporal convolutional encoder is developed, where each branch adopts a specific kernel size and dilation rate to capture temporal patterns at different receptive-field scales. The branch features are fused by a 1x1 convolution to obtain a unified representation.

2. **An enhanced temporal attention module with learnable gated fusion.** An enhanced temporal attention module is introduced by integrating multi-kernel temporal filtering and channel-wise reweighting derived from both average pooling and max pooling. A learnable gate is further employed to fuse the enhanced features with the original features in a residual manner, which strengthens informative channels and improves representation quality.

3. **A triplet-based metric embedding network for few-shot PV fault diagnosis.** A triplet-based metric embedding network is established by combining the proposed encoder with a normalized embedding head for distance-based recognition under limited supervision. Systematic ablation and robustness evaluations are provided in the revised manuscript to support the effectiveness of the proposed design.

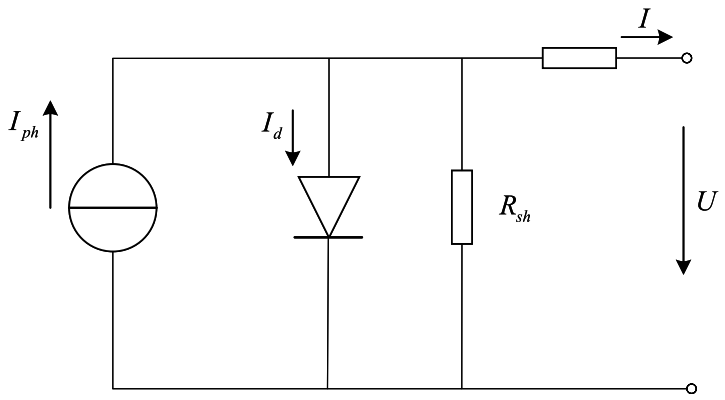


**Figure 1.** Equivalent circuit model of a PV cell.

## 2. PV cell equivalent circuit model

In the literature, the electrical characteristics of PV cells are commonly analyzed and simulated using equivalent-circuit models [22]. Among these models, the single-diode model (SDM) and the double-diode model (DDM) are the two most widely used approaches. Compared with the DDM, the SDM involves fewer parameters and a simpler mathematical formulation, resulting in higher computational efficiency. Therefore, it is often preferred in engineering applications and system-level simulations.

Since this study focuses on engineering applications, the SDM is adopted to describe the electrical behavior of PV cells. As illustrated in figure 1, the SDM provides a good balance between modeling accuracy and computational cost, while maintaining relatively low model complexity.

According to the equivalent circuit model of the PV cell, the output current can be expressed as follows [23]:

$$I = I_{\mathrm{ph}} - I_{\mathrm{d}}\left[\exp\left(\frac{U+IR_{\mathrm{s}}}{nU_{\mathrm{t}}}\right)-1\right]-\frac{U+IR_{\mathrm{s}}}{R_{\mathrm{sh}}} \tag{1}$$

where $I_{\mathrm{ph}}$ is the photocurrent, $I_{\mathrm{d}}$ is the diode reverse saturation current, $R_{\mathrm{s}}$ is the series resistance, $R_{\mathrm{sh}}$ is the shunt resistance, $n$ is the diode ideality factor, and $U_{\mathrm{t}}$ is the thermal voltage.

## 3. PV array fault analysis

### 3.1. Fault data features

Analysis of the characteristic curves of PV arrays under different operating conditions shows that the maximum power point (MPP), the voltage at the MPP, and the current at the MPP vary significantly when faults occur, and different fault types exhibit distinct variation patterns [24]. Moreover, these parameters are strongly affected by external environmental variables such as irradiance and temperature. Therefore, this study selects the open-circuit voltage ($V_{\mathrm{oc}}$), short-circuit current ($I_{\mathrm{sc}}$), voltage at the MPP ($V_{\mathrm{mp}}$), current at the MPP ($I_{\mathrm{mp}}$), output power ($P_{\mathrm{out}}$), array voltage ($V_{\mathrm{a}}$), array current ($I_{\mathrm{a}}$), irradiance ($G$), and temperature ($T$) as fault-related features for PV-array fault diagnosis.

### 3.2. PV array fault simulation and data sources

In real-world environments, irradiance and temperature are uncontrollable [25], and fault injection in operating PV arrays may cause irreversible damage. As a result, it is difficult to collect large amounts of labeled fault data. Since fault samples are essential for PV-array fault diagnosis and model training, their quality and reliability are critical to diagnostic accuracy.

To address this issue, many studies use simulation-based experiments to model PV-array faults and generate validation data [26]. In this work, a PV-array simulation model is developed using simulation software based on the Zytech Engineering Technology ZT185S module under standard test conditions (25°C and 1000 W m$^{-2}$). The electrical characteristics of the module are summarized in table 1, and a schematic of the fault simulation setup is shown in figure 2.

The effects of normal operation and eight fault conditions on the $I$–$V$ curve and the electrical characteristics of PV arrays are investigated. A short-circuit fault is simulated by connecting a very small

**Table 1.** Electrical characteristics of the ZT185S PV module under standard test conditions.

| Parameter | Value |
|---|---|
| Open-circuit voltage $V_{oc}$ (V) | 45 |
| Short-circuit current $I_{sc}$ (A) | 5.3 |
| Voltage at maximum power point $V_{mp}$ (V) | 38.09 |
| Current at maximum power point $I_{mp}$ (A) | 4.87 |

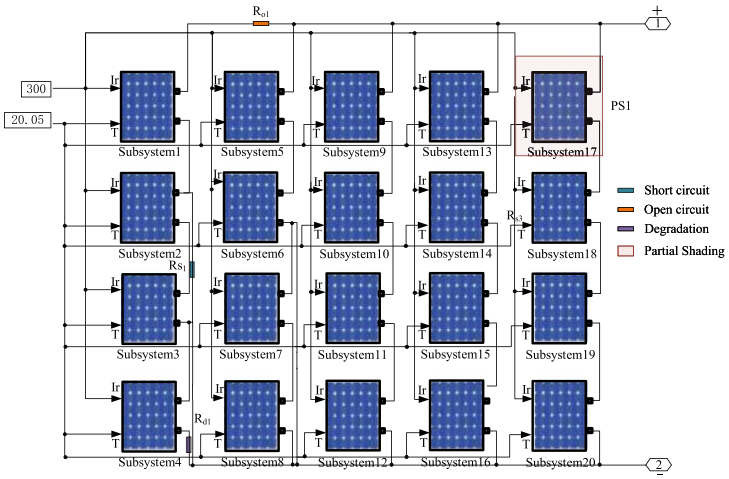


**Figure 2.** Schematic diagram of PV-array fault simulation.

**Table 2.** PV-array fault simulation settings.

| Fault type | Fault description |
|---|---|
| Open-circuit fault | A photovoltaic cell is disconnected (open-circuit failure) |
| Short-circuit fault | A photovoltaic cell is short-circuited |
| Shading fault | A photovoltaic cell is partially shaded |
| Shading and aging fault | A photovoltaic cell is shaded, and two cells experience aging |
| Shading and aging fault (two areas of shading) | Two photovoltaic cells are shaded, and two lines experience aging |
| Aging fault | Two photovoltaic cells experience aging |
| Short-circuit and aging fault | One photovoltaic cell is short-circuited, and two cells experience aging |
| Short-circuit, aging, and shading fault | One photovoltaic cell is short-circuited, two cells experience aging, and two cells are shaded |

resistor $R_{sc}$ ($R_{sc} = 0.001\,\Omega$) across the PV module terminals. An open-circuit fault is modeled by inserting a large series resistor $R_o$ into the circuit. A degradation fault is simulated by adding a small series resistor $R_d$ to the PV array. Partial shading is emulated by applying different irradiance levels to the PV modules. The detailed fault configurations are summarized in table 2.

Among PV arrays, the eight fault scenarios described above represent typical fault conditions. In the simulation, the irradiance is initially set to $300\,W\,m^{-2}$ and the temperature to 20.05 °C. The irradiance and temperature are then increased synchronously in steps of $2\,W\,m^{-2}$ and 0.05 °C, respectively, generating 300 sets of data for each operating condition. As an example, the representative model outputs and corresponding class labels obtained at an irradiance of $600\,W\,m^{-2}$ and a temperature of 27.55 °C are presented in table 3.

After dimensionality reduction, each sequence is segmented using a sliding window of length $L = 100$ with a step size of 1.

## 4. Model methodology

### 4.1. Siamese network

The Siamese network, originally proposed for metric learning and few-shot classification tasks [27], consists of two identical branches with shared weights that extract features from paired input samples.

**Table 3.** Representative model outputs and corresponding class labels.

| Fault type | Label | $V_{oc}$/V | $I_{sc}$/A | $V_{mp}$/V | $I_{mp}$/A | $P_{mp}$/W |
|---|---|---|---|---|---|---|
| Normal | 0 | 130.835 | 9.576 | 111.468 | 8.791 | 979.974 |
| Open-circuit fault | 1 | 130.835 | 6.384 | 111.468 | 5.861 | 653.316 |
| Short-circuit fault | 2 | 94.298 | 9.575 | 78.591 | 8.855 | 695.893 |
| Shading fault | 3 | 130.419 | 9.575 | 112.589 | 7.347 | 827.168 |
| Shading and aging fault | 4 | 130.612 | 9.462 | 97.058 | 5.734 | 556.499 |
| Shading and aging fault (two areas) | 5 | 130.443 | 9.404 | 94.380 | 5.821 | 549.365 |
| Aging fault | 6 | 130.834 | 9.481 | 90.109 | 6.711 | 604.728 |
| Short-circuit and aging fault | 7 | 127.826 | 9.453 | 83.428 | 4.993 | 416.533 |
| Short-circuit, aging, and shading fault | 8 | 126.134 | 9.377 | 77.682 | 4.505 | 349.980 |

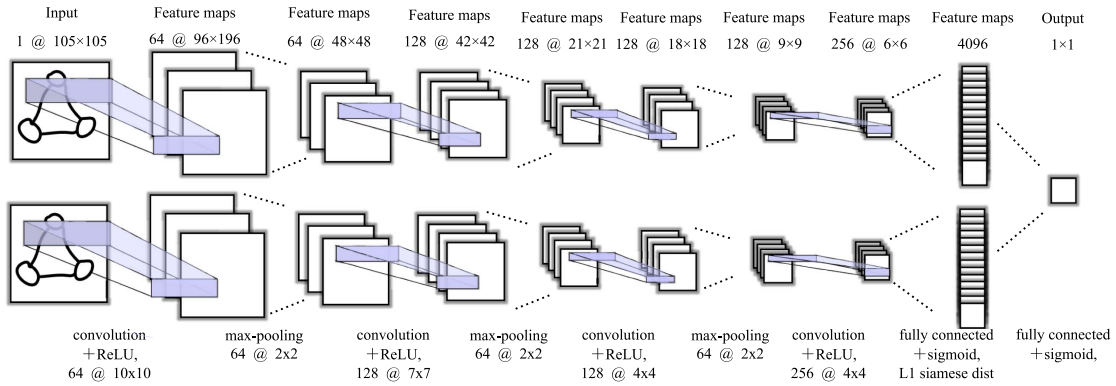


**Figure 3.** Siamese network architecture.

In fault diagnosis, it can distinguish different fault categories by learning the similarity relationships between samples.

Each branch processes one input sample and maps it into a high-dimensional feature representation. A small distance between the two feature vectors indicates that the corresponding samples are likely to belong to the same fault category, whereas a large distance suggests that they belong to different categories. In this study, the Siamese network is used to learn the similarity between PV-array fault samples under different operating conditions, thereby improving diagnostic performance when labeled fault data are limited. This structure enhances classification performance under small-sample conditions and alleviates overfitting, which is common in traditional deep learning methods.

The overall architecture of the Siamese network used in this study is illustrated in figure 3.

### 4.2. Triplet network

In small-sample multi-class fault diagnosis, a conventional Siamese network learns only pairwise similarity or distance, which is often insufficient to model the relative separations among multiple classes. As a result, it may struggle to form clear decision boundaries and may suffer from reduced diagnostic accuracy. In contrast, a triplet network exploits the relative distances among three samples, namely an anchor, a positive sample, and a negative sample, thereby enabling more precise class separation and more stable decisions. Therefore, a triplet network is adopted in this work.

The basic structure of the triplet network is shown in figure 4. As an extension of the Siamese network, it is particularly suitable for few-shot learning and metric-learning tasks. In fault diagnosis, training is performed using an anchor sample ($A$), a positive sample ($P$), and a negative sample ($N$) [28]. The objective is to pull $P$closer to $A$ in the embedding space while pushing $N$ farther away, thereby enhancing discrimination among fault categories.

This behavior is enforced by the triplet loss, which constrains the distances to satisfy $d(A,P) < d(A,N)$ with a margin. Accordingly, the loss is defined as

$$\mathcal{L} = \max\left(d\left(A,P\right) - d\left(A,N\right) + \alpha, 0\right), \tag{2}$$

where $d(A,P)$ and $d(A,N)$ denote the anchor–positive and anchor–negative distances, respectively, and $\alpha$ is a margin hyperparameter. Minimizing $\mathcal{L}$ encourages better separation among different fault categories in the embedding space, thereby improving fault-diagnosis accuracy.

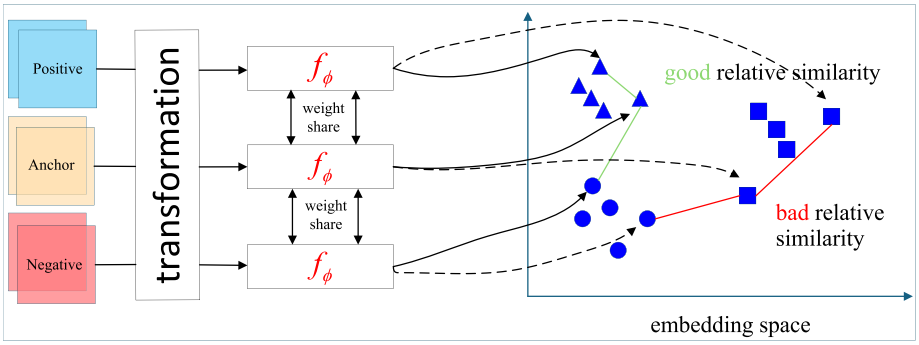


**Figure 4.** Basic structure of a triplet network.

Compared with the Siamese network, the triplet network offers significant advantages in fault-diagnosis tasks. While the Siamese network performs classification by comparing the similarity between two samples, the triplet network explicitly introduces negative samples and thus provides stronger discriminative capability. More specifically, it not only optimizes the similarity between anchor and positive samples, but also establishes clearer decision boundaries among different fault categories by leveraging negative samples. Consequently, even under limited-sample conditions, the triplet network can make more effective use of scarce fault data to learn more discriminative features, thereby improving the robustness of fault recognition. It enables the model to extract deep features from PV-array operating data and to identify different fault patterns more accurately, especially in few-shot learning scenarios with limited labeled data.

### 4.3. TCN

TCN employs causal, often dilated, one-dimensional convolutions to model temporal dependencies while maintaining parallel computation [29]. Its enlarged receptive field enables effective modeling of sequential patterns over temporal windows. In this study, the input fault samples are organized as short temporal sequences of PV-array operating data. Although the simulated irradiance and temperature profiles are generated through discrete, quasi-static step changes, the resulting fault data still contain informative local temporal structure, such as correlations among adjacent samples, transition patterns around operating-condition changes, and short-term persistence within each window. Therefore, the TCN is adopted to capture these local sequential dependencies, providing a more appropriate feature extraction mechanism than methods that treat the samples as unordered static vectors.

**Causal convolution.** Causal convolution enforces temporal causality by ensuring that the output at time $t$ depends only on the current and past inputs, i.e. $y_t = f(x_t, x_{t-1}, \ldots, x_{t-k})$, without accessing future samples [30]. Although this design prevents information leakage, the receptive field of a standard causal convolution is limited. Modeling long-range dependencies therefore requires deeper network stacks, which motivates the introduction of dilated convolution.

**Dilated convolution.** Dilated convolution enlarges the receptive field by inserting gaps between kernel elements, controlled by the dilation factor $d$. When $d = 1$, it reduces to standard convolution; larger values of $d$ allow the convolution to sample the sequence at wider intervals and thus capture longer-range dependencies. As shown in figure 5, a kernel of size 3 with dilation factors $d = (1, 2, 4)$ can effectively expand the receptive field. By combining different dilation factors, the network can capture short-term variations with small $d$ and long-term trends with large $d$ in time-series signals.

**Residual connection.** As the depth of the TCN increases, the receptive field expands rapidly, but optimization may suffer from vanishing or exploding gradients. To improve training stability, TCNs employ residual connections. A typical residual block consists of two stacked causal dilated convolutions, each followed by post-processing modules, such as weight normalization and dropout, for regularization and easier optimization.

In practice, directly adding the input to the block output may cause a channel mismatch. Therefore, a $1\times1$ convolution is applied on the residual path to align the feature dimensions, which facilitates stable gradient propagation and improves convergence. A TCN residual connection can be written as

$$\mathbf{y} = F(\mathbf{x}) + \mathrm{Conv}_{1\times1}(\mathbf{x}), \tag{3}$$

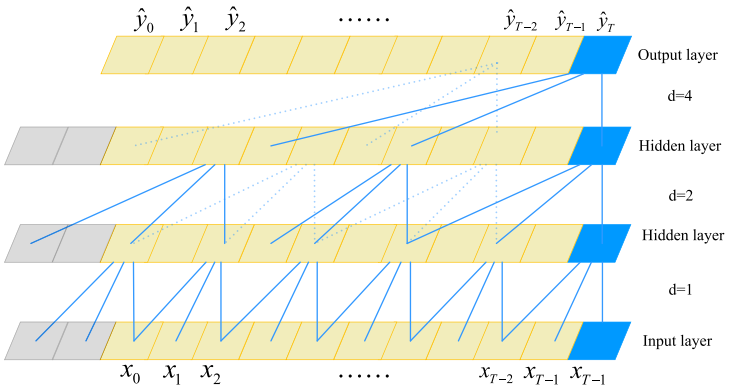


**Figure 5.** Dilated convolution.

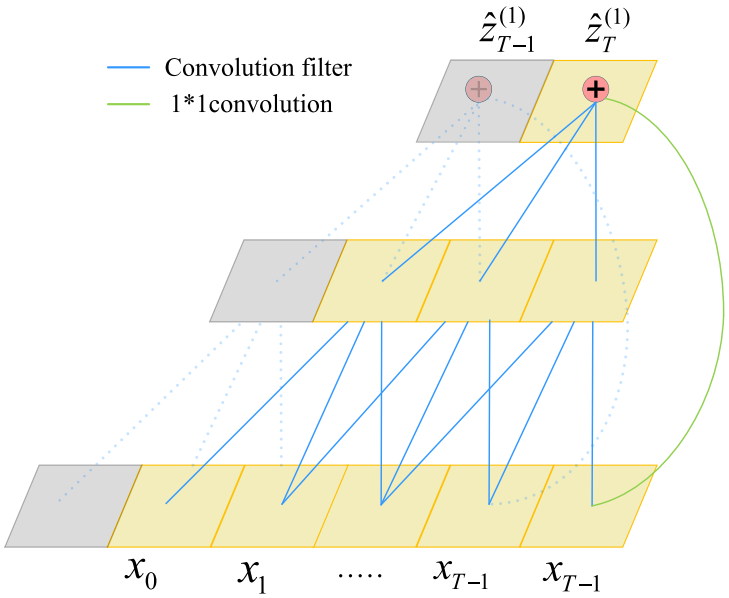


**Figure 6.** Residual connection structure.

where $F(\mathbf{x})$ denotes the stacked causal dilated convolutions with post-processing modules, such as weight normalization and dropout. The residual structure is illustrated in figure 6.

### 4.4. PCA-based dimensionality reduction

PCA is a widely used dimensionality reduction method that projects high-dimensional data into a lower-dimensional space while preserving as much variance as possible [31]. In this study, PCA is applied to the extracted PV-array fault features to reduce redundancy, suppress noise, and improve the efficiency of subsequent model training. The data are first standardized to ensure comparable scales across features. The covariance matrix is then calculated, and its eigenvalues and eigenvectors are used to determine the principal directions of variation. By selecting the top $k$ principal components, the data are projected into a lower-dimensional feature space. The dimensionality-reduced representation is given by:

$$\mathbf{Y} = \mathbf{ZW}, \tag{4}$$

where $\mathbf{Z}$ represents the standardized input data, $\mathbf{W}$ is the projection matrix formed by the eigenvectors of the selected principal components, and $\mathbf{Y}$ is the reduced-dimensional feature matrix.

The purpose of applying PCA is to enhance the training process by reducing redundancy and noise in the input data. This improves the discriminative power of the features under the triplet loss constraint, lowers computational costs, accelerates model convergence, and enhances the model's generalization ability and robustness.

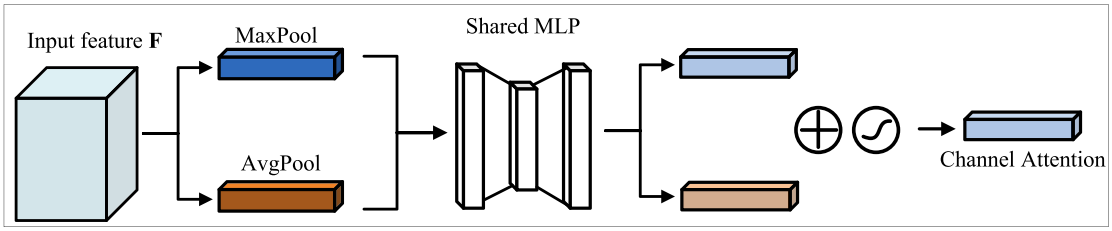


**Figure 7.** Architecture of the channel-attention module.

### 4.5. Channel attention module

In the multi-scale TCN (MSTCN), feature maps produced by convolution kernels of different sizes encode temporal patterns at multiple resolutions. However, the resulting channels are not equally informative: some capture fault-relevant signatures, whereas others may mainly reflect noise or redundant variations. To address this issue, a channel-attention mechanism is introduced to adaptively reweight different channels, thereby enhancing informative responses while suppressing less relevant ones.

Following the common channel-attention paradigm, the module preserves the channel dimension while summarizing global context through global average pooling and max pooling [32]. The pooled descriptors are then fed into a lightweight multilayer perceptron (MLP) to generate channel-wise attention weights, which are applied to the original feature map through multiplicative scaling. Finally, the reweighted features are fused with the original features through a residual connection, allowing the network to balance feature enhancement and information preservation. The architecture of the channel-attention module is illustrated in figure 7.

This design improves the discriminative quality of the learned representations by emphasizing fault-relevant channels while suppressing irrelevant or noisy responses, thereby enhancing diagnostic accuracy and robustness.

## 5. Model methodology

### 5.1. Multi-scale fusion module

To effectively capture multi-scale temporal dynamics in PV array operating data, a MSTCN fusion module is designed as the basic temporal feature extractor in the metric-learning framework. Unlike conventional single-scale convolution, this module employs multiple parallel temporal convolution branches with different kernel sizes and dilation rates, so as to jointly capture short-term fluctuations and long-term dependencies.

As shown in figure 8, the input feature $\mathbf{X}$ is processed by parallel dilated temporal convolution branches to construct representations with different temporal receptive fields. Specifically, the three branches use kernel sizes of 9, 11, and 15, with corresponding dilation rates of 1, 2, and 4, respectively. This design enables the model to capture both local transient variations and long-term temporal dependencies while keeping the parameter scale manageable, thereby producing multi-scale features $\mathbf{X}_0$, $\mathbf{X}_1$, and $\mathbf{X}_2$.

These multi-scale features are then concatenated along the channel dimension and fused through a $1 \times 1$ convolution for channel integration and feature refinement, yielding a unified representation:

$$\mathbf{X}_3 = \mathrm{Conv}_{1\times1}\left(\mathrm{Concat}\left(\mathbf{X}_0, \mathbf{X}_1, \mathbf{X}_2\right)\right), \tag{5}$$

where the $1 \times 1$ convolution enables cross-channel information interaction as well as feature compression and recalibration without altering the temporal resolution, thereby improving the representation efficiency and discriminability of the fused features.

On this basis, a channel-attention mechanism is further introduced to adaptively reweight $\mathbf{X}_3$ at the channel level. Specifically, global average pooling and global max pooling are first applied to obtain complementary channel-wise statistics. The pooled features are then passed through a shared MLP, summed, and activated by a Sigmoid function to generate channel-attention weights:

$$\mathbf{W}_{\mathrm{ch}} = \sigma\left(\mathrm{MLP}\left(\mathrm{AvgPool}\left(\mathbf{X}_3\right)\right) + \mathrm{MLP}\left(\mathrm{MaxPool}\left(\mathbf{X}_3\right)\right)\right), \tag{6}$$

and the attention-enhanced feature representation is obtained as

$$\tilde{\mathbf{X}}_3 = \mathbf{W}_{\mathrm{ch}} \odot \mathbf{X}_3, \tag{7}$$

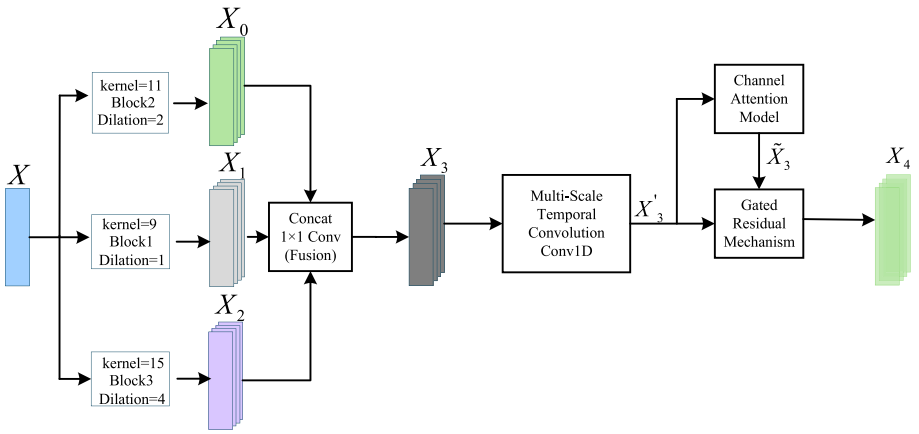


**Figure 8.** Multi-scale TCN-based fusion module.

where $\odot$ denotes element-wise multiplication. This mechanism emphasizes channel responses that are more critical for fault discrimination while suppressing redundant channel information, thereby providing more informative and robust temporal features for subsequent discriminative learning.

Finally, a gated residual fusion mechanism is introduced to adaptively combine the attention-enhanced features with the original temporal features, yielding the final output:

$$\mathbf{X}_4 = g\tilde{\mathbf{X}}_3 + (1-g)\,\mathbf{X}_3, \tag{8}$$

where $g$ is a learnable gate coefficient. This mechanism adaptively balances the contributions of the enhanced feature representation and the residual feature information. In this way, the model can further emphasize discriminative channel responses while preserving the underlying temporal structure of the original features, thereby improving the robustness and discriminative capability of the learned representation.

### 5.2. Overall framework and training mechanism

Let the original multivariate time series be denoted as $X \in \mathbb{R}^{T_0 \times D}$, where $T_0$ is the sequence length and $D$ is the feature dimension. First, the raw data are standardized, and PCA is applied to reduce the feature dimension to $D'$. Then, a sliding window of length $L$ is used to segment the reduced sequence into window samples $X_i \in \mathbb{R}^{L \times D'}$. To obtain discriminative representations for few-shot classification, an embedding mapping $f_\theta : \mathbb{R}^{L \times D'} \to \mathbb{R}^d$ is constructed so that samples from the same class are closer, while samples from different classes are farther apart in the embedding space.

The input samples are then fed into three parallel TCN branches with different kernel sizes and dilation-rate combinations. Each branch consists of residual TCN blocks. The outputs of the three branches are concatenated along the channel dimension, followed by a $1 \times 1$ convolution for feature fusion and channel compression. After that, multi-scale convolutions are employed to extract contextual information at different temporal scales, and a channel attention mechanism is introduced to adaptively enhance critical features. Finally, a learnable gated residual connection is used to combine the enhanced features with the original fused features, followed by activation, dropout, and normalization to obtain the final embedding representation $f_\theta(X_i)$.

During training, triplet samples $(X_\mathrm{a}, X_\mathrm{p}, X_\mathrm{n})$ are constructed, where $X_\mathrm{a}$ is the anchor sample, $X_\mathrm{p}$ is a positive sample from the same class, and $X_\mathrm{n}$ is a negative sample from a different class. In the embedding space, the distances between the anchor and the positive/negative samples are defined as

$$d_{\mathrm{ap}} = \left\| f_\theta\left(X_\mathrm{a}\right) - f_\theta\left(X_\mathrm{p}\right) \right\|_2 \tag{9}$$

$$d_{\mathrm{an}} = \left\| f_\theta\left(X_\mathrm{a}\right) - f_\theta\left(X_\mathrm{n}\right) \right\|_2 \tag{10}$$

The network parameters are optimized under triplet constraints to reduce the intra-class distance and enlarge the inter-class distance, thereby improving the discriminability of the learned embeddings. The training objective is optimized using the triplet loss. During inference, the same Euclidean distance metric in the embedding space is used to measure the similarity between the query sample and each support sample for nearest-neighbor diagnosis. The model architecture diagram is shown in figure 9.

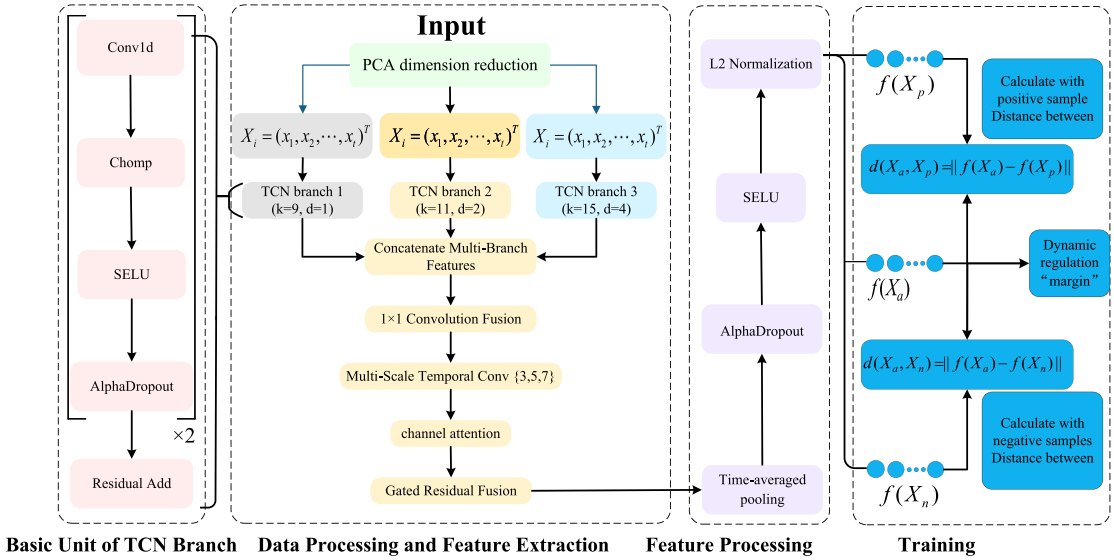


**Figure 9.** Model architecture diagram.

## 6. Data processing and diagnostic procedure

### 6.1. Data preprocessing

This study focuses on small-sample fault diagnosis under simulated PV-array operating conditions. For each operating condition, time-series samples are generated. To reduce the influence of scale differences among features and improve training stability, $Z$-score normalization is applied to the input data. This transformation converts each feature into a distribution with zero mean and unit variance. The normalization formula is given by

$$z = \frac{x - \mu}{\sigma}, \tag{11}$$

where $x$ denotes the original data, $\mu$ is the mean, $\sigma$ is the standard deviation, and $z$ is the normalized value.

### 6.2. Few-shot triplet construction

To adapt to the few-shot fault-diagnosis task, the original time-series samples are organized by category. Suppose the dataset contains $m$ classes, with $n$ samples in each class, resulting in a total of $m \times n$ samples. In the few-shot setting, $k \in \{5, 10, 15, 20, 25\}$ samples are randomly selected from each class to form the training set, while the remaining samples serve as a candidate pool for testing. Therefore, the number of training samples is given by

$$N_{\text{train}} = m \times k \in \{45, 90, 135, 180, 225\}, \tag{12}$$

where $m = 9$ denotes the number of classes. In addition, 150 samples are randomly selected from the remaining samples to form the test set. Based on the constructed training set, triplet constraints are generated under the triplet metric-learning framework. For each anchor sample, 8 triplets are randomly constructed by pairing the anchor with sampled positive and negative instances. Thus, the total number of generated triplets is:

$$N_{\text{triplets}} = 8 \times N_{\text{train}}. \tag{13}$$

In this study, triplet subsets of 200, 400, 600, 800, and 1000 are constructed for experimental analysis. Compared with full combinational construction, the random sampling strategy effectively reduces the number of triplets and the computational cost while preserving the supervisory information of intra-class similarity and inter-class separability for subsequent embedding representation learning.

**Table 4.** Overall diagnostic procedure of the proposed model.

| Fault diagnosis based on shared embedding and nearest neighbor |
|---|
| Input: class-wise labeled sample pool $\{W_c\}_{c=1}^{C}$, query set $Q$, query labels $Y_q$, embedding network $f_\theta(\cdot)$, and support size $K$ |
| Initialize support set $S \leftarrow \varnothing$ and label set $Y_s \leftarrow \varnothing$ |
| **for** $c = 1$ **to** $C$ **do** |
| Randomly select $K$ samples from $W_c$ and add them to $S$ |
| Record their labels in $Y_s$ |
| **end for** |
| Compute support embeddings $\{f_\theta(x_i^{(s)})\}_{i=1}^{N_s}$ |
| Initialize prediction set $\hat{Y} \leftarrow \varnothing$ |
| **for** each query sample $x_q^j \in Q$ **do** |
| Compute query embedding $f_\theta(x_q^j)$ |
| Compute distances $d(x_q^j, x_i^{(s)})$ to all support samples using the same metric as defined in equations (9) and (10) |
| Find the nearest support index $i^* = \arg\min_i d(x_q^j, x_i^{(s)})$ |
| Assign prediction $\hat{y}_q^j = y_{i^*}^{(s)}$ |
| Append $\hat{y}_q^j$ to $\hat{Y}$ |
| **end for** |
| Compute accuracy $Acc = \mathrm{Accuracy}(Y_q, \hat{Y})$ |
| Output: $\hat{Y}$ and $Acc$ |

### 6.3. Overall diagnostic procedure

The testing-stage fault-diagnosis process is summarized in table 4. In accordance with the few-shot setting, a small number of labeled samples from each class are selected to form the support set, and each sample to be identified is treated as a query sample. The support set and the query sample are both mapped into a shared embedding space, and the Euclidean distances between the query sample and all support samples are computed using the same distance metric defined in equations (9) and (10). The class label of the nearest support sample is then assigned to the query sample as the diagnostic result.

## 7. Experimental settings and hyperparameters

To facilitate reproducibility and provide a clear description of the proposed implementation, the training hyperparameters are summarized in table 5. The detailed network configuration of the proposed Tri-MSTCN is reported in tables 6 and 7, including the MSTCN branches with the fusion layer, as well as the attention module and the output head.

Table 5 summarizes the default training settings used in the experiments. Tables 6 and 7 present the architecture and key configurations of the proposed Tri-MSTCN. Unless otherwise specified, these settings are used throughout the paper. For fair comparison, all methods follow the same data preprocessing and evaluation protocol.

## 8. Experimental results and analysis

### 8.1. Ablation study

To quantify the contribution of each key component, component-wise ablation experiments were conducted by removing one component at a time and comparing the modified model with the full model. The experimental settings are summarized in table 8. In Group C2 ('No Multi-scale'), the proposed parallel multi-scale branches are replaced with a conventional single-path TCN composed of three sequential temporal convolution layers with (kernel = 9, dilation = 1), (kernel = 11, dilation = 2), and (kernel = 15, dilation = 4), respectively. In Group C3, the triplet-learning framework is replaced with a Siamese formulation using standard pairwise contrastive loss under the same MSTCN backbone, in order to isolate the contribution of triplet-based metric learning.

All ablation variants are implemented under the same data splitting, sliding-window configuration, and training hyperparameters to ensure a fair comparison. Figure 10 further illustrates the evolution of the positive- and negative-pair distances of the full model and its ablation variants during training.

**Table 5.** Training settings and hyperparameters.

| Hyperparameter | Setting |
|---|---|
| Optimizer | Adam |
| Learning rate | $1 \times 10^{-4}$ |
| Weight decay | $1 \times 10^{-5}$ |
| Batch size (train / test) | 32 / 16 |
| Epochs | 100 |
| Activation function | SELU |
| Dropout | AlphaDropout ($p = 0.1$) |
| Initial margin $\alpha$ | 0.2 |
| Margin range | $[0.1,\ 1.5]$ |

**Table 6.** Architecture of the proposed Tri-MSTCN: multi-scale TCN branches and feature fusion.

| Stage | Layer type | Kernel size | Dilation | Output channels |
|---|---|---|---|---|
| Branch 1 | 2× Conv1D (weight_norm) | $9 \times 1$ | 1 | 64 |
| Branch 1 | $1 \times 1$ Conv (residual) | $1 \times 1$ | 1 | 64 |
| Branch 2 | 2× Conv1D (weight_norm) | $11 \times 1$ | 2 | 64 |
| Branch 2 | $1 \times 1$ Conv (residual) | $1 \times 1$ | 1 | 64 |
| Branch 3 | 2× Conv1D (weight_norm) | $15 \times 1$ | 4 | 64 |
| Branch 3 | $1 \times 1$ Conv (residual) | $1 \times 1$ | 1 | 64 |
| Fusion | Concatenation (3 branches) | — | — | 192 |
| Fusion | $1 \times 1$ Conv | $1 \times 1$ | 1 | 64 |

**Table 7.** Architecture of the proposed Tri-MSTCN: attention module and output head.

| Stage | Module type | Key parameters | Output channels |
|---|---|---|---|
| Attn-1 | Multi-scale temporal Conv1D | Kernels = {3,5,7} | 64 |
| Attn-2 | Channel attention | AvgPool+MaxPool, reduction = 4 | 64 |
| Attn-3 | Gated residual | Learnable gate (init = 0.5) | 64 |
| Head-1 | Global average pooling | mean over time | 64 |
| Head-2 | Nonlinearity + dropout | SELU + AlphaDropout ($p = 0.1$) | 64 |
| Head-3 | Embedding normalization | $L_2$ normalization ($p = 2$) | 64 |

**Table 8.** Ablation experiment configuration.

| Group | PCA | Multi-scale structure | Channel attention mechanism | Triplet structure |
|---|---|---|---|---|
| C0 | ✓ | ✓ | ✓ | ✓ |
| C1 | × | ✓ | ✓ | ✓ |
| C2 | ✓ | × | ✓ | ✓ |
| C3 | ✓ | ✓ | × | ✓ |
| C4 | ✓ | ✓ | ✓ | × |

As shown in figure 10, the full model produces the most discriminative embedding during training: the positive-pair distance decreases faster and converges to a relatively low level, while the negative-pair distance remains at a relatively high level. Consequently, the full model maintains a large separation between positive and negative distances across epochs, indicating stronger intra-class compactness, inter-class separability, and a more stable convergence trend. In contrast, the no PCA variant exhibits a slower decrease and a higher converged positive distance, suggesting weaker feature aggregation. Both the no multi-scale and no channel attention variants can separate positive and negative pairs to some extent, but they converge with a higher positive distance and a smaller separation gap, implying that the multi-scale architecture and the channel-attention mechanism jointly contribute to a more discriminative embedding space. By contrast, the no Triplet structure variant shows the weakest discrimination ability, with the highest positive-pair distance and the smallest gap between positive and negative distances throughout training. This suggests that replacing the triplet-learning framework with a standard pairwise formulation weakens the relative distance constraints in the embedding space, thereby reducing intra-class compactness and inter-class separability.

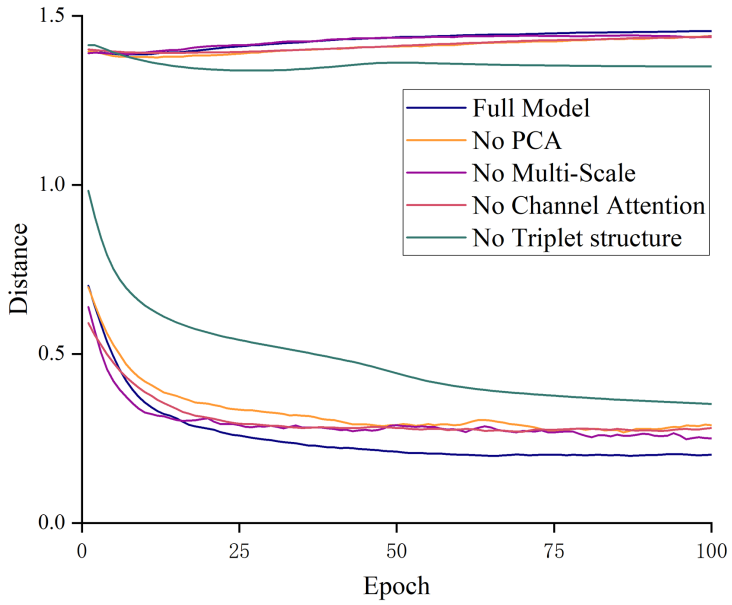


**Figure 10.** Positive- and negative-pair distances under ablation settings.

**Table 9.** Ablation results over 10 repeated evaluations (mean ± SD).

| Setting | Acc (%) | Macro-P (%) | Macro-R (%) | Macro-*F*1 (%) |
|---|---|---|---|---|
| C0 | 98.89 ± 0.11 | 98.95 ± 0.13 | 98.89 ± 0.11 | 98.88 ± 0.09 |
| C1 | 98.19 ± 0.38 | 98.25 ± 0.34 | 98.19 ± 0.38 | 98.19 ± 0.38 |
| C2 | 94.73 ± 0.70 | 95.42 ± 1.23 | 94.73 ± 0.70 | 94.66 ± 0.44 |
| C3 | 97.97 ± 0.61 | 98.27 ± 0.20 | 97.97 ± 0.61 | 97.98 ± 0.59 |
| C4 | 98.70 ± 2.78 | 99.02 ± 2.08 | 98.70 ± 2.78 | 98.60 ± 3.04 |

After the component-wise ablation analysis, further comparisons are conducted using the evaluation metrics of accuracy, precision, recall, and $F_1$ score. The quantitative results are summarized in table 9.

The results in table 9 show that the full model (C0) achieves the highest accuracy (98.89%) and stability. C1 shows a slight performance drop, with accuracy at 98.19%, indicating that PCA contributes to feature refinement and stable representation learning. C2 exhibits the most significant decline, with accuracy dropping to 94.73%, highlighting the importance of multi-scale feature extraction. C3 shows a moderate decrease in accuracy (97.97%), suggesting that the channel-attention mechanism provides an additional contribution to discriminative representation learning. By contrast, although C4 maintains a relatively high mean accuracy (98.70%), its substantially larger standard deviation indicates reduced stability across repeated evaluations, suggesting that the triplet-learning structure contributes not only to discriminative performance but also to more robust and consistent embedding learning.

### 8.2. Comparison with baseline methods

To evaluate the performance of the proposed Tri-MSTCN in small-sample fault diagnosis, it is compared with three widely used neural-network models, namely GRU [34], TCN [35], and CNN-BiGRU (CBG) [36], as well as representative metric-learning and few-shot learning baselines. The metric-learning baselines include SCBSN [21] and MSSASN [33], while the few-shot learning baselines include DVPN [37].

Following the data-partitioning scheme described in the previous section, each model is trained for 100 epochs under different training-sample settings and then evaluated on a test set of 1350 samples. The corresponding fault-diagnosis accuracies are reported in table 10.

As shown in table 10, the diagnostic accuracies of all compared models generally improve as the number of training samples increases, and the proposed Tri-MSTCN achieves the best performance under all training-sample settings. When the total number of training samples is 45, the proposed model already reaches an accuracy of 96.67%. When the number of training samples increases to 225, its accuracy further improves to 100.00%, indicating strong diagnostic capability in few-shot scenarios. Among the traditional deep-learning models, TCN outperforms GRU and CBG. Although SCBSN and MSSASN achieve better overall performance than the traditional models, they remain inferior to the proposed

**Table 10.** Results under different numbers of training samples.

| Training samples | GRU | CBG | TCN | SCBSN | MSSASN | DVPN | Tri-MSTCN |
|---|---|---|---|---|---|---|---|
| 45 | — | — | — | 92.83 | 91.89 | 94.95 | 96.67 |
| 90 | — | — | — | 94.01 | 93.07 | 97.58 | 97.38 |
| 135 | 93.04 | 95.04 | 95.36 | 96.45 | 95.51 | 98.20 | 98.89 |
| 180 | 94.21 | 97.30 | 96.93 | 98.90 | 96.78 | 99.51 | 100.00 |
| 225 | 97.93 | 98.07 | 98.78 | 99.21 | 97.93 | 99.76 | 100.00 |

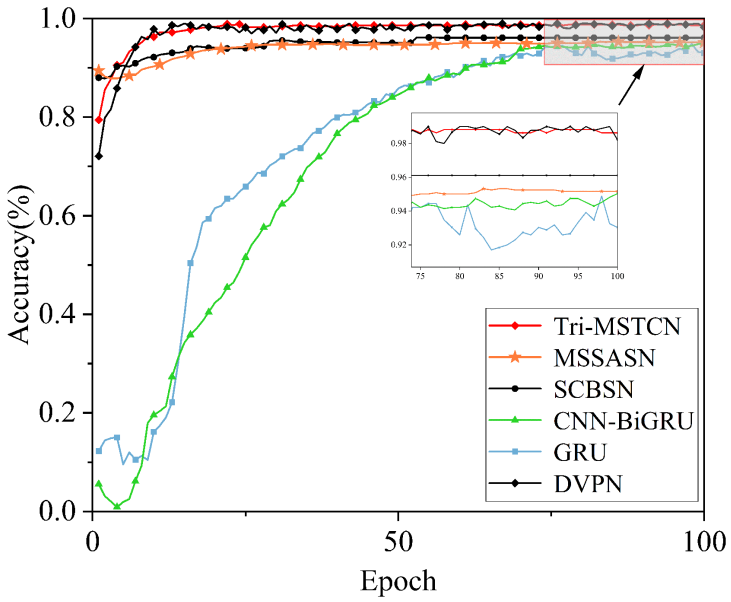


**Figure 11.** Test accuracy curves of different models.

model. DVPN performs better than SCBSN, MSSASN, and the traditional models, but it is still consistently outperformed by Tri-MSTCN across all training-sample settings. This performance gap is particularly evident under extremely low-sample conditions. Overall, the proposed model demonstrates clear performance advantages across different sample sizes.

To further validate the superiority of the proposed Tri-MSTCN, additional comparative experiments are conducted under the same small-sample setting. Four representative baseline models, namely GRU, CNN-BiGRU, SCBSN, and MSSASN, are selected for comparison. All models are trained and tested with 15 samples per fault category under identical conditions. The corresponding test-accuracy curves are shown in figure 11, which provide further evidence for the comparative performance analysis.

As shown in figure 11, the proposed Tri-MSTCN achieves the best overall performance. Its accuracy increases rapidly in the early stage and quickly converges to a high and stable level close to 1.0, demonstrating superior learning efficiency and generalization capability.

By contrast, GRU and CNN-BiGRU converge more slowly and yield lower final accuracy. SCBSN exhibits a relatively smooth improvement during training, but its final performance remains inferior to that of Tri-MSTCN. MSSASN performs better than GRU and CNN-BiGRU under the small-sample setting, yet its final accuracy and stability are still slightly lower than those of Tri-MSTCN. Overall, Tri-MSTCN shows clear advantages in convergence speed, steady-state accuracy, and stability.

Although accuracy provides an overall measure of model performance, it cannot fully reflect the class-wise classification behavior of different models. Therefore, confusion matrices are further introduced under the setting of 15 samples per class to analyze the recognition and misclassification patterns of different models in detail. The corresponding comparison results are shown in figure 12.

As shown in figure 12, all models classify classes 0–3 and 6–8 almost perfectly, whereas classes 4 and 5 remain the main source of error. The proposed Tri-MSTCN achieves the highest overall accuracy and the best recognition rate for class 4 (98.68%), with only 1.32% of samples misclassified as class 5. Its performance on class 5 is also competitive (86.09%). Among the baseline models, DVPN is the strongest competitor, reaching 98.20% overall accuracy and delivering strong recognition for both class 4 (87.3%) and class 5 (96.0%), although its overall performance is still slightly inferior to that of

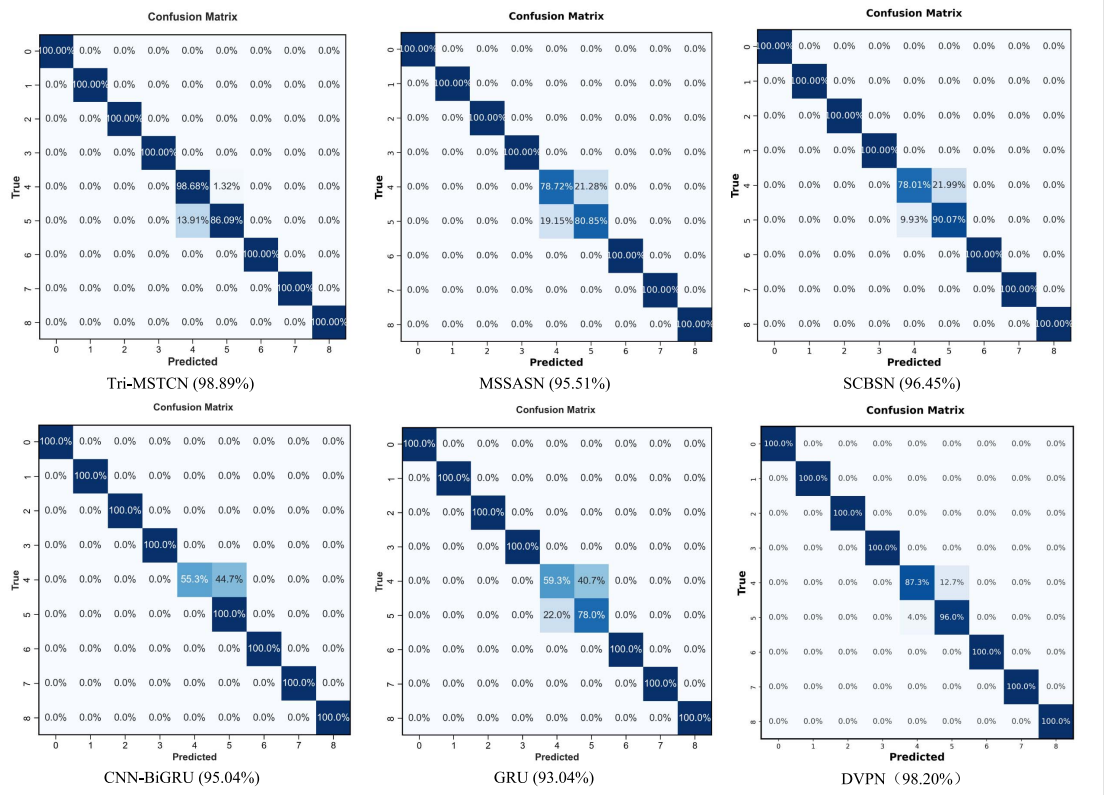


**Figure 12.** Comparison of confusion matrices for different models under the 15-shot setting.

**Table 11.** Performance comparison over 10 repeated evaluations (mean ± SD).

| Model | Acc (%) | Macro-P (%) | Macro-R (%) | Macro-*F*1 (%) |
|---|---|---|---|---|
| GRU | 92.74 ± 0.20 | 92.74 ± 0.20 | 92.74 ± 0.20 | 92.74 ± 0.20 |
| CBG | 94.69 ± 0.11 | 95.64 ± 0.06 | 94.69 ± 0.11 | 94.58 ± 0.13 |
| TCN | 95.38 ± 0.18 | 95.75 ± 0.13 | 95.38 ± 0.18 | 95.36 ± 0.19 |
| SCBSN | 96.45 ± 0.56 | 96.53 ± 0.53 | 95.45 ± 0.56 | 95.44 ± 0.56 |
| MSSASN | 95.00 ± 0.50 | 95.35 ± 0.51 | 95.00 ± 0.50 | 94.97 ± 0.51 |
| DVPN | 98.14 ± 0.22 | 98.15 ± 0.22 | 98.14 ± 0.22 | 98.14 ± 0.23 |

Tri-MSTCN. MSSASN and SCBSN outperform GRU and CNN-BiGRU, but both still exhibit noticeable mutual confusion between classes 4 and 5. By contrast, GRU and CNN-BiGRU show weaker discrimination for these difficult classes, with substantial cross-class misclassification. Overall, the results indicate that Tri-MSTCN provides the most robust and balanced performance, especially for hard-to-distinguish fault categories.

Using the same repeated-evaluation protocol as in the ablation study, we further compare Tri-MSTCN with representative baseline methods. Table 11 reports Acc, Macro-P, Macro-R, and Macro-*F*1 in terms of mean ± SD over 10 runs. To further evaluate the overall performance and stability of different methods, table 11 presents the average results of 10 repeated experiments.

This observation is further confirmed by the quantitative results in tables 9 and 11. Over 10 repeated runs, Tri-MSTCN achieves 98.89 ± 0.11% accuracy, 98.95 ± 0.13% Macro-P, 98.89 ± 0.11% Macro-R, and 98.88 ± 0.09% Macro-*F*1, all of which are higher than those of the competing methods listed in table 11. Compared with the strongest baseline, DVPN, Tiri-MSTCN consistently delivers better results across all metrics while maintaining relatively small standard deviations, demonstrating superior diagnostic accuracy, stability, and robustness.

### 8.3. Robustness analysis under different noise conditions

Although the previous experiments were conducted under relatively ideal conditions, practical industrial environments are often affected by background noise. To assess the robustness of the proposed

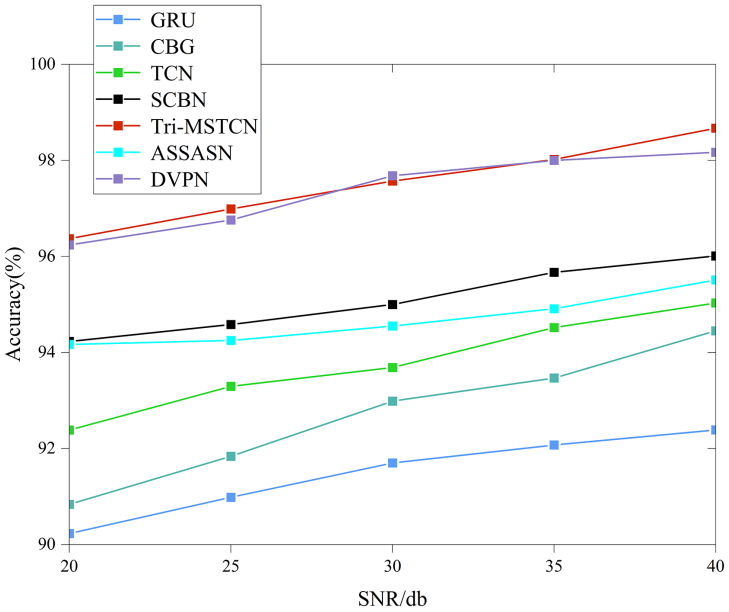


**Figure 13.** Accuracy comparison of different models under different noise conditions.

Tri-MSTCN, Gaussian white noise is added to the signals with SNRs ranging from 40 dB to 20 dB. The diagnostic accuracies of the compared models under different noise levels are shown in figure 13.

As shown in figure 13, the accuracy of all models increases as the SNR rises from 20 dB to 40 dB, indicating improved diagnostic performance under lower noise levels. The proposed Tri-MSTCN consistently achieves high accuracy across the entire SNR range, demonstrating strong robustness to noise. SCBSN and MSSASN rank second and maintain relatively good performance under noisy conditions, whereas TCN and CNN-BiGRU show only moderate improvement. By contrast, GRU performs the worst at all SNR levels, revealing limited robustness to noise. DVPN also shows relatively stable performance over the tested SNR range and remains competitive, especially at 30 dB and 35 dB. Overall, the proposed Tri-MSTCN exhibits the most stable and reliable classification performance under different noise conditions.

To further illustrate the differences in classification behavior under limited-sample and high-noise conditions, a comparative analysis is conducted based on the t-SNE scatter plots of the classification results under the setting of 15 samples per class with 20 dB noise. The corresponding visualization results are shown in figure 14.

As shown in figure 14, t-SNE is used to visualize the feature distributions of GRU, CNN-BiGRU, TCN, SCBSN, MSSASN, DVPN, and the proposed Tri-MSTCN. The results indicate noticeable differences in feature discrimination among the compared models. The proposed Tri-MSTCN shows relatively compact intra-class clustering and clearer inter-class separation, with fewer misclassified samples, suggesting strong noise robustness and stable feature representations. MSSASN also performs relatively well and exhibits improved cluster compactness and reduced overlap compared with TCN and CNN-BiGRU, although some confusion remains between class 4 and class 5. DVPN likewise presents comparatively clear class boundaries and competitive clustering behavior, but its feature separability is still slightly weaker than that of Tri-MSTCN. SCBSN and TCN show a moderate degree of separability, whereas CNN-BiGRU and GRU exhibit more evident feature entanglement and misclassification. Overall, these results suggest that the proposed Tri-MSTCN provides strong robustness, feature separability, and classification reliability under high-noise and small-sample conditions.

## 9. Discussion and future work

To assess the practical applicability of the proposed method, we report its computational complexity and briefly discuss the limitations of simulation-only evaluation, as well as future deployment directions.

Table 12 presents the computational complexity of the proposed method. It has 247 505 parameters (0.94 MB in FP32), and an average inference latency of 2.505 ms per sample at batch size 1 with

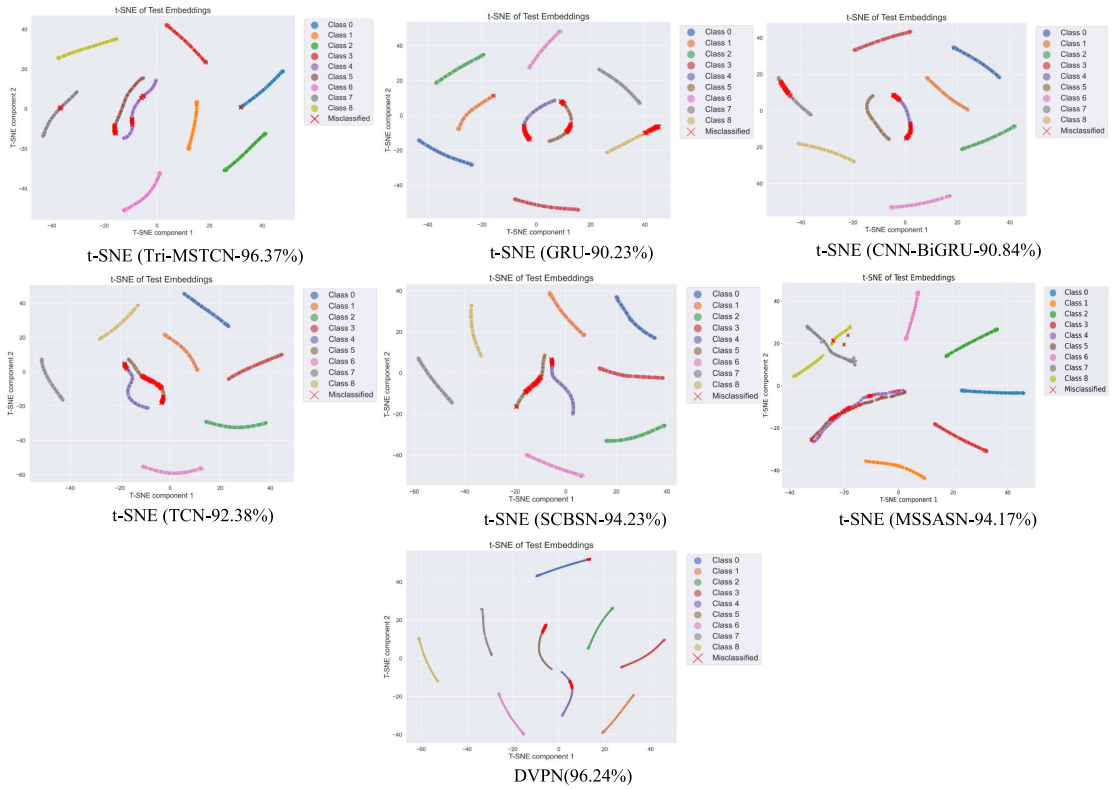


**Figure 14.** t-SNE comparison of different models under 20 dB noise.

**Table 12.** Model complexity report of the proposed method.

| Metric | Value |
|---|---|
| Total parameters | 247 505 |
| Trainable parameters | 247 505 |
| Model size (FP32 params only) | 0.94 MB |
| Inference time (batch = 1, input 6 × 100) | 2.505 ms |

an input size of 6 × 100. These results indicate that the model is compact and efficient, with millisecond-level inference time, making it well-suited for edge deployment.

The current evaluation is based on simulated PV data using a small-sample protocol, which simulates the scarcity of labeled measured fault data. However, it cannot replace a rigorous Sim-to-Real validation with real-world datasets. In practical PV plants, additional challenges may arise, such as sensor drift, missing data, environmental variability, operating-condition changes, and evolving fault signatures.

Despite these limitations, robustness experiments under noise and perturbations suggest that the learned representations are effective under certain controlled distribution shifts. Therefore, the reported results should be seen as evidence obtained under controlled simulation conditions rather than a guarantee of performance in real-world deployments. Future work will incorporate measured PV-station data and explore simulation-to-real transfer, including few-shot adaptation, to improve generalization to real-world scenarios.

## 10. Conclusion

This paper proposes a fault-diagnosis method for PV arrays under small-sample conditions based on the Tri-MSTCN. The proposed framework integrates multi-scale temporal convolution, channel attention, and gated residual fusion to learn discriminative representations from limited labeled data. By introducing triplet-based metric learning, the model is able to enhance intra-class compactness and inter-class separability in the embedding space, thereby improving fault-class discrimination under data-scarce conditions.

The experimental results show that Tri-MSTCN outperforms the compared baseline models under different training-sample settings and exhibits clear advantages in few-shot fault diagnosis. Ablation studies confirm the positive contributions of the main modules in the proposed architecture. Moreover, experiments under different noise levels indicate that the proposed model maintains stable diagnostic performance and strong robustness in noisy environments.

Overall, the proposed Tri-MSTCN provides an effective and robust solution for PV-array fault diagnosis under small-sample conditions. Future work will investigate semi-supervised and self-supervised learning methods to make better use of unlabeled operational data and to improve the recognition of unknown or emerging fault patterns in practical PV systems.

## Acknowledgments

This work is supported by the National Natural Science Foundation of China (61563032, 61963025, 62203196, 62463018,62473167), the Gansu Provincial Science and Technology Program (22YF7GA164, 22CX8GA131, 22CX3GD005) and he Natural Science Foundation of Jilin Province, China under Grant No. 20240402079GH.

## Data availability statement

The data cannot be made publicly available upon publication because no suitable repository exists for hosting data in this field of study. The data that support the findings of this study are available upon reasonable request from the authors.

## ORCID iD

Yanbo Jian 0009-0006-6512-1518

## References

[1] Hossain M D S, Wadi Al-Fatlawi A, Kumar L, Yan Y R and Assad M E-H 2024 Solar PV high-penetration scenario: an overview of the global PV power status and future growth *Energy Syst.* **17** 809–65
[2] Lei Z, Zhang P and Chen Y *et al* 2023 Prior knowledge-embedded meta-transfer learning for few-shot fault diagnosis under variable operating conditions *Mech. Syst. Signal Process.* **200** 110491
[3] Wu J, Zhao Z, Sun C and Xiong M 2020 Few-shot transfer learning for intelligent fault diagnosis of machine *Measurement* **166** 108202
[4] Che C, Zhang Y, Wang H and Xiong M 2024 Interpretable multi-domain meta-transfer learning for few-shot fault diagnosis of rolling bearing under variable working conditions *Meas. Sci. Technol.* **35** 076103
[5] Liu J and Dai W 2023 Constructive incremental learning for fault diagnosis of rolling bearings with ensemble domain adaptation (arXiv:2308.14983)
[6] Lin C, Kong Y, Han Q, Wang T, Dong M, Liu H and Chu F 2024 An information fusion-based meta transfer learning method for few-shot fault diagnosis under varying operating conditions *Mech. Syst. Signal Process.* **220** 111652
[7] Li X, Zhang W, Ding Q and Sun J-Q 2020 Intelligent rotating machinery fault diagnosis based on deep learning using data augmentation *J. Intell. Manuf.* **31** 433–52
[8] Li W, Zhong X, Shao H, Cai B and yang X 2022 Multi-mode data augmentation and fault diagnosis of rotating machinery using modified ACGAN designed with new framework *Adv. Eng. Inf.* **52** 101552
[9] Tian J, Jiang Y, Zhang J, Luo H and Shen Y 2024 A novel data augmentation approach to fault diagnosis with class-imbalance problem *Reliab. Eng. Syst. Saf.* **243** 109832
[10] Fan C, Zhang Y, Ma H, Yu K and Ma Z 2025 A novel lightweight DDPM-based data augmentation method for rotating machinery fault diagnosis with small sample *Mech. Syst. Signal Process.* **232** 112741
[11] Yang X, Ye T, Yuan X, Zhu W, Mei X and Zhou F 2024 A novel data augmentation method based on denoising diffusion probabilistic model for fault diagnosis under imbalanced data *IEEE Trans. Ind. Inf.* **20** 7820–31
[12] Fu Q and Wang H 2020 A novel deep learning system with data augmentation for machine fault diagnosis from vibration signals *Appl. Sci.* **10** 5765
[13] Yu S, Li Z, Gu J, Wang R, Liu X, Li L, Guo F and Ren H 2025 CWMS-GAN: a small-sample bearing fault diagnosis method based on continuous wavelet transform and multi-size kernel attention mechanism *PLoS One* **20** e0319202
[14] Sun S, Han S, Luan Z, Qin X, Yin J, Zhao Z, Cao J and Wang H 2025 An advanced convolutional neural network for bearing fault diagnosis under limited data (arXiv:2509.11053)
[15] Finn C, Abbeel P and Levine S 2017 Model-agnostic meta-learning for fast adaptation of deep networks *Int. Conf. on Machine Learning PMLR* pp 1126–35
[16] Li Z, Zhou F, Chen F and Li H 2017 Meta-SGD: learning to learn quickly for few-shot learning (arXiv:1707.09835)
[17] Nichol A, Achiam J and Schulman J 2018 Reptile: a scalable metalearning algorithm 2(3) 4 (arXiv:1803.02999)

[18] Raghu A, Raghu M, Bengio S and Vinyals O 2019 Rapid learning or feature reuse? towards understanding the effectiveness of maml (arXiv:1909.09157)
[19] Li C, Li S, Zhang A, He Q, Liao Z and Hu J 2021 Meta-learning for few-shot bearing fault diagnosis under complex working conditions *Neurocomputing* **439** 197–211
[20] Lin J, Shao H, Zhou X, Cai B and Liu B 2023 Generalized MAML for few-shot cross-domain fault diagnosis of bearing driven by heterogeneous signals *Expert Syst. Appl.* **230** 120696
[21] Zhao Z, Wu D and Dou G 2023 Bearing fault diagnosis method based on a CNN-BiGRU Siamese network *J. Vib. Shock* **42** 166–71 (available at: https://jvs.sjtu.edu.cn/EN/Y2023/V42/I6/166)
[22] Gobichettipalayam Shanmugam S K, Sakthivel T S, Gaftar B A, Iyappan P and Sathyamurthy R 2022 Modeling and simulation of single-and double-diode PV solar cell model for renewable energy power solution *Environ. Sci. Pollution Res.* **29** 4414–30
[23] Kumar R and Kumar A 2024 Effective-diode-based analysis of industrial solar photovoltaic panel by utilizing novel three-diode solar cell model against conventional single and double solar cell *Environ. Sci. Pollut. Res.* **31** 25356–72
[24] Zhang P and Zhou B 2021 Research on general model and parameter characteristics of photovoltaic array *Trans. Electr. Electron. Mater.* **22** 346–56
[25] Hajji M, Yahyaoui Z, Mansouri M, Nounou H and Nounou M 2023 Fault detection and diagnosis in grid-connected PV systems under irradiance variations *Energy Rep.* **9** 4005–17
[26] Joseph E, Singh B S M and Ching D L C 2023 Developing a simple algorithm for photovoltaic array fault detection using MATLAB/Simulink simulation *Int. J. Energy Prod. Manag.* **8** 235–40
[27] Jimeng Li, Gao J, Huang Q and Meng Z 2025 Few-Shot fault diagnosis of rolling bearings based on category-enhanced Siamese attention hierarchical-embedded UNet network *IEEE Trans. Instrum. Meas.* **74** 3543611
[28] Yang K, Zhao L and Wang C 2022 A new intelligent bearing fault diagnosis model based on triplet network and SVM *Sci. Rep.* **12** 5234
[29] Liu H, Zhang F, Tan Y, Huang L, Li Y, Huang G, Luo S and Zeng A 2024 Multi-scale quaternion CNN and BiGRU with cross self-attention feature fusion for fault diagnosis of bearing *Meas. Sci. Technol.* **35** 086138
[30] Kong L, Wengen Li, Yang H, Zhang Y and Guan J 2025 CausalFormer: an interpretable transformer for temporal causal discovery *IEEE Trans. Knowl. Data Eng.* **37** 102–15
[31] Akbari E and Shadlu Milad S 2024 An intelligent ANFIS-based fault detection, classification, and location model for VSC-HVDC systems based on hybrid PCA-DWT signal processing technique *Comput. Electr. Eng.* **120** 109763
[32] Wang Y, Wang W, Li Y, Jia Y, Xu Y, Ling Y and Ma J 2024 An attention mechanism module with spatial perception and channel information interaction *Complex Intell. Syst.* **10** 5427–44
[33] Zhang X, Chen Y, Ni H, Zhang D and Abdulaal M 2024 Multi-subspace self-attention Siamese networks for fault diagnosis with limited data *Signal, Image Video Process.* **18** 2465–72
[34] Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H and Bengio Y 2014 Learning phrase representations using RNN encoder-decoder for statistical machine translation (arXiv:1406.1078)
[35] Bai S, Kolter J Z and Koltun V 2018 An empirical evaluation of generic convolutional and recurrent networks for sequence modeling (arXiv:1803.01271)
[36] Xu Z, Li Y F, Huang H Z, Deng Z and Huang Z 2024 A novel method based on CNN-BiGRU and AM model for bearing fault diagnosis *J. Mech. Sci. Technol.* **38** 3361–9
[37] Chuang P, Lei C, Kuangrong H, Chen S, Cai X and Wei B 2024 A novel dimensional variational prototypical network for industrial few-shot fault diagnosis with unseen faults *Comput. Ind.* **162** 104133