Periodica Polytechnica Electrical Engineering and Computer Science, 69(4), pp. 409–427, 2025

Enhanced CNN-LSTM Feature Extraction and Ensemble Learning for Anomaly Detection in Photovoltaic Data

Hongsheng Su[1*], Yongcang Hu[1], Zhensheng Teng[1]

[1] School of Automation and Electrical Engineering, Lanzhou Jiaotong University, 88 Anning West Road, 730070 Lanzhou, China
[*] Corresponding author, e-mail: 11240361@stu.lzjtu.edu.cn

Received: 15 October 2025, Accepted: 11 December 2025, Published online: 09 January 2026

Abstract

This study proposes an anomaly detection framework that combines CNN–LSTM feature extraction with a boosting-based ensemble strategy to improve the reliability of photovoltaic (PV) system monitoring. Real multi-source PV operational data are first preprocessed using the ISODATA clustering algorithm, which automatically adjusts the number of clusters and reduces redundancy. Principal component analysis (PCA) is then applied to lower data dimensionality while retaining key variability. A hybrid CNN-LSTM network is developed, where CNNs extract spatial features from heterogeneous PV measurements and LSTMs capture temporal dependencies in power sequences. Based on the learned representations, an ensemble model integrates the outputs of Gaussian Mixture Models (GMM), Isolation Forest (IF), and Interquartile Range (IQR) through a boosting-inspired weighting mechanism to enhance robustness under complex operating conditions. Experiments conducted on real PV datasets show that the proposed method achieves nearly 97% anomaly detection accuracy, with an average F1-score of 0.89 ± 0.03 and a recall rate of 0.91 ± 0.02. Compared with single-model baselines, the framework provides more stable performance and maintains a false positive rate below 2.1%, demonstrating its practical value for real-world PV anomaly detection.

Keywords

photovoltaic anomaly detection, ensemble learning, CNN-LSTM, boosting framework

1 Introduction

With the accelerating global transition toward low-carbon and sustainable energy systems, PV technology has become an indispensable pillar of the modern renewable-energy portfolio. High-efficiency, green, and scalable distributed PV installations now play a central role in smart-grid and micro-grid architectures, contributing directly to energy mix optimization, carbon neutrality objectives, and the long-term transformation of electricity markets [1–3]. The rapid progress in PV module efficiency, the integration of battery-energy-storage systems, and recent advances in power electronics have further promoted widespread PV deployment across urban, industrial, and rural energy infrastructures [4]. Consequently, the operational characteristics of PV systems—including energy conversion efficiency, structural reliability, and their complex interactions with environmental factors—have attracted increasing attention from both researchers and industry practitioners. These characteristics not only determine the long-term performance and economic viability of PV installations but also play a critical role in ensuring stable power output under diverse and highly dynamic environmental conditions [5–7]. Ensuring stable, safe, and optimal performance in increasingly complex environments has thus become a fundamental challenge within renewable-energy engineering [8].

However, as PV installations scale up and system architectures become more sophisticated, their operational stability is also exposed to a wider range of potential disturbances [9]. Component aging, partial shading, soiling, module mismatch, inverter degradation, electrical imbalance, and external environmental fluctuations may all induce anomalous operating behaviors. If not detected and handled promptly, such anomalies can lead to reduced power quality, accelerated equipment wear, or even cascading system failures [10–12]. Real-world PV systems consist of heterogeneous components with diverse dynamic behaviors. Consequently, their monitoring data often contain strong noise, non-stationarity, nonlinearity, multimodality,

Cite this article as: Su, H., Hu, Y., Teng, Z. "Enhanced CNN-LSTM Feature Extraction and Ensemble Learning for Anomaly Detection in Photovoltaic Data", Periodica Polytechnica Electrical Engineering and Computer Science, 69(4), pp. 409–427, 2025. https://doi.org/10.3311/PPee.42509

and time-varying statistics [13–15]. These characteristics severely limit the effectiveness of traditional anomaly-detection approaches. Prior research has further emphasized that PV data frequently display non-Gaussian distributions, nonlinear irradiance–temperature–performance couplings, and intricate environmental interactions, all of which complicate the development of highly accurate and robust anomaly-detection models [16–18].

Existing anomaly-detection techniques for PV systems can generally be classified into two categories: statistical modeling and traditional machine learning. Statistical methods—such as 3-sigma rules, Copula dependency modeling, and various robust statistical estimators—are valued for their simplicity and interpretability. Nevertheless, they depend heavily on explicit assumptions about underlying data distributions and often fail to capture the nonlinear relationships among environmental factors, hardware conditions, and electrical behaviors [19]. Traditional machine-learning algorithms, including SVM, kNN, and basic isolation forests, perform reasonably well in scenarios with low-dimensional features. Yet, their expressive capacity is limited when encountering high-dimensional, multi-source PV datasets, and they typically lack mechanisms to model complex temporal dependencies [20]. Consequently, many real-world PV systems still suffer from high false-alarm rates, unstable anomaly recognition, and weak cross-site generalization—particularly when models trained on one plant are deployed at another with different climatic, geographic, or structural conditions [21].

The rapid evolution of artificial intelligence and deep learning has opened new opportunities for data-driven PV monitoring, forecasting, and anomaly management, enabling more accurate modeling of system behavior and more efficient identification of abnormal operating conditions [22]. CNNs excel at extracting local spatial patterns, while LSTMs effectively model long-range temporal correlations, making them well-suited for handling heterogeneous PV time series. Even so, selecting an appropriate deep-learning architecture requires careful consideration of computational cost, model complexity, and practical deployability. Although ConvLSTM offers strong spatiotemporal modeling capabilities, its convolution-gate structure incurs high computational overhead at each recurrent step, making it more appropriate for dense grid-like data such as videos [23]. Transformers, on the other hand, handle long sequences efficiently and capture global dependencies, yet their quadratic complexity with respect to sequence length and substantial data requirements limit their applicability to resource-constrained PV monitoring environments [24]. Additionally, the lack of spatial inductive bias in Transformers reduces their effectiveness in modeling local operational patterns from heterogeneous sensor data. By comparison, hybrid CNN-LSTM architectures combine spatial and temporal modeling in an efficient manner, ensuring favorable performance in terms of model complexity, real-time inference, and robustness to noise—key properties for practical PV monitoring [25].

Despite encouraging progress, current PV anomaly-detection methods continue to face several critical challenges. First, the intrinsic spatiotemporal coupling structure of multi-source PV datasets has not been fully exploited. Second, single-model approaches often struggle to maintain robustness under varying weather patterns, hardware conditions, and environmental disturbances. Third, most existing models rely on large, high-quality labeled datasets, yet anomaly labels in real PV systems are extremely costly and difficult to obtain. Although recent studies have introduced unsupervised-learning approaches, advanced imaging techniques (e.g., electroluminescence and infrared thermography), and dynamic thresholding strategies to enhance detection performance [26], these methods remain constrained by high hardware costs, sensitivity to noise, and limited scalability.

To address the challenges of noise interference, distributional variability, and diverse anomaly patterns in PV operational data, this study proposes an unsupervised multidimensional anomaly-detection framework that integrates deep learning–based feature extraction with ensemble learning to achieve high accuracy, robustness, and cross-site generalizability. The framework begins with ISODATA clustering, which performs dynamic grouping and redundancy elimination, overcoming the limitations of fixed clustering schemes and uncovering latent structural relationships within PV data. Subsequently, a CNN–LSTM network is employed to jointly capture spatial correlations and temporal dependencies, enabling a more comprehensive and stable representation of PV time-series behavior. Three complementary unsupervised detectors—GMM (statistical modeling), IF (structural sparsity), and IQR (distribution deviation)—are then used to independently assess anomaly likelihood from heterogeneous perspectives. Finally, a Boosting-inspired adaptive weighting strategy is introduced to integrate the outputs of these detectors. By dynamically adjusting the contribution of each detector according to its reliability under varying operating conditions, the proposed fusion

mechanism suppresses weak predictions, enhances strong ones, and significantly improves overall performance and cross-plant stability.

Compared with existing literature, the novelty of this work can be summarized as follows:

1. Unlike prior studies on hybrid ConvLSTM or ConvTransformer models that rely primarily on end-to-end deep architectures for feature fusion, this work establishes a three-stage unsupervised framework—feature extraction, multi-detector evaluation, and Boosting-based adaptive fusion—where each component contributes independently and transparently, offering better modularity and interpretability.
2. In contrast to existing hybrid models that often adopt static or uniform weighting schemes, this study introduces a Boosting-style dynamic weighting strategy that adjusts detector weights according to their context-dependent credibility. This represents a key methodological innovation, substantially improving robustness under noisy conditions, imbalanced datasets, and cross-site deployment scenarios.
3. The multi-detector design emphasizes not only model heterogeneity but also complementary statistical principles, enabling the unified detection of both abrupt failures and gradual degradation—an aspect that single-model deep learning approaches such as ConvLSTM or Transformer-based methods generally struggle to handle simultaneously.
4. The incorporation of adaptive clustering and dimensionality-reduction steps makes the framework more suitable for real-world PV systems characterized by diverse devices and varying environmental conditions, thereby enhancing scalability and practical applicability for large-scale deployments.

In summary, the proposed unsupervised anomaly-detection framework—integrating CNN-LSTM feature extraction with a Boosting-based ensemble fusion strategy—advances beyond the limitations of existing hybrid approaches in terms of methodological design, adaptive integration, and cross-station stability. This provides a novel, effective, and scalable pathway toward intelligent and high-reliability PV operation-and-maintenance systems.

2 Theories

2.1 ISODATA clustering algorithm

In the context of photovoltaic power generation data preprocessing, the ISODATA clustering algorithm is extensively employed for handling complex and high-dimensional datasets due to its dynamic adjustment of cluster numbers and adaptive splitting-merging mechanisms [27–29]. Unlike the conventional K-means algorithm, the ISODATA exhibits the capability to adaptively modify the number of clusters during the clustering process and contingent upon the characteristics of the data distribution. Thus, it facilitates more effective identification and mitigation of heterogeneity and noise interference within PV power generation data.

The central tenet of ISODATA is to minimize the sum of the squared distances between data points and their respective cluster centroids. Given a dataset in Eq. (1):

$$\boldsymbol{X} = \{x_1, x_2, \ldots, x_n\}. \tag{1}$$

Each data point $x_i \in \mathbb{R}^d$ is assigned to one of K clusters. The objective function of ISODATA can be expressed as depicted in Eq. (2):

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2 . \tag{2}$$

In Eq. (2), C_k denotes the k^{th} cluster, $\mu_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i$ represents the centroid of cluster C_k, and $\|\cdot\|$ signifies the Euclidean distance.

During the iterative process, ISODATA dynamically assesses the validity of clusters by computing intra-cluster variance and inter-cluster distance. Equation (3) defines the variance of cluster C_k.

$$\sigma_k^2 = \frac{1}{|C_k|} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2 \tag{3}$$

Based on this metric, the algorithm dynamically adjusts the number of clusters in each iteration using the following rule:

1. Cluster splitting

   If the variance σ_k^2 of a cluster exceeds a predefined threshold $\sigma_{\max}^2$, and the number of samples $|C_k|$ within the cluster surpasses the minimum split size $N_{\min}$, the cluster is partitioned into two new clusters. The split typically occurs along the principal direction, which is determined by the eigenvector $\boldsymbol{v}_{\max}$ corresponding to the maximum eigenvalue of the cluster covariance matrix $\boldsymbol{\Sigma}$:

   $$\mu_{k,1} = \mu_k + \alpha \cdot \boldsymbol{v}_{\max}, \quad \mu_{k,2} = \mu_k - \alpha \cdot \boldsymbol{v}_{\max} . \tag{4}$$

   In Eq. (4), α denotes the step size parameter for the splitting.

2. Cluster merging rule
   If the distance between the centroids of two clusters, C_i and C_j, is less than the minimum distance threshold $d_{\min}$, then:
   $$\left\|\mu_i - \mu_j\right\| < d_{\min}. \tag{5}$$
   As illustrated in Eq. (6), the combined variance:
   $$\sigma^2_{i\cup j} = \frac{1}{|C_i|+|C_j|}\sum_{x\in(C_i\cup C_j)}\left\|x-\mu_{i\cup j}\right\|^2. \tag{6}$$
   If the value does not exceed the predefined upper limit, merge clusters C_i and C_j to form a new cluster $C_{i\cup j}$.
3. Iteration termination criteria
   The algorithm terminates and outputs the final clustering results when the number of clusters remains stable across successive iterations, or when the magnitude of all cluster center updates $\left\|\mu_k^{(t+1)}-\mu_k^{(t)}\right\|$ falls below the convergence threshold ϵ.

To ensure the effectiveness and stability of the ISODATA algorithm in processing photovoltaic power generation data, key parameters were set based on extensive experimentation. Specifically, the maximum allowable intra-cluster variance threshold $\sigma_{\max}$ was set to 0.5; when the variance of a cluster exceeds this threshold, a splitting operation is triggered to maintain compactness and representativeness of clusters. This threshold was tuned to accommodate the heterogeneous characteristics of PV data. To prevent over-segmentation due to insufficient samples, the minimum cluster size for splitting $N_{\min}$ was set to 10, ensuring that splits occur only when enough samples are present, thereby enhancing the stability and reliability of clustering results. During cluster splitting, the step size parameter α was set to 0.5 to control the displacement of new centroids, allowing for refined cluster structures without introducing excessive noise. For cluster merging, the minimum centroid distance threshold $d_{\min}$ was set to 0.3, and the variance upper limit after merging σ^2_{merge} was set to 0.4. Clusters are merged only if both distance and variance criteria are met, effectively reducing redundant clusters while avoiding over-partitioning. Finally, the iteration termination criteria were defined as either the number of clusters remaining unchanged across consecutive iterations or the movement of all cluster centroids falling below 1×10^{-4}, ensuring convergence and stability.

Leveraging the aforementioned mechanism, ISODATA exhibits adaptive clustering capabilities, obviating the need for pre-defined cluster numbers. This attribute is particularly salient in the context of photovoltaic power generation data, which frequently manifests significant heterogeneity and uncertainty. The split-merge rules inherent to ISODATA substantially enhance the robustness and interpretability of the clustering process.

2.2 Spatiotemporal feature extraction optimized by CNN-LSTM

It is crucial to extract spatial features from the input data while also capturing its temporal dependencies. To this end, this paper introduces a CNN-LSTM joint architecture. This architecture leverages CNNs to extract local features from the input data and employs LSTMs to model temporal relationships, ultimately generating feature vectors that represent the spatiotemporal information of the data.

2.2.1 Feature extraction capabilities of the CNN module

As illustrated in Fig. 1, the CNN serves as the model's front-end module. It is responsible for feature extraction from the input data, where T denotes the time steps, H and W represent the spatial dimensions and C signifies the number of channels. Leveraging the local receptive fields and parameter sharing inherent to convolutional layers, the CNN effectively extracts spatial features at each time step.

Assuming a convolutional kernel $\boldsymbol{K}$, the output feature map $\boldsymbol{Y}$ resulted from the convolution operation which can be formulated as

$$\boldsymbol{Y}_t^{(l)} = f\left(\sum_{c=1}^{C}\boldsymbol{K}^{(l)} * \boldsymbol{X}_t^{(c)} + b^{(l)}\right). \tag{7}$$

Herein, * denotes the convolution operation, l represents the layer number of the convolutional layer, f signifies the activation function (ReLU activation function) and $b^{(l)}$ is the bias term. Following further dimensionality reduction via the pooling layer, the feature vector $\boldsymbol{z}_t$ is obtained for each time step.

2.2.2 Temporal modeling capabilities of the LSTM module

As illustrated in Fig. 2, the LSTM network represents an advanced variant of Recurrent Neural Networks (RNNs) and

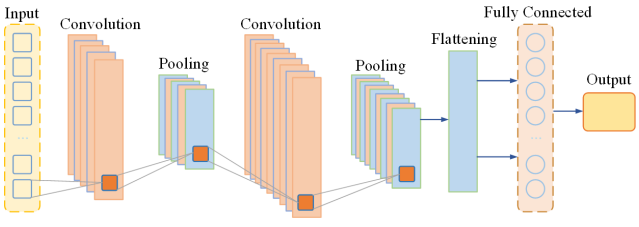

Fig. 1 Network architecture of the CNN layer

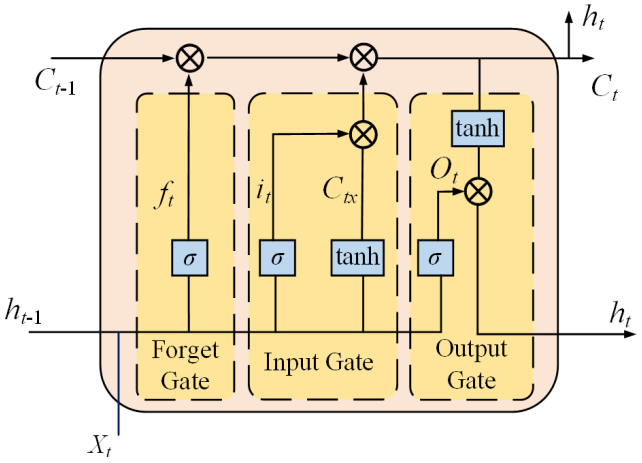

Fig. 2 Long short-term memory network architecture

designed to manage long-range dependencies via its gating mechanisms [30]. The feature sequence $\{z_t\}_{t=1}^{T}$, extracted by a CNN is subsequently fed into the LSTM. The update equations for its hidden state are defined as follows:

$$i_t = \sigma\left(W_i z_t + U_i h_{t-1} + b_i\right) \tag{8}$$

$$f_t = \sigma\left(W_f z_t + U_f h_{t-1} + b_f\right) \tag{9}$$

$$o_t = \sigma\left(W_o z_t + U_o h_{t-1} + b_o\right) \tag{10}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh\left(W_c z_t + U_c h_{t-1} + b_c\right) \tag{11}$$

$$h_t = o_t \odot \tanh\left(c_t\right). \tag{12}$$

In Eqs. (8)–(12), σ denotes the tanh activation function. The $\odot$ represents element-wise multiplication and c_t is the cell state. The h_t is the hidden state and i_t,f_t, o_t are corresponding to the input gate, forget gate and output gate, respectively.

2.2.3 The CNN-LSTM hybrid architecture

Based on Sections 1 and 2, it is evident that CNNs are capable of extracting spatial features from each timestep's input while LSTMs can model the dynamic evolution of these features over time. The fusion mechanism lies in the following: CNNs initially map the raw input sequence from multidimensional spatiotemporal data to a one-dimensional time series feature vector which is then subjected to temporal modeling by the LSTM. Specifically, the input sequence can be represented as

$$\boldsymbol{X} = \{x_1, x_2, \ldots, x_T\}, \quad x_t \in \mathbb{R}^{H \times W \times C}. \tag{13}$$

In Eq. (13), T denotes the timestep. Following CNN feature extraction and pooling, the input at each timestep t is mapped to a low-dimensional feature vector:

$$z_t = f_{\text{CNN}}\left(x_t\right), \quad z_t \in \mathbb{R}^d. \tag{14}$$

In Eq. (14), $f_{\text{CNN}}(\cdot)$ denotes the non-linear mapping of the CNN, and d represents the feature dimension. The feature vectors from all timesteps are concatenated to generate a feature sequence:

$$\boldsymbol{Z} = \left[z_1, z_2, \ldots, z_T\right]^{\mathrm{T}}, \quad \boldsymbol{Z} \in \mathbb{R}^{T \times d}. \tag{15}$$

Subsequently, the sequence $\boldsymbol{Z}$ is inputted into the LSTM, leveraging its gating mechanisms to model temporal dependencies. The hidden state update equations are presented in Eqs. (8)–(12), and the resulting hidden state sequence is:

$$H = \{h_1, h_2, \ldots, h_T\}, \quad h_t \in \mathbb{R}^m. \tag{16}$$

In Eq. (16), m denotes the number of hidden units within the LSTM architecture. The LSTM's capacity to retain long-term dependencies via the memory cell, c_t enables the model to capture not only the local spatial patterns of the input signal but also to effectively model the dynamic evolutionary trends across time steps.

Furthermore, the final hidden state h_T of the LSTM can be fed into subsequent fully connected layers to perform anomaly detection or prediction tasks as given in Eq. (17).

$$y = f_{\text{FC}}\left(h_T\right) \tag{17}$$

In summary, the core fusion of CNN-LSTM lies in the decomposition of spatiotemporal signals into spatial features by CNN which are then modeled by LSTM across the temporal dimension, thus enabling efficient learning of intricate spatiotemporal dependencies.

2.3 Anomaly detection in traditional machine learning

PV power generation data may contain substantial anomalies caused by weather fluctuations and equipment malfunctions or sensor errors [31]. To efficiently identify these outliers, this paper proposes an ensemble detection framework integrating GMM, IF and IQR methods [32].

2.3.1 Anomaly detection with IF

The IF algorithm constructs isolation trees through random partitioning of the sample space whose core premise is that anomalous instances are more likely to be isolated during the splitting process. The mathematical formulation of IF is as follows:

- Each isolation tree is constructed by randomly selecting a feature f and a split point θ where the splitting rule $g(x_i)$ is defined as

$$g(x_i)=\begin{cases}1, & \text{if } x_i[f]<\theta\\ 0, & \text{otherwise}\end{cases}. \tag{18}$$

- A sample x_i traverses the tree's splitting path until its partitioned subspace contains only a single data point.
- The path length $h_t(x_i)$ represents the number of splits required to isolate the sample and has its expected value $\mathbb{E}\left[h(x_i)\right]$ calculated through averaging across multiple isolation trees:

$$\mathbb{E}\left[h(x_i)\right]=\frac{1}{T}\sum_{t=1}^{T}h_t(x_i). \tag{19}$$

- Anomaly scores are computed via normalization based on path length:

$$\text{Score}_{\text{IF}}(x_i)=2^{-\frac{\mathbb{E}\left[h(x_i)\right]}{c(n)}}. \tag{20}$$

- In Eqs. (18)–(20), $c(n)=2H(n-1)-\frac{2(n-1)}{n}$, with $H(n)$ denoting the n^{th} harmonic number $H(n)=\sum_{i=1}^{n}\frac{1}{i}$. A higher anomaly score suggests that a sample is more likely to be an outlier.

2.3.2 GMM anomaly detection

GMM represents a probabilistic clustering approach, where the modeling data distribution by assuming data points originated from a mixture of Gaussian distributions. Within anomaly detection, the GMMs aim to identify outliers by establishing the probability density function of normal data and subsequently identifying points that deviate from these distributions [33].

Within the GMM framework, as given in a dataset $x = \{x_1, x_2, \ldots, x_n\}$, the data points x_i are assumed to be sampled from a mixture of Gaussian distributions. The objective of the model is to detect anomalies by learning the underlying mixture distribution of the data.

Model Definition: GMMs are typically defined as a weighted sum of several Gaussian distributions, mathematically represented in Eq. (21):

$$p(x)=\sum_{k=1}^{K}\pi_k\mathcal{N}\left(x|\mu_k,\Sigma_k\right). \tag{21}$$

In Eq. (21), π_k denotes the weight associated with the k^{th} Gaussian component, $\mathcal{N}\left(x|\mu_k,\Sigma_k\right)$ represents a Gaussian distribution characterized by a mean of μ_k and a covariance matrix of Σ_k. The k signifies the total number of Gaussian components within the mixture model.

To train the model, the Expectation-Maximization (EM) algorithm is typically employed. This iterative algorithm estimates model parameters through alternating Expectation (E) and Maximization (M) steps. Specifically, the E-step computes the probability of each data point belonging to each Gaussian distribution while the M-step updates the model parameters based on these probabilities.

Anomaly Detection: following the training of the GMM, the likelihood of a novel observation sample x_{new} can be computed within the established model as described in Eq. (22).

$$p(x_{\text{new}})=\sum_{k=1}^{K}\pi_k\mathcal{N}\left(x_{\text{new}}|\mu_k,\Sigma_k\right) \tag{22}$$

An observation is flagged as an outlier if its likelihood falls below a predefined threshold. Typically, a threshold is denoted as τ which is established to differentiate between normal and anomalous data points. The formal representation of the decision rule is expressed in Eq. (23):

$$\text{if } p(x_{\text{new}})<\tau, \text{ then } x_{\text{new}} \text{ is considered an anomaly}. \tag{23}$$

2.3.3 IQR-based anomaly detection

The IQR method identifies outliers by assessing the distribution range of a dataset. The core principle involves calculating the first quartile Q_1 and the third quartile Q_3 of the data and defining the normal range of the data using their difference in IQR [34].

The IQR is computed as

$$\text{IQR}=Q_3-Q_1. \tag{24}$$

In Eq. (24), Q_1 and Q_3 denote the 25$^{\text{th}}$ and 75$^{\text{th}}$ percentiles of the data, respectively. The IQR quantifies the dispersion of the central 50% of the dataset and serves as a crucial metric for assessing the data's variability.

In the IQR outlier detection method, the outliers are defined as data points that fall outside the boundaries of the data distribution. Specifically, a threshold (typically 1.5 times the IQR) is used to establish the lower and upper bounds of the data. As demonstrated in Eqs. (25) and (26), the upper and lower bounds of the data can be definitively established.

$$\text{Lower Bound}=Q_1-1.5\times\text{IQR} \tag{25}$$

$$\text{Upper Bound}=Q_3+1.5\times\text{IQR} \tag{26}$$

Any data point exhibiting a value below the lower bound or exceeding the upper bound is classified as an outlier. Consequently, a data point is denoted as x_i and the

conditions specified in Eq. (27) are satisfied. Hence, the data point is classified as an outlier:

$$x_i < Q_1 - 1.5 \times \mathrm{IQR} \text{ or } x_i > Q_3 + 1.5 \times \mathrm{IQR} . \quad (27)$$

2.4 A Boosting-based anomaly scoring and thresholding

To leverage the complementary advantages of the detectors, this study proposes a lightweight ensemble fusion mechanism inspired by Boosting principles [35]. Traditional Boosting methods typically rely on ground-truth labels for training samples to compute error residuals; however, in photovoltaic anomaly detection, data are predominantly unlabeled or weakly labeled, rendering direct application challenging. Consequently, this research introduces a custom unsupervised iterative weighting strategy to enable adaptive weight allocation among multiple base learners [36].

Firstly, as illustrated in Fig. 3, the stability of individual base learners in an unsupervised environment exhibits noticeable variation. To quantitatively assess this behavior, this study introduces the concept of the Consistency Score. For each training sample, the three models separately produce anomaly detection results, which are combined through majority voting to generate pseudo-labels PL_i. Subsequently, the consistency between each model's output and the pseudo-labels is computed, with the consistency score of the t^{th} base learner defined according to Eq. (28).

$$c_t = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[M_t(i) = PL_i \right] \quad (28)$$

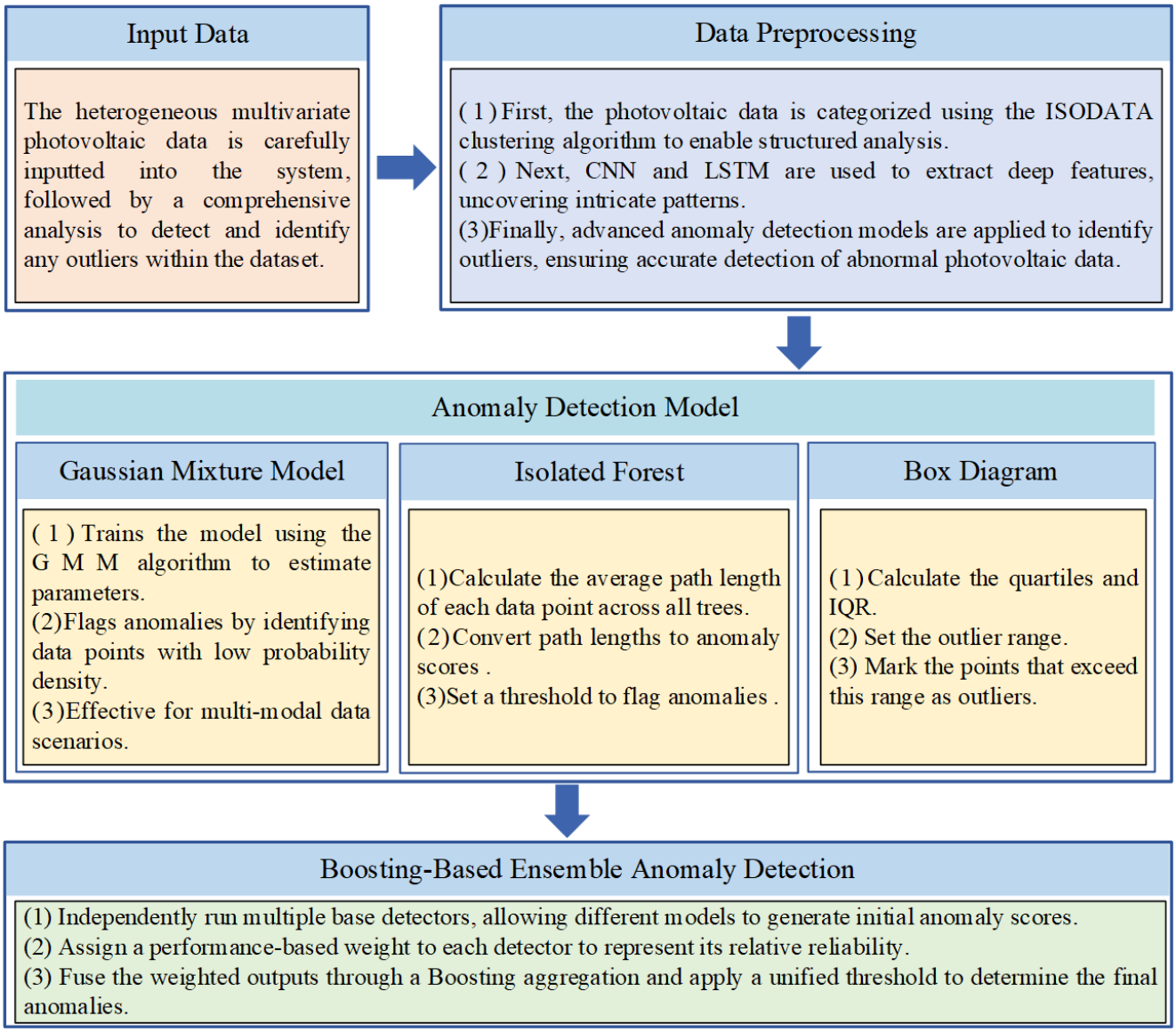

Fig. 3 Flowchart of the ensemble learning–based anomaly detection model

Where $M_t(i)$ denotes the judgment of the t^{th} model on the i^{th} sample, N represents the total number of samples, and $\mathbb{I}[\cdot]$ is an indicator function.

After obtaining the consistency scores, the models are adaptively weighted according to their scores using a linear proportional scheme as described in Eq. (29).

$$\alpha_t = \frac{c_t}{\sum_{k=1}^{T} c_k} \tag{29}$$

And by normalization, ensure consistency across $\sum_{t=1}^{T} \alpha_t = 1$. The final integrated anomaly score is obtained by weighted summation of the outputs from each model, resulting in the sample's anomaly score as expressed in Eq. (30).

$$S(i) = \sum_{t=1}^{T} \alpha_t \times S_t(i) \tag{30}$$

Where $S_t(i)$ represents the anomaly score assigned by the t^{th} base learner to sample i.

To convert the weighted anomaly score $S(i)$ into a final anomaly/normal classification, this study employs an Adaptive Thresholding technique for decision-making. The specific methodology is as follows:

1. Perform a kernel density estimation of the distribution of the aggregate anomaly scores $S(i)$.
2. Using the upper bound of the standard distribution as the threshold:

$$\theta = \mu_s + k\sigma_s . \tag{31}$$

   Where μ_s and σ_s denote the mean and standard deviation of the composite scores, respectively, and k represents the sensitivity coefficient determined based on the validation set.
3. Conduct final classification for each sample.

$$\text{Label}(i) = \begin{cases} 1, & S(i) > \theta \ \ (\text{Anomaly}) \\ 0, & S(i) \le \theta \ \ (\text{Normal}) \end{cases} \tag{32}$$

Based on this methodology, when a sample's composite anomaly score significantly deviates from the normal distribution range, it is classified as anomalous data. This adaptive thresholding approach dynamically adjusts according to variations in illumination, temperature, and operational conditions, effectively mitigating false positives associated with fixed thresholds under complex operational scenarios.

3 Experimental section

3.1 Dataset characterization

The dataset employed in this study originates from the 2024 Digital China Innovation Contest[1]. It encompasses the operational monitoring data from nine photovoltaic power stations within a specific region over a three-month period. The data comprises a total of 79,488 records as sampled at 15-minute intervals. Feature variables are categorized into two primary groups: meteorological features (4 dimensions) including atmospheric pressure (hPa), relative humidity (%), cloud cover (okta) and 10-meter wind speed (m/s). It is characterized by the external climatic conditions influencing PV generation variability and operational features (6 dimensions). It encompasses the DC output power, AC output power, inverter voltage, current, frequency and equipment operational status. It reflects the operational performance and equipment health of the power stations. Overall, the dataset incorporates 10 feature dimensions and exhibits a multi-source and heterogeneous nature. Specifically, variations in monitoring equipment and data precision across different power stations lead to distributional differences in similar features. Furthermore, meteorological factors and operational statuses are derived from environmental monitoring and power equipment, respectively. It represents data dimensions with distinct physical attributes and necessitates the unified preprocessing and feature fusion techniques. Following expert annotation, 4,947 anomalous samples were identified within the dataset, resulting in an anomaly rate of approximately 6.22%. It provides a robust data foundation for subsequent anomaly detection and predictive modeling of PV systems.

3.2 Multi-model anomaly detection algorithms and parameter optimization elaboration

The study initially employs deep learning techniques for feature extraction from PV data as illustrated in Fig. 4.

Subsequently, anomaly detection is performed using GMM, IF and IQR. An ensemble learning approach is incorporated a weighted voting mechanism and is implemented to enhance the accuracy and robustness of anomaly detection. The selection and optimization of key

1 The "2024 Digital China Innovation Contest" is a public competition. The dataset used in this study was obtained by the authors through participation in this competition in compliance with its regulations. If readers require access to the experimental data used in this paper, they may request it by contacting the corresponding author of this article, subject to compliance with relevant data usage policies.

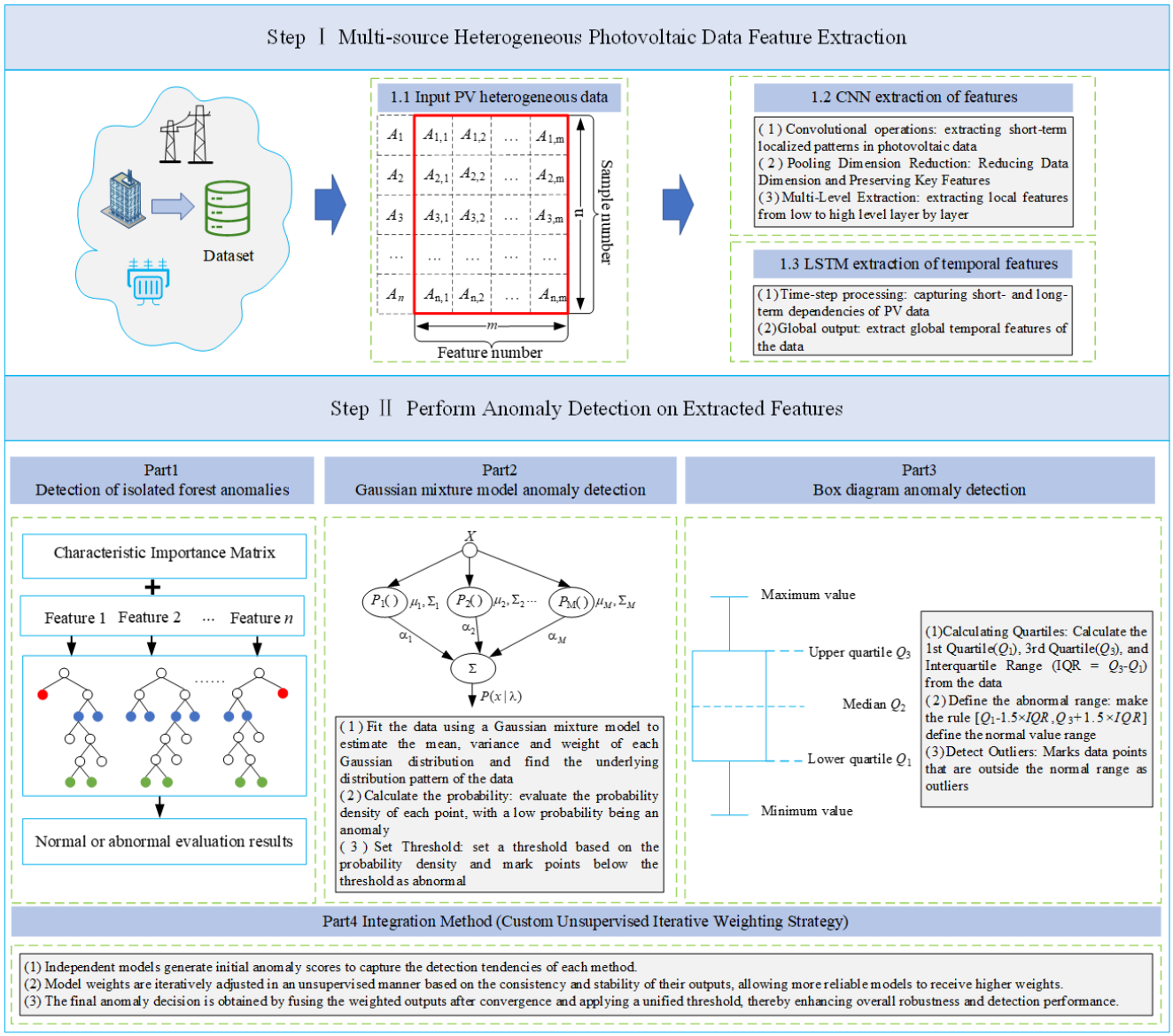

Fig. 4 Flowchart of the multi-model anomaly detection algorithm

parameters for each algorithm are detailed. Initially, the PV data distribution is posited to be a composite of two components within the GMM framework and normal and anomalous. Consequently, the number of mixture components is set to two. Furthermore, the "full" covariance type is selected to accommodate the intricate distributional characteristics of the data.

Subsequently, a boosting-based ensemble learning framework is employed for training within the IF algorithm. The maximum number of splits for the decision trees (MaxNumSplits) is configured to 20, aiming to balance computational efficiency and model accuracy. The classification of samples is performed with samples exhibiting higher scores being identified as outliers based on the anomaly scores generated by the model. Finally, the boxplot method leverages the distributional statistics of the data. It computes the 25th percentile (Q_1) and the 75th percentile (Q_3) in conjunction with 1.5 times the IQR to establish upper and lower bounds for outlier detection. Any data points exceeding these bounds are classified as anomalies. In the weighted fusion phase, the outputs of the multiple detectors were combined using the converged weights obtained from the unsupervised iterative adjustment process. These adaptive weights reflect the relative reliability of each model under unlabeled conditions. The final anomaly assessment for each sample was then determined by applying a unified threshold to the fused anomaly score, thereby improving the overall detection accuracy and robustness.

3.3 Evaluation metrics

Comprehensive evaluation of model performance is critical in anomaly detection. This study adopts accuracy, precision, recall and F1-score as core metrics.

3.3.1 Accuracy

Accuracy, a commonly used metric in classification tasks, quantifies the proportion of correctly classified instances relative to the total number of instances. Within the context of anomaly detection, the accuracy serves as a measure of the model's overall classification efficacy across all samples. The formula for calculating accuracy is as delineated in Eq. (33).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{33}$$

In this context, TP denotes the count of true positives. It further represents the instances correctly identified as anomalous. TN signifies the number of true negatives and indicates normal instances accurately classified into several terms such as FP corresponds to false positives where normal instances are erroneously flagged as anomalous and FN represents false negatives. It denotes the anomalous instances incorrectly classified as normal. A higher accuracy value suggests superior overall model performance.

3.3.2 Precision

Precision quantifies the proportion of actual anomalies among the samples predicted as anomalous by the model. Within the context of anomaly detection, precision reflects the accuracy of the model's predictions regarding anomalous instances. The formula for calculating precision is as delineated in Eq. (34).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{34}$$

A higher precision value suggests superior accuracy in anomaly detection and reflects a reduced rate of false alarms.

3.3.3 Recall

Recall quantifies the proportion of actual anomalous instances correctly identified by the model. Within the context of anomaly detection, it recalls and emphasizes the model's sensitivity specifically its capacity to detect as many anomalous samples as possible. The formula for calculating recall is as delineated in Eq. (35).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{35}$$

A higher recall value suggests and enhances anomaly detection efficacy. Also, it is used potentially at the expense of increased false positives.

3.3.4 F1-score

The F1-score represents the harmonic meaning of precision and recall. It integrates the strengths and mitigates the weaknesses of these metrics. This is particularly pertinent in scenarios of class imbalance such as anomaly detection where the ratio of normal to anomalous instances is often skewed. The formula for calculating F1-score is as delineated in Eq. (36).

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{36}$$

A higher F1-score indicates a more optimal balance between Precision and Recall. It is commonly considered a critical metric for evaluating the performance of classification models.

3.4 Algorithm of the proposed framework

To enhance reproducibility and to provide a clear overview of the proposed anomaly detection framework, Section 3.4 summarizes the complete training and testing pipeline in the form of Algorithm 1. The workflow consists of four major stages:

1. data preprocessing,
2. CNN-LSTM based feature extraction,
3. training of the ensemble anomaly detection models, and
4. inference using a weighted decision fusion strategy.

Algorithm 1 Training workflow of the proposed CNN-LSTM ensemble model

```
// ----- Data Preprocessing -----
Clean X by removing missing and corrupted samples
Normalize features using min-max or z-score scaling
Segment X into fixed-length temporal windows
Split the segmented dataset into training and validation sets
// ----- CNN-LSTM Feature Extraction -----
for each window x_i in training set do
    h_i = CNN(x_i)            // extract local spatial patterns
    f_i = LSTM(h_i)           // capture temporal dependencies
    Store feature vector f_i
end for
Construct feature matrix F = {f_i}
// ----- Train Base Anomaly Detectors -----
Train IF model using F                  → IF_model
Train GMM using F                       → GMM_model
Compute IQR-based thresholds using F → IQR_model
// ----- Compute Ensemble Weights -----
Evaluate three detectors on the validation set
Convert accuracy scores into normalized weights α1, α2, α3
return F(·), IF_model, GMM_model, IQR_model, α1, α2, α3
```

Algorithm 1 summarizes how raw photovoltaic (PV) power data are transformed into model-ready feature embeddings and processed through the multi-model ensemble to produce final anomaly decisions.

3.4.1 Training workflow of the CNN-LSTM–based ensemble model

To provide a clearer and more reproducible description of the proposed method, Section 3.4.1 presents the Algorithm 1 for the training workflow. The training phase aims to

1. extract spatiotemporal representations from raw PV time-series data using the CNN-LSTM module,
2. train three unsupervised anomaly detectors based on the extracted features, and
3. compute adaptive ensemble weights according to model consistency on the validation set.

Algorithm 1 summarizes the complete training pipeline, where:

- Input: Raw PV time-series data X;
- Output: a Trained CNN-LSTM encoder $F(\cdot)$; trained IF, GMM, and IQR detectors;
- Normalized Ensemble Weights $\alpha_1, \alpha_2, \alpha_3$.

3.4.2 Testing workflow for anomaly detection

For completeness, we also provide Algorithm 2 describing the testing (inference) stage. During testing, the trained CNN-LSTM network generates feature embeddings for incoming PV data windows, after which each anomaly detector produces an anomaly score. The final detection result is obtained by applying the learned ensemble weights to fuse these scores. Algorithm 2 outlines the complete testing workflow, where:

- Input: New PV data window x; trained CNN-LSTM encoder $F(\cdot)$;
- IF_model, GMM_model, IQR_model; weights α_1, α_2, α_3;
- Output: Final anomaly decision y_{pred}.

Algorithm 2 Testing workflow

```
// ----- Feature Extraction -----
h = CNN(x)
f = LSTM(h)
// ----- Individual Model Predictions -----
y_IF = IF_model.predict (f)
y_GMM = GMM_model.predict (f)
y_IQR = IQR_model.detect (f)
// ----- Weighted Ensemble Fusion -----
score = α1 × y_IF + α2 × y_GMM + α3 × y_IQR
y_pred = (score > threshold) ? "Anomaly" : "Normal"
return y_pred
```

Algorithm 2 provides a clear and comprehensive summary of the proposed end-to-end anomaly detection workflow. In the training phase, the CNN-LSTM feature extractor converts raw photovoltaic time-series data into compact spatiotemporal embeddings, which subsequently serve as inputs for training the Isolation Forest, GMM, and IQR detectors. The weights of the ensemble are determined automatically based on validation consistency, ensuring that models with more stable detection behavior contribute more significantly. During testing, each data window is encoded into a feature vector and independently evaluated by the three detectors, after which their weighted outputs are fused to produce the final anomaly decision. This unified and structured pipeline enhances the interpretability, reproducibility, and robustness of the entire framework, providing a reliable basis for deploying the model in real-world PV monitoring scenarios.

4 Anomaly detection results

4.1 Dataset partitioning

To ensure fair and effective evaluation, the dataset was split into 80% training set and 20% test set. The dataset consists of nine independent photovoltaic data sequences, where the training set is used for model calibration, and the test set is employed for validation and anomaly detection performance assessment.

The proposed model adopts an end-to-end joint training strategy, integrating a CNN and a LSTM into a unified deep learning framework optimized simultaneously via back-propagation. During training, the Adam optimizer was used with an initial learning rate of 0.001, which was decayed by a factor of 0.1 at the 30th and 60th epochs to improve convergence speed and generalization capability. The model was trained for a total of 10 epochs with a batch size of 64, balancing training efficiency and GPU memory constraints. Experiments were conducted on a computing platform equipped with an NVIDIA GeForce RTX 3080 GPU and 32 GB of RAM, implemented using the PyTorch framework. This training configuration effectively enables the model to capture the spatiotemporal features of photovoltaic data, thereby enhancing anomaly detection performance.

4.2 Baseline model configuration and fairness guarantees

To ensure fairness in comparing the proposed model with baseline models, this study standardized parameter tuning procedures for all baseline methods (CNN-LSTM, IF, GMM, IQR). All methods strictly employed identical data

preprocessing techniques and identical training/validation set splits to eliminate performance bias from inconsistent data partitioning.

For the deep learning baseline CNN-LSTM, this study fixed parameters based on prior experimental experience and commonly used settings in the literature: a convolution kernel size of 3, 32 channels per convolution layer, 128 hidden units in the LSTM, Adam optimizer, learning rate of 0.001, and 100 training epochs. All parameters were selected based on validation set performance to determine the final configuration. For IF and GMM, grid search within common parameter ranges was employed to tune the number of trees (n_estimators) and Gaussian components (n_components), respectively, selecting the combination yielding optimal validation set performance. IQR has fewer parameters and uses the standard threshold of 1.5 × IQR.

To further ensure objectivity in comparisons, all models were run on identical hardware (NVIDIA RTX 3080 GPU + 32 GB RAM), with complete training and inference times recorded. Parameters and runtime metrics are detailed in Table 1. Additionally, this study employed one-way ANOVA to test the statistical significance of F1 scores across multiple models. Results indicate that the performance improvement of the proposed method over baseline methods is statistically significant ($p < 0.05$), confirming that the performance gains are not due to random fluctuations.

4.3 Comparative evaluation of anomaly detection methods

A comparative analysis of the accuracy rates across photovoltaic plant samples 1–9 reveals significant disparities among the various methodologies as illustrated in Fig. 5. The boosting-based ensemble learning approach demonstrated superior performance across all test scenarios, achieving an average accuracy of 0.973. This method not only outperformed others in terms of precision but also exhibited a high degree of consistency across different PV plants.

In contrast, the Z-score and Transformer models yielded average accuracies of 0.899 and 0.904, respectively. The IQR and IF methods exhibited weaker overall performance with average accuracies of 0.808 and 0.803, respectively. Most baseline methods achieved accuracy below 0.88 across the PV plants. Overall, the ensemble method demonstrated enhanced robustness in the presence of multi-source heterogeneity and noise interference, making it more suitable for anomaly detection in large-scale PV plants.

The comparative precision results are further underscoring the superior anomaly localization capabilities of the integrated framework as illustrated in Fig. 6. The average precision achieved is 0.923 and demonstrates the near-perfect performance across all power stations.

In contrast, the Z-score and Transformer methods exhibit average precisions of 0.854 and 0.889, respectively. However, the IQR and IF methods yielded results of only 0.838 and 0.795 with several power stations falling below 0.800. These findings indicate that the ensemble method significantly outperforms traditional statistical methods and tree-based models in the precise identification of anomalous data points.

The recall results indicate that the IF and IQR methods exhibit a performance advantage in this metric with average recall rates of 0.916 and 0.939, respectively as depicted in Fig. 7. Thereby, it captures a greater number of anomalies. However, their precision is suboptimal which limits their overall reliability. The ensemble method achieves an average recall of 0.911, slightly below the aforementioned methods. So far, it demonstrates more consistent performance across different power stations. In contrast, the Z-score method constrained by distributional assumptions yields an average recall of only 0.781 which indicates a significant performance decline in complex scenarios. The Transformer model achieves an average recall of 0.876 with overall performance remaining unsatisfactory.

The F1 score results further validate the integrated method's advantage in balancing precision and recall as illustrated in Fig. 8. The integrated method achieved an average F1 score of 0.887 and maintained a leading position across all power stations. It further demonstrates a comprehensive "high precision-high recall" advantage of this system.

Table 1 Summary of baseline model parameter settings and runtime

Model	Core parameter settings	Parameter selection method	Training/Runtime duration (s)
CNN-LSTM (Baseline)	Conv channels = 32; kernel size = 3; LSTM hidden units = 64; learning rate = 0.001; Batch = 64	Validation set parameter tuning	148
IF	n_estimators = 100; contamination = 0.1	Grid search (50/100/150 Trees)	9.6
GMM	n_components = 3; covariance_type = "full"	Grid search (2/3/4 Components)	7.3
IQR	Threshold = 1.5 × IQR	Experience standard	0.04
Ensemble learning methods	Employ a custom unsupervised iterative weighting strategy	Adaptive weight Assignment	159

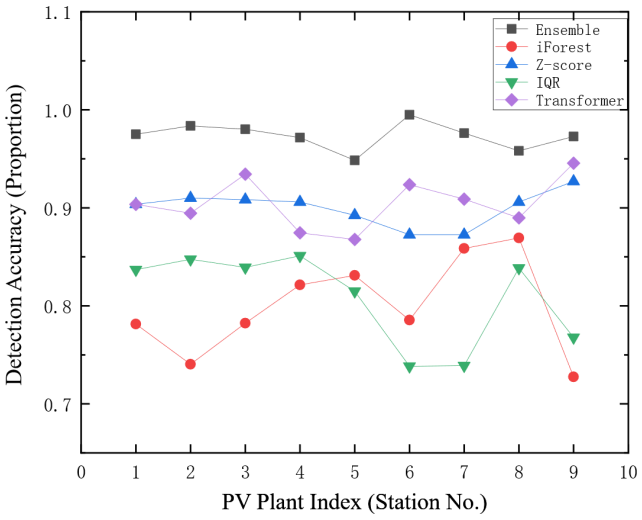

Fig. 5 Comparison of accuracy for different methods

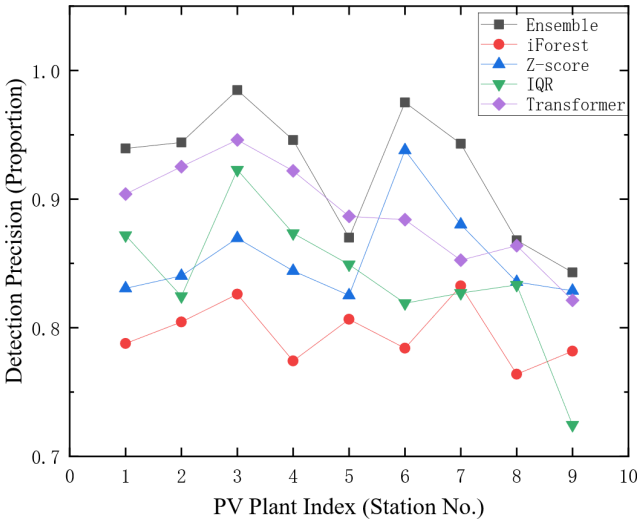

Fig. 6 Comparison of precision rates for different methods

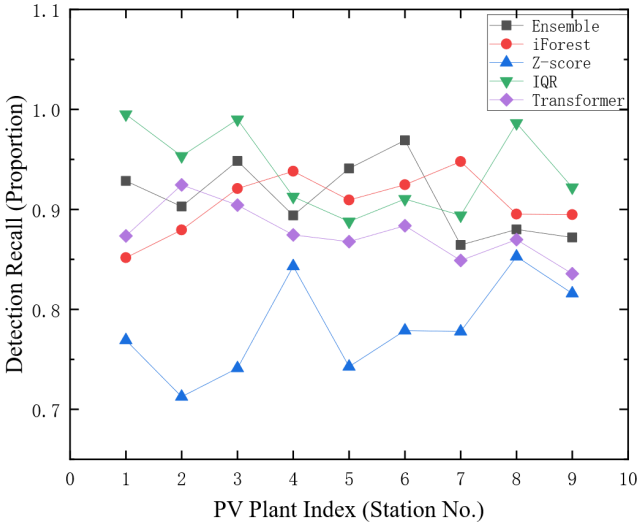

Fig. 7 Comparison of recall rates for different methods

The Transformer method followed its stability which was insufficient in some power stations with an average F1 score of 0.876. The average F1 scores for IQR and IF were 0.784 and 0.758, respectively exhibit significant overall fluctuation. The Z-score method had the lowest average F1 score (at only 0.749) with a noticeable degradation in detection performance under non-normal data scenarios.

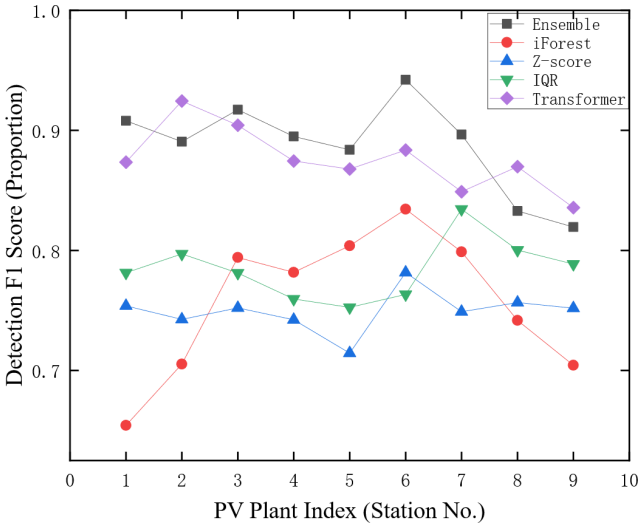

Fig. 8 Comparison of F1 scores for different methods

In summary, the ensemble learning approach demonstrates a comprehensive advantage across the key metrics of accuracy, precision, recall and F1-score. It effectively achieves a synergistic optimization of detection accuracy and generalization capability through feature fusion and dynamic weighting mechanisms. The overall performance significantly surpasses both traditional statistical methods and single-model baselines as well as the Transformer framework. These findings underscore the practical value of ensemble methods in complex photovoltaic scenarios, offering robust support for the intelligent operation and maintenance of large-scale power plants.

4.4 Ablation study on model parameters

To further validate the model's sensitivity to key parameters, we designed an ablation study to examine the impact of the number of CNN layers. The number of LSTM hidden units and the number of iterations on detection performance were determined. Hence, this experiment summarizes the results under different parameter combinations as given in Table 2. Overall, the model exhibited high accuracy (96.7%–97.3%) across all configurations, indicating that the proposed CNN-LSTM framework has good robustness.

Regarding data splitting, we tested two schemes: 70% training set and 30% test set, and 80% training set and 20% test set. Experiments revealed minimal performance differences between the two, with the 80/20 split showing slight advantages across multiple metrics. Therefore, this

Table 2 Impact of key parameters on model performance

Number of CNN layers	LSTM units	Number of iterations	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Average training time (s)
1	64	100	96.7	92.1	91.7	88.8	120
2	64	100	96.8	92.1	91.9	88.9	170
3	64	100	97.0	92.2	92.0	89.0	210
2	32	100	97.2	92.2	92.1	89.1	140
2	128	100	97.3	92.4	92.1	88.7	160
2	64	50	97.1	92.3	92.0	88.8	90
2	64	150	97.3	92.1	91.8	88.7	250

paper adopts the 80/20 split to ensure sufficient training data and stable test evaluation.

Initially, the model's accuracy increased from 96.7% to 97.0% as the number of layers increased from 1 to 3 in the comparison of CNN layer numbers. However, the improvement was limited which indicated that excessively deep convolutional structures do not significantly enhance the extraction of boundary features from photovoltaic data. Subsequently, the model's accuracy slightly improved (97.2%–97.3%) in the selection of LSTM hidden units when the number of units increased from 32 to 128. Nevertheless, this was accompanied by minor fluctuations in the F1-score. It further suggested that an excessive number of hidden units did not substantially improve the modeling of temporal dependencies. Finally, there was little performance difference between 50, 100 and 150 iterations (97.1%–97.3%) regarding the setting of iteration numbers. It indicated that the model essentially converged after approximately 50 iterations. Further increasing the number of iterations primarily was increased by training costs rather than improving performance.

Comparative analysis reveals that the optimal outcome is achieved with a configuration of a 2-layer CNN, 128 hidden units and 100 iterations. However, it yields an accuracy of 97.3%. Nevertheless, the improvement is marginal, which ranges from 0.1% to 0.3% and does not significantly outperform other configurations. Consequently, this study adopts a 2-layer CNN with 64 hidden units and 100 iterations as a compromise to balance performance, computational cost and model complexity.

In addition to accuracy evaluation, this study analyzed the computational complexity and runtime performance of all methods. The results, summarized in Tables 1 and 2, show that the baseline models exhibit lower computational overhead, whereas the proposed ensemble CNN-LSTM framework requires slightly higher training time due to deeper feature extraction and iterative weighting mechanisms. However, the increase in computation is moderate and proportionate to the performance improvement (1.8%–3.1% in F1-score). Furthermore, inference time remains within an acceptable range for real-world photovoltaic monitoring applications, confirming that the model is computationally feasible. Overall, the statistical validation and complexity analysis together provide strong evidence supporting the robustness, reliability, and practical applicability of the proposed anomaly detection method.

4.5 Confusion matrix analysis

The ensemble learning model demonstrates robust performance in distinguishing between normal and anomalous data across the majority of power plants, as illustrated in Fig. 9. In Fig. 9 (a), the model achieved 1530 TN, 187 TP, 15 FP and 34 FN has yielded an accuracy of 94.9% corresponding to Plant 1. Similarly, Fig. 9 (b) (Plant 2) shows 1607 TN, 115 TP, 8 FP and 36 FN corresponding to an accuracy of 96.2% with a minimal false-positive rate. The model produced 1535 TN, 158 TP, 20 FP and 53 FN yield as shown in Fig. 9 (c) (Plant 3). It maintains a balanced accuracy of 94.6%.

The advantage of the ensemble approach is further highlighted in Fig. 9 (d) (Plant 4), where the results reached 1567 TN and 158 TP. So, 2 FP and 39 FN are yielded an accuracy exceeding 97% and an exceptionally low false-positive rate of 0.1%. In Fig. 9 (e) (Plant 5), the model recorded 1539 TN, 182 TP, 9 FP and 36 FN corresponding to a detection accuracy of 96.3%. The best overall performance was observed in Plant 6 with 1592 TN, 152 TP, 8 FP and 14 FN as shown in Fig. 9 (f). It achieves an accuracy of 98.2%, precision above 95% and recall above 91%.

Consistent performance was also observed and obtained in 1542 TN, 176 TP, 16 FP and 32 FN as shown in Fig. 9 (g) (Plant 7). It results in an accuracy of 96.9%. Hence, the ensemble model remained robust under more challenging conditions. For instance, Fig. 9 (h) (Plant 8)

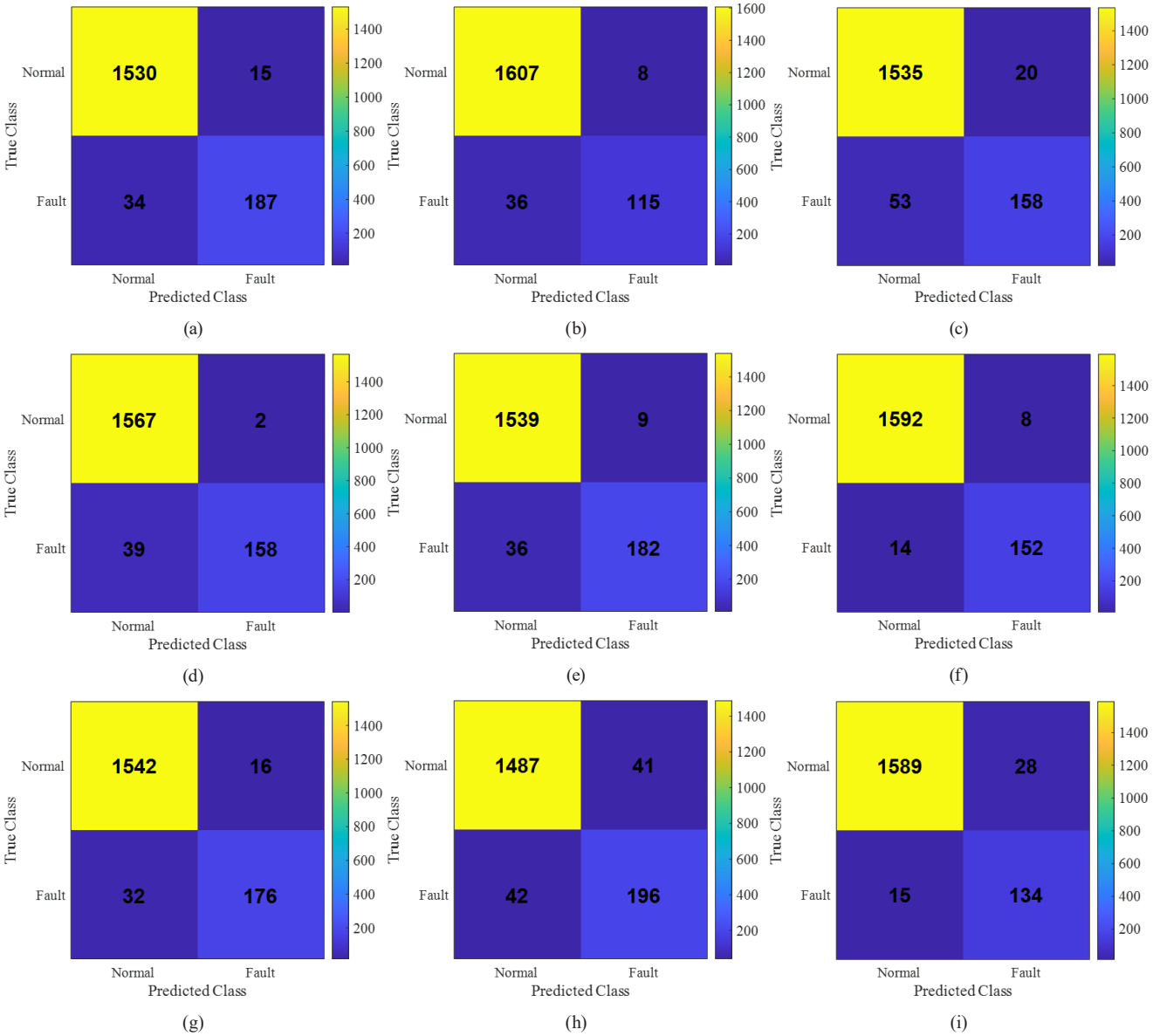

Fig. 9 The confusion matrices for anomaly detection across various PV plants: (a) photovoltaic power station 1, (b) photovoltaic power station 2, (c) photovoltaic power station 3, (d) photovoltaic power station 4, (e) photovoltaic power station 5, (f) photovoltaic power station 6, (g) photovoltaic power station 7, (h) photovoltaic power station 8, (i) photovoltaic power station 9

shows 1487 TN, 196 TP, 41 FP, and 42 FN with accuracy of 94.0% and a low false-positive rate of 2.7%. Finally, it demonstrates 1589 TN, 134 TP, 28 FP, and 15 FN with corresponding to an accuracy of 97.3% as shown in Fig. 9 (i) (Plant 9). It further confirms the stable performance across different operational environments.

Taken together, these results confirm that the ensemble learning methodology consistently achieves high accuracy across diverse PV plants (Fig. 9 (a)–(i)). Moreover, the ensemble approach effectively mitigates overfitting to noise by aggregating the predictions of multiple models. It further addresses class imbalance and enhances both robustness and generalization in anomaly detection.

4.6 Visualization of anomaly distributions

The ensemble learning framework demonstrates robust anomaly detection across most PV plants, effectively distinguishing normal (blue) from anomalous (red) data, as shown in Fig. 10 (a). Normal data cluster in concentrated regions, whereas anomalous points deviate significantly, reflecting clear separability.

In Fig. 10 (b), the distribution of anomalies forms several distinct clusters corresponding to another PV plant. These clusters are likely attributable to specific operational states or environmental influences that introduce notable fluctuations. Similarly, the anomalies are grouped within localized regions as shown in Fig. 10 (c). Further,

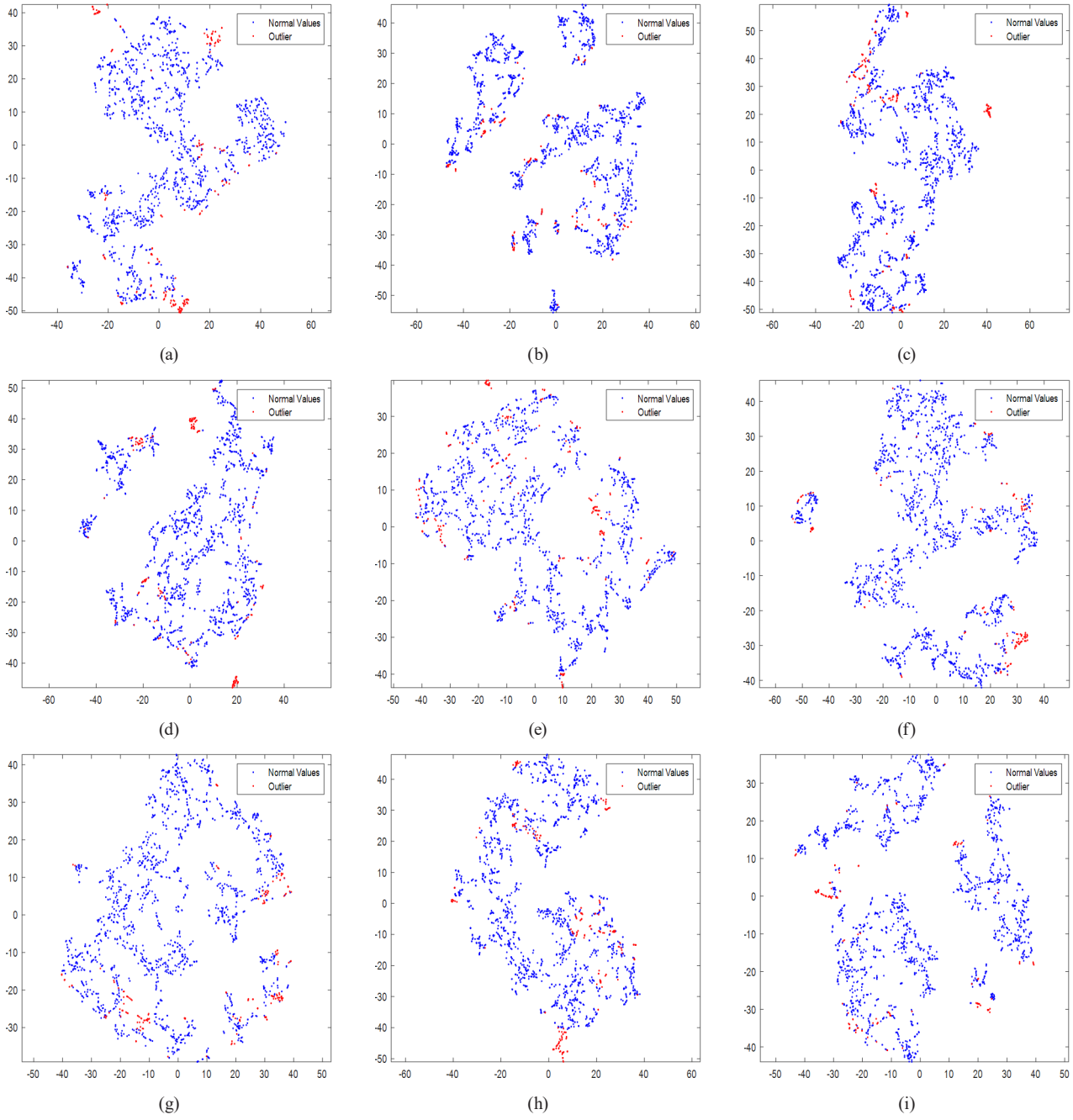

Fig. 10 Ensemble learning anomaly detection outcomes, where red and blue points indicating anomalous and normal values, respectively, for the following cases: (a) photovoltaic power station 1, (b) photovoltaic power station 2, (c) photovoltaic power station 3, (d) photovoltaic power station 4, (e) photovoltaic power station 5, (f) photovoltaic power station 6, (g) photovoltaic power station 7, (h) photovoltaic power station 8, (i) photovoltaic power station 9

it reinforces the ability of the ensemble approach to highlight plant-specific abnormal behaviors.

A clearer separation is observed and the anomalies are sparsely distributed around the periphery of dense normal data regions as shown in Fig. 10 (d). It shows the ensemble model's sensitivity to subtle deviations. Meanwhile, Fig. 10 (e) reveals multiple compact clusters of anomalies. It suggests repeated abnormal operating states under consistent environmental conditions. By contrast, a more dispersed anomaly distribution has been indicated on a broader range of abnormal events and potential variability in detection thresholds as shown in Fig. 10 (f).

The anomalies appear widely scattered across the feature space and yet the ensemble method still effectively distinguishes them from the dense core of normal samples as shown in In Fig. 10 (g). Similarly, as shown in Fig. 10 (h), a hybrid pattern emerges: some anomalies cluster closely together while others remain more dispersed, reflecting complex and heterogeneous abnormal behavior. This is primarily attributed to the significant fluctuations

in meteorological conditions within the station's vicinity. Local cloud cover, rapid changes in wind speed and temperature, and other factors contribute to increased short-term fluctuations in the power generation data from this station, thereby complicating anomaly detection. Finally, a scenario in which anomalous samples are relatively scattered but remain consistently separable from the normal population as shown in Fig. 10 (i). It underscores the robustness of the ensemble approach.

Additionally, variations in sensor equipment and maintenance conditions across different power plants also impact data quality. Noise or calibration errors in some sensors lead to more unstable data collection, causing the distribution of anomaly samples to exhibit a scattered trend. While filtering and smoothing techniques in data preprocessing can mitigate meteorological noise interference to some extent, they struggle to fully eliminate distribution anomalies caused by short-term weather fluctuations. To address this, the proposed ensemble anomaly detection model integrates multiple approaches—including isolated forests, Gaussian mixture models, and box plots—to extract anomalous features from diverse perspectives, thereby enhancing the model's ability to capture complex anomaly patterns. To address long-term changes in photovoltaic systems and environmental conditions—known as concept drift—the model design incorporates a dynamic threshold adjustment mechanism and supports sliding-window-based parameter update strategies. This enables the model to adapt promptly to shifts in data distribution, maintaining stable and robust detection performance.

Taken together, the visualization results confirm that ensemble learning not only enhances anomaly detection precision but also adapts effectively to diverse operational environments as shown in Fig. 10 (a)–(i). By aggregating the decisions of multiple models, the framework ensures robustness against variability in anomaly distribution patterns. Thereby, it achieves reliable identification of potential anomalous samples across heterogeneous PV plants.

4.7 Limitations and future work

Although the anomaly proportion in this study's dataset is 6.22%, reflecting the scarcity of abnormal events in real-world operations, the data imbalance issue still poses challenges for model training and generalization. Future work will explore employing more sample augmentation techniques or cost-sensitive learning methods to enhance anomaly detection capabilities. The current model demonstrates excellent performance in accuracy and robustness, but it exhibits high computational complexity and training time. Subsequent research will focus on model lightweighting and online deployment strategies to reduce resource consumption in practical applications. Furthermore, future plans include extending the proposed method to fault detection in image or multimodal data domains. For instance, integrating infrared thermal imaging and visual sensor data will enrich the input features for anomaly detection, thereby enhancing the model's detection capability and adaptability under complex operating conditions.

5 Conclusions

This study proposes a boosting-based ensemble learning method for anomaly detection in PV power plants. By integrating multiple models through feature fusion and dynamic weighting strategies, it achieves high-precision identification of anomalies in PV data. Experimental results on multi-plant datasets demonstrate that the proposed method attains an accuracy, precision, recall, and F1-score of 0.973, 0.923, 0.911, and 0.887, respectively. This significantly outperforms traditional statistical methods such as Z-score and IQR, as well as single machine learning models like IF and Transformer. Specifically, the Z-score model suffers from distributional assumption limitations, resulting in an F1-score below 0.80; the IF model achieves high recall (>0.95) but has relatively low precision, causing its F1-score to fall below 0.85; the Transformer exhibits instability in certain plant scenarios. In contrast, the proposed ensemble approach maintains high stability, robustness, and generalization across multiple plants, demonstrating its ability to adapt effectively to the complex operating conditions and strong noise inherent in PV systems.

From an engineering perspective, the proposed framework can be directly applied to online monitoring systems of PV plants for screening abnormal power points, early detection of equipment degradation, and supporting operation and maintenance (O&M) decision-making. By reducing false alarms and missed detections, it can lower O&M costs and improve power generation revenue. Moreover, the method is extendable to anomaly diagnosis in other renewable energy scenarios such as wind power and energy storage, providing a technical foundation for intelligent multi-energy monitoring.

Future work will focus on several directions:

1. Model lightweighting — employing pruning, quantization, and knowledge distillation to reduce inference latency and improve deployment efficiency;
2. Feature enhancement and domain adaptation — incorporating heterogeneous features (e.g., weather, component parameters, plant geographical data)

and using transfer learning to improve cross-plant generalization;
3. Online learning and adaptive weighting mechanisms — dynamically updating model weights to quickly respond to changing operating conditions;
4. Multi-source data association analysis — integrating SCADA, IV curve, infrared thermography, and meteorological data to model anomaly propagation mechanisms;
5. Joint detection-scheduling optimization — embedding anomaly detection results into virtual power plant and multi-plant coordinated scheduling frameworks, developing intelligent optimization strategies considering O&M cost constraints to enhance overall system economy and reliability.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No.52467008) and Gansu Provincial Department of Education Youth Doctoral Support Project (2024QB-051).

References

[1] Sánchez-Hernández, G., Jiménez-Garrote, A., López-Cuesta, M., Galván, I. M., Aler, R., Pozo-Vázquez, D. "A novel method for modeling renewable power production using ERA5: Spanish solar PV energy", Renewable Energy, 240, 122120, 2025. https://doi.org/10.1016/j.renene.2024.122120

[2] Yang, S., Hu, X., Tu, Z., Yin, R., Zhang, Y., Li, J., Wang, L., Meng, L. "A Distributed Photovoltaic Aggregation Method Considering Photovoltaic Output Characteristics and Active Voltage Sensitivity", Electronics, 14(5), 917, 2025. https://doi.org/10.3390/electronics14050917

[3] Tina, G. M., Maione, G., Stefanelli, D. "The Impact of Grid-Forming vs. Grid-Following Converters on Frequency Regulation: Comparing Centralised or Distributed Photovoltaic Generation", Energies, 17(23), 5827, 2024. https://doi.org/10.3390/en17235827

[4] Zhao, C., Wang, X., Yao, X., Tian, M. "A background refinement method based on local density for hyperspectral anomaly detection", Journal of Central South University, 25(1), pp. 84–94, 2018. https://doi.org/10.1007/s11771-018-3719-6

[5] Barraz, Z., Sebari, I., Oufettoul, H., Ait El Kadi, K., Lamrini, N., Ait Abdelmoula, I. "A holistic multimodal approach for real-time anomaly detection and classification in large-scale photovoltaic plants", Energy and AI, 21, 100525, 2025. https://doi.org/10.1016/j.egyai.2025.100525

[6] Laguna, G., Moreno, P., Cipriano, J., Mor, G., Gabaldón, E., Luna, A. "Detection of abnormal photovoltaic systems' operation with minimum data requirements based on Recursive Least Squares algorithms", Solar Energy, 274, 112556, 2024. https://doi.org/10.1016/j.solener.2024.112556

[7] Chen, C., Li, Q., Chen, L., Liang, Y., Huang, H. "An improved GraphSAGE to detect power system anomaly based on time-neighbor feature", Energy Reports, 9(Supplement 1), pp. 930–937, 2023. https://doi.org/10.1016/j.egyr.2022.11.116

[8] Hachemi, A. T., Kamel, R. M., Hashem, M., Arif, S., Ebeed, M. "Towards reliable and secure distribution network: Assessing energy management system using integrated renewables energy-driven hydrogen storages", International Journal of Hydrogen Energy, 189, 152140, 2025. https://doi.org/10.1016/j.ijhydene.2025.152140

[9] Yin, H., Zhang, G., Wu, Q., Cui, F., Yan, B., Yin, S., Soltanian, M. R., Thanh, H. V., Dai, Z. "Transfer Learning with Transformer-Based Models for Mine Water Inrush Prediction: A Multivariate Analysis Using Sparse and Imbalanced Monitoring Data", Mine Water and the Environment, 43(4), pp. 707–726, 2024. https://doi.org/10.1007/s10230-024-01011-2

[10] Taghavi, M., Lee, C.-J. "Development of a novel hydrogen liquefaction structure based on liquefied natural gas regasification operations and solid oxide fuel cell: Exergy and economic analyses", Fuel, 384, 133826, 2025. https://doi.org/10.1016/j.fuel.2024.133826

[11] Taghavi, M., Yoon, H.-J., Choi, J.-U., Lee, C.-J. "Innovative Structure of a Liquefied Natural Gas (LNG) Process by Mixed Fluid Cascade Using Solar Renewable Energy, Photovoltaic Panels (PV), and Absorption Refrigeration System", Computer Aided Chemical Engineering, 53, pp. 2071–2076, 2024. https://doi.org/10.1016/B978-0-443-28824-1.50346-X

[12] Taghavi, M., Lee, C.-J. "Development of novel hydrogen liquefaction structures based on waste heat recovery in diffusion-absorption refrigeration and power generation units", Energy Conversion and Management, 302, 118056, 2024. https://doi.org/10.1016/j.enconman.2023.118056

[13] Taghavi, M., Salarian, H., Ghorbani, B. "Thermodynamic and exergy evaluation of a novel integrated hydrogen liquefaction structure using liquid air cold energy recovery, solid oxide fuel cell and photovoltaic panels", Journal of Cleaner Production, 320, 128821, 2021. https://doi.org/10.1016/j.jclepro.2021.128821

[14] Taghavi, M., Salarian, H., Ghorbani, B. "Economic Evaluation of a Hybrid Hydrogen Liquefaction System Utilizing Liquid Air Cold Recovery and Renewable Energies", Renewable Energy Research and Applications, 4(1), pp. 125–143, 2023. https://doi.org/10.22044/rera.2022.11899.1122

[15] Soonmin, H., Taghavi, M. "Solar Energy Development: Study Cases in Iran and Malaysia", International Journal of Engineering Trends and Technology (IJETT), 70(8), pp. 408–422, 2022. https://doi.org/10.14445/22315381/IJETT-V70I8P242

[16] Ebrahimi, A., Ghorbani, B., Taghavi, M. "Novel integrated structure consisting of CO_2 capture cycle, heat pump unit, Kalina power, and ejector refrigeration systems for liquid CO_2 storage using renewable energies", Energy Science & Engineering, 10(8), pp. 3167–3188, 2022. https://doi.org/10.1002/ese3.1211

[17] Ghorbani, B., Salehi, G., Ebrahimi, A., Taghavi, M. "Energy, exergy and pinch analyses of a novel energy storage structure using post-combustion CO_2 separation unit, dual pressure Linde-Hampson liquefaction system, two-stage organic Rankine cycle and geothermal energy", Energy, 233, 121051, 2021. https://doi.org/10.1016/j.energy.2021.121051

[18] Afrouzy, Z. A., Taghavi, M. "Thermo-economic analysis of a novel integrated structure for liquefied natural gas production using photovoltaic panels", Journal of Thermal Analysis and Calorimetry, 145(3), pp. 1509–1536, 2021. https://doi.org/10.1007/s10973-021-10769-4

[19] Yepmo, V., Smits, G., Lesot, M.-J., Pivert, O. "Leveraging an Isolation Forest to Anomaly Detection and Data Clustering", Data & Knowledge Engineering, 151, 102302, 2024. https://doi.org/10.1016/j.datak.2024.102302

[20] Liu, D., Huang, D., Chen, X., Dou, J., Tang, L., Zhang, Z. "An Unsupervised Abnormal Power Consumption Detection Method Combining Multi-Cluster Feature Selection and the Gaussian Mixture Model", Electronics, 13(17), 3446, 2024. https://doi.org/10.3390/electronics13173446

[21] Marcelli, E., Barbariol, T., Sartor, D., Susto, G. A. "Active Learning-based Isolation Forest (ALIF): Enhancing anomaly detection with expert feedback", Information Sciences, 678, 121012, 2024. https://doi.org/10.1016/j.ins.2024.121012

[22] Tufail, S., Iqbal, H., Tariq, M., Sarwat, A. "Multi-variable false data injection attack detection and classification in solar photovoltaic grids with ensemble learning", Engineering Applications of Artificial Intelligence, 160, 111755, 2025. https://doi.org/10.1016/j.engappai.2025.111755

[23] Parvin, M., Yousefi, H., Mohammadi-Ivatloo, B. "Photovoltaic fault detection algorithm using ensemble learning enhanced with deep neural network feature engineering", Results in Engineering, 27, 106491, 2025. https://doi.org/10.1016/j.rineng.2025.106491

[24] Alqaraghuli, O., Ibrahim, A. "Optimizing Photovoltaic System Diagnostics: Integrating Machine Learning and DBFLA for Advanced Fault Detection and Classification", Electronics, 14(8), 1495, 2025. https://doi.org/10.3390/electronics14081495

[25] Venkatesh, S. N., Jeyavadhanam, B. R., Sizkouhi, A. M. M., Esmailifar, S. M., Aghaei, M., Sugumaran, V. "Automatic detection of visual faults on photovoltaic modules using deep ensemble learning network", Energy Reports, 8, pp. 14382–14395, 2022. https://doi.org/10.1016/j.egyr.2022.10.427

[26] Ma, S., Huang, Y., Liu, Y., Liu, H., Chen, Y., Wang, J., Xu, J. "Big data-driven correlation analysis based on clustering for energy-intensive manufacturing industries", Applied Energy, 349, 121608, 2023. https://doi.org/10.1016/j.apenergy.2023.121608

[27] Nsaif, Y. M., Hossain Lipu, M. S., Hussain, A., Ayob, A., Yusof, Y. "Island detection for grid connected photovoltaic distributed generations via integrated signal processing and machine learning approach", International Journal of Electrical Power & Energy Systems, 154, 109468, 2023. https://doi.org/10.1016/j.ijepes.2023.109468

[28] Wang, H., Sui, L., Bian, J., Yu, H., Li, G. "Integrated operation risk assessment of distribution network based on improved subjective and objective combination weighting and ISODATA", Electric Power Systems Research, 233, 110469, 2024. https://doi.org/10.1016/j.epsr.2024.110469

[29] Liu, D., Zhen, H., Kong, D., Chen, X., Zhang, L., Yuan, M., Wang, H. "Sensors Anomaly Detection of Industrial Internet of Things Based on Isolated Forest Algorithm and Data Compression", Scientific Programming, 2021(1), 6699313, 2021. https://doi.org/10.1155/2021/6699313

[30] Chen, Y., Zhao, Q., Lu, L. "Combining the outputs of various k-nearest neighbor anomaly detectors to form a robust ensemble model for high-dimensional geochemical anomaly detection", Journal of Geochemical Exploration, 231, 106875, 2021. https://doi.org/10.1016/j.gexplo.2021.106875

[31] Carletti, M., Terzi, M., Susto, G. A. "Interpretable Anomaly Detection with DIFFI: Depth-based feature importance of Isolation Forest", Engineering Applications of Artificial Intelligence, 119, 105730, 2023. https://doi.org/10.1016/j.engappai.2022.105730

[32] Xiong, Z., Zhu, D., Liu, D., He, S., Zhao, L. "Anomaly Detection of Metallurgical Energy Data Based on iForest-AE", Applied Sciences, 12(19), 9977, 2022. https://doi.org/10.3390/app12199977

[33] Liu, X., Zhu, S., Yang, F., Liang, S. "Research on unsupervised anomaly data detection method based on improved automatic encoder and Gaussian mixture model", Journal of Cloud Computing, 11(1), 58, 2022. https://doi.org/10.1186/s13677-022-00328-z

[34] Yu, B., Zhang, Y., Xie, W., Zuo, W., Zhao, Y., Wei, Y. "A Network Traffic Anomaly Detection Method Based on Gaussian Mixture Model", Electronics, 12(6), 1397, 2023. https://doi.org/10.3390/electronics12061397

[35] Ding, N., Ma, H., Gao, H., Ma, Y., Tan, G. "Real-time anomaly detection based on long short-Term memory and Gaussian Mixture Model", Computers & Electrical Engineering, 79, 106458, 2019. https://doi.org/10.1016/j.compeleceng.2019.106458

[36] Sepúlveda-Oviedo, E. H., Travé-Massuyès, L., Subias, A., Pavlov, M., Alonso, C. "An ensemble learning framework for snail trail fault detection and diagnosis in photovoltaic modules", Engineering Applications of Artificial Intelligence, 137, 109068, 2024. https://doi.org/10.1016/j.engappai.2024.109068