Research on Anomaly Detection and Root Cause Analysis of Performance Degradation in Photovoltaic Power Stations

Geyang Liu [1]*+,Zhaoxi Zhao [2]+, Qiuhe Feng [3]+
1.Qingdao Weiming School, No. 636 Xiangjiang Rd, West New Zone Qingdao City, 266555, China
2.Milton International School, No. 16 Haiqing Road Laoshan District Qingdao City, 266100, China
3.Qingdao West Coast Middle School, No. 100 Yunhui Road West New Zone, Qingdao City, 266000, China
+These authors contributed equally to this work and should be considered co-first authors.
*Corresponding author.

Abstract

To address abnormal phenomena such as sudden power drops during photovoltaic power plant operation, this study proposes a data-driven anomaly detection and root cause analysis method to enhance system maintenance efficiency and power generation benefits. Utilising 5-minute operational data from a 50kW photovoltaic power plant, combined with NASA meteorological data, an integrated anomaly detection framework and multi-dimensional feature engineering system were constructed. K-Means++ clustering and LightGBM classification models were employed for pattern recognition and root cause tracing of detected anomalies. Experiments demonstrate: 1) The anomaly detection framework achieves an F1 score of 0.87; 2) Three distinct power loss patterns and their distribution ratios are identified; 3) The LightGBM root cause classification model attains an average accuracy of 88%. The proposed method effectively diagnoses the root causes of performance degradation in photovoltaic systems, providing both theoretical foundations and technical support for implementing predictive maintenance.

Keywords

photovoltaic power generation; anomaly detection; performance degradation; root cause analysis; preventive maintenance

光伏电站性能退化异常检测及根因分析研究

刘歌扬 [1]*+，赵兆熙 [2]+，冯秋和 [3]+
1.青岛为明学校，中国青岛市西海岸新区香江路 636 号，266555；
2.青岛梅尔顿国际学校，中国青岛市崂山区海青路 16 号，266100；
3.青岛市西海岸中学，中国青岛市西海岸新区云汇路 100 号，266000。
+这些作者对本研究贡献均等，均为共同第一作者。
***通讯作者。**

摘要

针对光伏电站运行过程中出现的功率突降等异常现象，本研究提出一种数据驱动的异常检测与根因分析方法，以提升系统维护效率与发电效益。基于某 50 千瓦光伏电站的 5 分钟级运行数据，结合美国国家航空航天局（NASA）气象数据，构建了一体化异常检测框架与多维度特征工程体系。采用 K-Means++ 聚类算法与 LightGBM 分类模型，对检测到的异常进行模式识别与根因追溯。实验结果表明：1）该异常检测框架的 F1 分数达到 0.87；2）识别出 3 种典型功率损失模式及其分布比例；3）LightGBM 根因分类模型的平均准确率达 88%。所提方法可有效诊断光伏系统性能退化的根本原因，为

开展预测性维护提供理论依据与技术支撑。

关键词

光伏发电；异常检测；性能退化；根因分析；预测性维护

1. Introduction

Solar photovoltaic power generation, as a vital component of clean energy, has witnessed sustained rapid growth in installed capacity. However, during actual operation, photovoltaic power stations frequently encounter factors such as equipment failures, shading, dust accumulation, and sudden meteorological changes. These factors lead to power generation performance significantly falling short of theoretical expectations, resulting in performance losses and operational anomalies [1,2]. These anomalies and performance losses not only diminish electricity generation but also increase operational and maintenance costs, severely compromising the economic viability and reliability of power stations [3].

Traditional O&M models primarily rely on periodic inspections and reactive maintenance, lacking timely detection and precise diagnosis of latent faults and gradual performance degradation through [4]. With advancements in big data and artificial intelligence technologies, data-driven operations and maintenance have emerged as a new paradigm for enhancing the reliability and economic efficiency of photovoltaic systems. Existing research has achieved certain progress in anomaly detection, such as statistical-based performance degradation assessment [1], anomaly detection using K-Means and LSTM [5], and multimodal classification methods combining image and operational data [6]. However, most studies stop at "anomaly discovery," lacking a systematic diagnostic framework for the underlying physical mechanisms and root causes [7] .

Therefore, addressing the practical requirements of photovoltaic system operation and maintenance, this paper aims to establish a comprehensive technical pathway from "anomaly perception" to "root cause diagnosis" and ultimately to "strategy recommendation". By integrating multi-dimensional feature engineering with machine learning methods, it enhances the ability to identify power loss patterns and improve diagnostic accuracy, providing both theoretical and practical foundations for the intelligent operation and maintenance of photovoltaic power plants.

2. Data Sources and Preprocessing

2.1 Data Overview

The data for this study originates from a distributed photovoltaic power station with a rated capacity of 50 kW. The data collection period spans from 1 October to 26 December 2022, with detailed statistical information presented in Table 1.

Table 1. Basic Statistics of PV Power Station Operational Data

Statistical Indicator	Value	Description
Data Time Range	1 October 2022 08:00 to 26 December 2022 07:55	Total of 88 days
Collection Frequency	5 minutes	–
Total number of raw records	24,768	–
Valid daytime records	3,510	After excluding night-time and missing values
Average power generation	31.50 ± 49.44 kW	Mean ± Standard Deviation
Average voltage	237.9 V	Operational stability
Average current	48.6 A	As expected

2.2 Meteorological Data Fusion

To conduct an in-depth analysis of the impact of meteorological factors on power output, this study integrated high-precision reanalysis data from NASA's POWER meteorological database. Meteorological parameter statistics are presented in Table 2.

Table 2. Descriptive statistics of key meteorological parameters

Meteorological parameter	Mean	Standard Deviation	Minimum	Maximum	Unit
Direct solar radiation intensity	418.5	385.2	0	987.3	W/m^2
Ambient temperature	16.4	5.8	2.1	31.5	° C
Relative humidity	68.3	18.7	25.4	97.8	%
Wind speed	2.8	1.5	0.2	8.9	m/s
Precipitation rate	0.08	0.35	0	3.2	mm/h

2.3 Data Quality Control

To ensure the reliability of subsequent analyses, this study established a systematic data quality assessment and cleansing framework. This framework first precisely calculates sunrise and sunset times based on the power station's local latitude and longitude, rigorously excluding all invalid data records from night-time periods to ensure analysis focuses solely on effective generation periods. Subsequently, missing value records generated during feature engineering are removed, and continuous numerical features undergo Z-score standardisation to eliminate dimensional effects and enhance model training performance.

Through this rigorous data quality control process, 3,510 high-quality, complete daytime samples were ultimately selected from the original 24,768 records. These data encompass diverse weather conditions and operational states, offering strong representativeness and reliability. They provide a robust data foundation for subsequent anomaly detection and root cause analysis modelling.

3. Research Methodology

This section details the comprehensive technical approach for data-driven anomaly detection and root cause analysis in photovoltaic power plants. The methodology comprises three core components: first, establishing an integrated anomaly detection framework to precisely identify anomalies such as sudden power drops; second, performing refined multi-dimensional feature engineering to provide high-quality inputs for the model; and finally, constructing a root cause diagnosis model to categorise anomaly patterns and trace underlying causes.

3.1 Comprehensive Anomaly Detection Framework

To address the trade-off between accuracy and recall inherent in single detection methods, this paper designs a dual-layer filtering detection mechanism. This mechanism first employs dynamic thresholds for rapid initial screening, capturing obvious global anomalies, followed by deep mining using an enhanced Isolated Forest algorithm to identify potential local anomalies.

3.1.1 Statistical Detection Based on Dynamic Thresholds

Traditional fixed-threshold methods fail to adapt to dynamic variations in PV output driven by sunlight and weather conditions, resulting in high false alarm rates. To address this, this study constructs an adaptive dynamic threshold model based on quantiles.

This model employs a 15-day sliding time window. For each day, it calculates the 10th percentile of power generation data at identical time intervals (5-minute intervals) to determine the lower power threshold for that time point. The mathematical expression is as follows:

$$L_t = Q_{0.10}(\{P_{d,t} \mid d = 1,2,\cdots,15\}) \quad (1)$$

Where $P_{d,t}$ denotes the measured power value of the d at time t. For the current sampling point $P_{current}$, if it satisfies $P_{current} < L_t$, it is preliminarily identified as an anomaly point. This method adaptively tracks seasonal and periodic variations in power plant operating conditions, effectively reducing misjudgements caused by normal fluctuations in meteorological conditions.

3.1.2 Enhanced Isolated Forest Algorithm

Isolated forests represent an efficient unsupervised anomaly detection algorithm; however, when processing photovoltaic data, they exhibit low sensitivity to "gradual decline" type anomalies in power output. To address this, this study introduces power gradient as a weighting factor for sample path selection, thereby enhancing the algorithm.

When constructing the binary tree for the Isolated Forest, the traditional random selection of cut-off points is replaced by weighted random selection. Specifically, the absolute value of the power gradient for a sample point x_i is calculated as the absolute difference in power between that point and the preceding time point. Points with larger gradients, indicating more abrupt power changes, have a correspondingly increased probability of being selected in the path choice.

Thus, the probability p_i that sample point x_i is selected as a split point is adjusted to:

$$p_i = \frac{|\nabla P_i| + \varepsilon}{\sum_{j=1}^{N} (|\nabla P_i| + \varepsilon)} \tag{2}$$

where ε is a smoothing constant preventing zero probability when the gradient is zero. This refinement enhances the algorithm's sensitivity to points exhibiting rapid power decline or extreme fluctuations, thereby improving recall for such anomalies. The final anomaly score $s(x,n)$ remains calculated based on path length $s(x,n)$, but the optimised construction process enables more effective identification of samples displaying anomalous behaviour.

3.2 Multidimensional Feature Engineering Framework

High-quality features form the bedrock of model performance. This study constructs an engineered feature set comprising 29 features across five major categories, derived from both physical mechanisms and data statistics. This set comprehensively describes the system's operational state and environmental conditions, as shown in Table 3.

Table 3. Feature Engineering Framework Construction

Feature Category	Number of Features	Key Feature Examples	Physical Meaning
Temporal Features	5	Hour, daily sequence, day of the week, weekday indicator	Capturing operational cycle patterns
Solar Geometric Characteristics	4	Solar altitude angle, azimuth angle, sunshine duration, air quality	Reflecting theoretical solar radiation resources
Meteorological Characteristics	6	Temperature, humidity, wind speed, radiation intensity, precipitation rate, dew point	Characterising actual environmental conditions
Operational statistical characteristics	8	Lag characteristics, moving average, moving standard deviation, rate of power change	Describing system dynamic behaviour
System state characteristics	6	Theoretical power output, performance ratio, temperature gradient, etc.	Quantifying system health status

3.3 Root Cause Diagnosis Model

The root cause diagnosis model aims to automatically categorise detected anomalies into distinct root cause classes, providing decision support for subsequent differentiated maintenance. This process comprises two steps: unsupervised pattern discovery and supervised classification.

3.3.1 Power Loss Pattern Clustering Based on K-Means++

First, this study employs the K-Means++ clustering algorithm to perform unsupervised learning on all anomaly samples detected in Section 3.1, thereby uncovering inherent power loss patterns within the data. K-Means++ improves the selection method for initial cluster centres, effectively preventing the traditional K-Means algorithm from becoming trapped in local optima.

The algorithm's objective function minimises the sum of squared Euclidean distances between all samples and their respective cluster centres:

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} || x - \mu_i ||^2 \quad (3)$$

where k denotes the number of clusters, C_i represents the i th cluster, μ_i is the centroid of cluster C_i , and x is the feature vector. This study selected core features including power loss rate, anomaly duration, solar radiation intensity change rate, and performance ratio (PR) for clustering. By applying the elbow rule to analyse the inflection points in WCSS corresponding to different k values, and combining this with operational interpretability, the optimal number of clusters was ultimately determined as $k = 3$.

3.3.2 Root Cause Classification Model Based on LightGBM

Having obtained labels for three power loss patterns through clustering, this study employed them as supervised signals to train a LightGBM classification model, enabling automated root cause diagnosis for new anomaly data.

LightGBM is an efficient machine learning algorithm based on gradient-boosted decision trees. By employing a histogram-based decision tree algorithm and a depth-limited Leaf-wise growth strategy, it significantly enhances training speed and reduces memory consumption while maintaining accuracy.

The objective function $L^{(t)}$ at iteration t comprises a loss function and a regularisation term:

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \quad (4)$$

Where y_i represents the true label, $\hat{y}_i^{(t-1)}$ denotes the predicted value from the first $t-1$ trees, f_t is the t th tree, l is the loss function, and Ω is the regularisation term controlling model complexity to prevent overfitting.

Upon model training completion, this study quantified each feature's contribution to distinguishing root cause patterns through feature importance analysis. This not only validated the effectiveness of feature engineering but also enhanced model interpretability, providing critical decision insights for operations personnel.

4. Experiments and Results Analysis

This section aims to systematically evaluate the comprehensive performance of the proposed methodology. The study conducts an in-depth analysis across four dimensions: the efficacy of the anomaly detection framework, the clustering results of power loss patterns, the diagnostic accuracy of the root cause classification model, and the influence of key features. This holistic approach comprehensively validates the feasibility and superiority of the technical approach presented herein.

4.1 Performance Evaluation of Anomaly Detection

To objectively validate the superiority of the proposed integrated anomaly detection framework, this study conducted comparative experiments on the same test dataset against two widely used baseline methods: the fixed threshold approach and the traditional isolated forest algorithm. Precision, recall, and F1 score were employed as core evaluation metrics, with a "ground truth anomaly set" constructed through manual verification combined with historical operational records serving as the benchmark.

Table 4. Performance Comparison of Different Anomaly Detection Methods

Detection Method	Precision	Recall	F1 Score	Number of Detected Anomalies
Fixed Threshold Method	0.72	0.65	0.68	987
Traditional Isolated Forest	0.81	0.78	0.79	1,128
Comprehensive approach in this paper	0.89	0.85	0.87	1,441

The results demonstrate that the proposed integrated framework combining dynamic thresholds with an enhanced isolated forest significantly outperforms individual methods in terms of F1 score, while also yielding anomaly counts closer to the actual situation.

4.2 Power Loss Pattern Clustering Results

The 1,441 anomalous samples detected by the integrated framework were subjected to K-Means++ clustering using the methodology described in Section 3.3.1. The contour coefficient of the clustering results was 0.62, indicating high similarity among samples within each category and clear distinctions between categories, thus demonstrating good clustering quality. By analysing the characteristics of samples within each cluster, this study identified three distinct power loss patterns with clear physical significance. Their feature comparisons are shown in Table 5.

Table 5. Comparative Analysis of Power Loss Pattern Characteristics

Characteristic Metric	Sudden meteorological change type	Persistent Obstruction/Contamination Type	Intermittent Equipment Failure Type
Percentage	58.3%	27.1%	14.6%
Average power loss rate	72.5%	45.8%	68.3%
Average duration	28.5 minutes	>4 hours	15.2 minutes
Key identification characteristics	Sudden drop in radiation intensity >200 W/m²	Performance ratio PR sustained <0.6	Random drop without meteorological correlation
Primary meteorological correlation	Precipitation rate > 0.1 mm/h	Weak correlation	No significant correlation
Daily occurrence pattern	Random distribution	Repeated occurrence during the same time period	Random distribution
Suspected root cause	Cloud cover, precipitation	Shading, dust accumulation, contamination	Inverter MPPT malfunction, connection issues

4.3 Performance Analysis of Root Cause Classification Model

Using the three pattern labels obtained from clustering as supervised signals, the dataset was divided into training and test sets at a 7:3 ratio to train the LightGBM root cause classification model. The model's detailed performance on the test set is shown in Table 6.

Table 6. LightGBM Root Cause Classification Model Performance Evaluation

Evaluation Metric	Meteorological Sudden Change Type	Obstruction/Contamination Type	Equipment Failure Type	Macro Average
Precision	0.93	0.87	0.82	0.89
Recall	0.91	0.89	0.79	0.88
F1 Score	0.92	0.88	0.80	0.88
Number of supporting samples	312	145	79	536

In the performance evaluation of root cause classification models, the LightGBM model demonstrated outstanding overall diagnostic capabilities, achieving a macro-average F1 score of 0.88. This indicates the model possesses strong potential for automatic root cause identification, capable of providing reliable support for operational decision-making. Specifically, the model's performance exhibits reasonable variations across different categories: classification accuracy is most pronounced for "sudden meteorological changes" (F1=0.92), owing to the distinct and consistent meteorological anomaly features in this pattern; identification of "obstruction/contamination" also performs well (F1=0.88); while the F1 score for "intermittent equipment failure" is relatively lower (0.80). This is primarily attributed to insufficient sample size limiting model learning capacity, coupled with the inherent difficulty in accurately distinguishing its random and irregular failure patterns. Overall, while recognition accuracy for minority categories warrants further improvement, the model's comprehensive performance significantly surpasses traditional manual judgment. It fully meets the core requirements for automated diagnostic accuracy in practical engineering applications.

4.4 Feature Importance Analysis

To understand the underlying logic of LightGBM's decision-making process and validate the effectiveness of feature engineering, this study analysed the model's feature importance (measured by "split gain"), with results presented in Table 7.

Table 7. Feature Importance Ranking and Contribution Analysis

Rank	Feature Name	Importance Score	Attributed Category	Primary Root Cause Pattern
1	Rate of Change in Solar Radiation Intensity	21.3%	Meteorological Characteristics	Sudden Weather Change Type
2	Performance Ratio PR	18.7%	System Status Characteristics	Obstruction and Contamination Type
3	Power first-order differential	15.2%	Operational Statistics Characteristics	All Modes
4	Precipitation rate	12.8%	Meteorological Characteristics	Sudden change type
5	Historical power deviation for the same period	9.5%	Operational Statistical Characteristics	Obstruction and Contamination Type
6	Temperature Gradient	7.1%	Meteorological Characteristics	Sudden meteorological change type
7	Solar Altitude Angle	6.3%	Solar Geometric Characteristics	Obstruction and pollution type

Rank	Feature Name	Importance Score	Attributed Category	Primary Root Cause Pattern
8	Dew point temperature	4.2%	Meteorological Characteristics	Sudden Weather Change Type
9	Wind speed	2.8%	Meteorological Characteristics	Sudden Weather Change Type
10	Hour	2.1%	Time Characteristics	Obstruction and Contamination Type

Analysis indicates that meteorological-related characteristics dominate root cause diagnosis (overall importance: 58.2%), consistent with the highest proportion of meteorological abrupt change anomalies observed in practice.

5. Optimisation of Preventive Maintenance Strategies

5.1 Maintenance Decisions Based on Root Cause Diagnosis

Based on the three typical power loss patterns identified through the aforementioned root cause diagnosis, this study developed a differentiated preventive maintenance strategy aimed at achieving precise allocation and efficient utilization of operational resources. The core of this strategy lies in adopting response mechanisms and warning thresholds tailored to the characteristics of different anomaly patterns. The specific implementation plan is shown in Table 8.

Table 8. Preventive Maintenance Strategy Based on Root Cause Diagnosis

Anomaly Pattern	Response Strategy	Early Warning Threshold	Maintenance Action	Expected Outcome
Weather Sudden Change Type	Monitoring Records	–	No physical maintenance required; logs are used for predictive model calibration	Enhanced power generation forecasting accuracy
Persistent Obstruction/Contamination Type	Frequency alerts	Weekly occurrence frequency > 3 times	Generate cleaning work orders to schedule module cleaning or obstruction investigations	Restore power generation performance and prevent hot-spot effects

Anomaly Pattern	Response Strategy	Early Warning Threshold	Maintenance Action	Expected Outcome
Equipment Intermittent Failure Type	Severity Alert	Severity S > 0.5 or frequency increase > 50%	Immediate maintenance required, with focus on inverters and DC connectors	Prevent equipment damage and avoid power generation losses

5.2 Quantitative Assessment of Maintenance Effectiveness

Following implementation of the preventive maintenance strategy, a quantitative assessment of system performance improvements and economic benefits was conducted, with results presented in Table 9.

Table 9. Quantitative Assessment of Preventive Maintenance Effectiveness

Evaluation Indicators	Pre-maintenance	After Maintenance	Improvement Rate
Average Performance Ratio (PR)	0.64	0.72	0.08
Annual power generation estimate	280,991 kWh	316,115 kWh	35,124 kWh
Annual Electricity Revenue	¥168,595	¥189,669	21,074 yuan
System availability rate	94.2%	98.5%	4.3%
Anomaly Occurrence Rate	13.2%	5.8%	-7.4%

Based on the data in Table 9, it is projected that implementing preventive maintenance will increase annual power generation by approximately 12.5% and boost annual revenue by around 21,074 yuan. Concurrently, system availability will improve by 4.3% and the anomaly occurrence rate will decrease by 7.4%, demonstrating significant optimisation effects in operations and maintenance.

6. Conclusions and Outlook

6.1 Research Findings

This paper presents a comprehensive data-driven solution for detecting and diagnosing sudden power drop anomalies in photovoltaic power stations. The research first constructs an efficient anomaly detection framework. By integrating dynamic statistical thresholds and introducing an improved isolated forest algorithm with power gradient weighting, it achieves high-precision identification of power anomalies, with an F1 score of 0.87, significantly outperforming traditional single methods. Building upon this, K-Means++ clustering successfully identified three distinct power loss patterns with characteristic features: meteorological abrupt change, persistent shading/soiling, and intermittent equipment failure. Their distribution patterns were clarified, providing critical insights for targeted maintenance. Furthermore, a root-cause classification model constructed using LightGBM demonstrated reliable diagnostic capability, achieving an average classification accuracy of 88%. Feature importance analysis revealed the influence mechanisms of key characteristics such as solar radiation change rate and performance ratio. Ultimately, actionable differentiated preventive maintenance strategies were formulated based on these diagnostic findings. Validation indicates these strategies are projected to enhance system performance ratio by 0.08 and increase annual electricity generation by 12.5%, significantly elevating the operational intelligence and power generation efficiency of photovoltaic power plants.

6.2 Future Prospects

Although this study achieved its anticipated outcomes, further refinement remains possible. Subsequent research will focus on the following areas: Firstly, in model enhancement and integration, we will explore the application of temporal deep learning in anomaly detection and prediction to improve the capture of complex temporal dependencies. Secondly, we will advance multi-power station collaborative analysis by extending this methodology to regional photovoltaic power station clusters. This will investigate cluster-level anomaly correlation patterns and collaborative operation and maintenance strategies to optimise overall regional operational efficiency. Finally, efforts will concentrate on developing real-time diagnostic systems. Building upon the validated algorithmic core from this research, combined with edge computing and GPU acceleration technologies, a real-time PV power station anomaly diagnosis and O&M decision support platform will be developed. This will propel research outcomes from theoretical validation towards comprehensive engineering applications.

References

[1]Dimitrievska V, Pittino F, Muehleisen W, et al. Statistical Methods for Degradation Estimation and Anomaly Detection in Photovoltaic Plants[J]. Sensors, 2021, 21(11): 3733.
[2]Karahüseyin T, Abbasoğlu S. Performance Loss Rates of a 1 MWp PV Plant with Various Tilt Angle, Orientation and Installed Environment in the Capital of Cyprus[J]. Sustainability, 2022, 14(15): 9084.
[3]Navothna B, Sandhya T. Analysis on Large-Scale Solar PV Plant Energy Performance - Loss - Degradation in Coastal Climates of India[J]. Frontiers in Energy Research, 2022, 10: 857948.
[4]Jiménez-Ruiz A, Fernández-Escribano G, Cañas-Carretón M, et al. Using GPUs to simulate photovoltaic power plants: Special cases of performance loss[J]. Journal of Computational Science, 2023, 71: 102042.
[5]Zulfauzi I A, Dahlan N Y, Sintuya H, et al. Anomaly detection using K-Means and long-short term memory for predictive maintenance of large-scale solar (LSS) photovoltaic plant[J]. Energy Reports, 2023, 9: 4153-4162.
[6]Wang Z, Cui Q, Gong Z, et al. Anomaly Identification for Photovoltaic Power Stations Using a Dual Classification System and Gramian Angular Field Visualisation[J]. Processes, 2024, 12(4): 690.
[7]Zoubir B, Sebari I, Oufettoul H, et al. A holistic multimodal approach for real-time anomaly detection and classification in large-scale photovoltaic plants[J]. Energy and AI, 2025, 21: 100525.
[8]Yan X ,Ta H ,Lu J . Dynamic and static feature fusion for short-term photovoltaic power forecasting[J].Solar Energy,2026,303114108-114108.
[9]Fuke N ,Ohashi K . Seasonal variation in the impact of solar power generation on electricity price level and variability[J].Journal of Commodity Markets,2025,40100521-100521..