Improving Solar Energy Reliability with Data-Driven Anomaly Detection Techniques

Zakiyyan Zain Alkaf[1,*], Bhre Wangsa Lenggana[2], A'isya Nur Aulia Yusuf[3], Elsa Sari Hayunah Nurdiniyah[3], Tri Wisudawati[1], Ameliyana Rizky Syamara Putri Akhmad Yani[2]

[1]Department of Industrial Engineering, Faculty of Engineering, Universitas Jenderal Soedirman, Indonesia

[2]Department of Mechanical Engineering, Faculty of Engineering, Universitas Jenderal Soedirman, Indonesia

[3]Department of Electrical Engineering, Faculty of Engineering, Universitas Jenderal Soedirman, Indonesia

Received 07 December 2025; received in revised form 09 February 2026; accepted 11 February 2026

DOI: https://doi.org/10.46604/aiti.2026.15951

Abstract

This study investigates unsupervised machine learning (ML) for anomaly detection in solar photovoltaic (PV) power generation data from 2019 to 2023. An unsupervised approach is selected to overcome the absence of pre-labeled fault data, enabling the autonomous identification of operational patterns. Following data preparation, K-means clustering (k=3) identifies distinct operational patterns, specifically characterizing regimes such as optimal performance (Cluster 2) and low energy output attributed to adverse weather conditions (Cluster 1). These clusters are subsequently visualized using principal component analysis (PCA) to validate their distinct separation. An isolation forest model is then employed for anomaly detection, identifying 17 significant deviations. These anomalies occur most frequently in 2020, coinciding with the COVID-19 pandemic period. Many fall outside the typical energy range of 2.0–3.2 kWh/day and are associated with non-ideal weather conditions. This finding demonstrates that unsupervised ML provides a scalable framework for monitoring PV system health, enhancing reliability, and supporting preventive strategies.

Keywords: solar photovoltaic (PV) systems, anomaly detection, K-means clustering, isolation forest, renewable energy monitoring.

1. Introduction

The increasing need to mitigate climate change and reduce global reliance on finite fossil fuels drives the global transition toward sustainable energy systems. Renewable energy sources—including solar, wind, hydro, and biomass—emerge as cornerstones of this transformation, given their capacity to provide clean, low-carbon electricity at scale. Governments, industries, and international organizations respond by implementing supportive policies, offering financial incentives, and investing in research and development to accelerate the deployment of these technologies. Among the various forms of renewable energy, solar power achieves particular prominence due to its inherent scalability, continuously declining costs, and broad geographic applicability. Solar energy is primarily harvested through multiple technologies, with photovoltaic (PV) systems being the most widely adopted [1]. These systems directly convert sunlight into electricity, offering a flexible solution applicable in both grid-connected and off-grid environments. PV technology is integrated into diverse applications, ranging from residential rooftops to utility-scale solar farms.

As the global deployment of PV systems rapidly expands, the necessity for reliable monitoring and data-driven performance assessment becomes increasingly critical. PV plants generate large volumes of operational data, encompassing

* Corresponding author. E-mail address: zakiyyan.alkaf@unsoed.ac.id

energy output and environmental indicators, which are essential for assessing performance and detecting system faults [2]. Recent studies emphasized the critical role of simulation-based comparative analysis and material parameter investigation in optimizing the operational reliability of photovoltaic modules [3-4]. Monitoring the operational performance of PV systems relies heavily on the continuous recording and analysis of sensor and meter data streams. These data capture valuable information on the system's energy output, efficiency, and overall health. However, raw data collected from real-world PV plants are often incomplete, noisy, or contain anomalies that obscure meaningful patterns [5]. This underscores the importance of data preprocessing and anomaly detection as prerequisite steps for robust data-driven analysis in renewable energy systems.

Anomalies in PV plant data can arise from a wide range of sources, including equipment degradation, sensor malfunctions, unexpected weather events, or human intervention (e.g., maintenance). Detecting these anomalies is paramount, as they may signal inefficiencies, potential safety hazards, or crucial opportunities for system optimization. Traditional rule-based systems often prove insufficient due to the complex, non-linear, and dynamic nature of PV operations, and they struggle to adapt to changing environmental and operational conditions over time [6]. To circumvent these limitations, recent studies have increasingly leverage machine learning (ML) techniques, which offer a more flexible and scalable approach to pattern recognition and fault detection [7]. Specifically, unsupervised learning methods are exceptionally well-suited to PV plant data, which typically lacks pre-labeled fault instances or established ground truth. Existing literature often treats anomalies primarily as noise to be filtered out to improve forecasting accuracy [7]. In contrast, this study treats anomalies as critical diagnostic signals that focus on classifying deviations to trigger preventive maintenance.

Although ML applications in PV anomaly detection yield promising results, a gap remains in effectively interpreting these anomalies within noise-heavy operational data. Many studies focus on algorithmic accuracy but fail to clearly link detected anomalies back to underlying environmental factors. This study bridges this gap by contrasting clustering-based methods with isolation-based methods. The primary novelty lies in analyzing how combining these distinct methodological assumptions allows operators to distinguish between weather-induced low production (contextual anomalies) and true system faults (point anomalies). This approach offers a level of operational interpretability often missing in "black-box" anomaly removal techniques.

The remainder of this manuscript is structured into three main sections: Section 2, Methodology, outlines the research design, detailing the data preparation steps, and the model training and evaluation process. Section 3, Results and Discussion, presents data insights via the correlation matrix heatmap, followed by the interpretation of the operational patterns identified through K-means clustering (visualized via principal component analysis). Finally, Section 4, Conclusion, summarizes the key findings and offers implications for system reliability and future work.

2. Methodology

The methodology is structured into three distinct subsections to ensure a systematic analysis. Research design establishes the study's quantitative exploratory framework, justifying the selection of unsupervised ML—specifically K-means clustering and isolation forest—to address the absence of labeled ground truth. Subsequently, Data preparation details the acquisition and processing of historical inputs from nine PV plants, focusing on data sanitization, normalization via Specific Energy, and the engineering of rolling temporal features to capture system dynamics. Finally, Model training and evaluation describe the configuration and validation of the algorithms, delineating the criteria used to optimize cluster separation (k=3) and calibrate the anomaly detection threshold (-0.05) for maximum reliability.

2.1. Research design

This research employs a quantitative exploratory design grounded in data-driven analysis of historical operational data from a PV power plant. The quantitative nature of the study is reflected in the systematic collection, preprocessing, and

computational analysis of time-series data generated by the PV system. An unsupervised learning approach is adopted primarily due to the absence of pre-labeled fault instances and established ground truth within the PV operational dataset, which renders supervised or semi-supervised methodologies impractical. By leveraging unsupervised techniques such as K-means clustering and isolation forest, this study autonomously identifies inherent patterns and isolates anomalies without the need for predefined labels. This approach ensures a scalable and flexible detection framework for real-world applications.

Notwithstanding these advantages, exclusive reliance on unsupervised learning entails specific limitations. These primarily the inherent challenge of validating detection accuracy without ground truth labels and the complexity of distinguishing between rare environmental variability and genuine technical malfunctions.

2.2. Data preparation

The dataset comprises PV production data collected from plants across the Lisbon region of Portugal, made available by the non-profit organization Coopérnico and published by Mendeley Data [8]. The data spans from 2019 to 2023. To ensure technological representativeness, the selected PV plants vary in installed capacity (kW_p) and connection power (kW_p), representing typical distributed generation systems in residential and commercial sectors.

Regarding climatic representativeness, the plants are geographically distributed in Lisbon. This geographical spread captures a range of microclimates characteristic of the Mediterranean region, ensuring the model is tested against diverse weather patterns, including varying irradiance levels, cloud cover, and seasonal temperature fluctuations. Building on this representative dataset, the proposed framework is designed to be highly generalizable to other PV systems: the utilization of specific energy (kWh/kWp) effectively normalizes the output (applicable to systems of varying capacities), while the unsupervised nature of the methodology allows it to autonomously learn operational patterns in different locations.

The dataset captures key performance indicators that are critical to evaluating the system's efficiency and reliability over time. Specifically, the dataset comprises three primary features: Produced energy (kWh), representing the total electrical energy generated; Specific energy (kWh/kWp), indicating the energy output normalized by installed capacity; and CO_2 Avoided (tons), reflecting the estimated reduction in carbon emissions due to solar energy generation. These variables are recorded at regular intervals, offering a temporal view of the system's performance. Such periodic data is essential for monitoring trends, identifying inefficiencies, and detecting operational irregularities that may require intervention. The dataset's granularity enables detailed temporal analysis, particularly for time-series modeling and anomaly detection.

Extended durations of zero energy production are observed, which could result from system shutdowns, sensor failures, or adverse environmental conditions. Extreme values—both unusually high and low—are identified in the produced energy and specific energy features. These outliers, if left unaddressed, distort model outcomes or obscure meaningful insights.

During the initial preprocessing phase, the CO_2 Avoided (tons) feature is excluded from the modeling process despite its strong correlation with energy variables. This exclusion is justified because the column contained a substantial number of malformed and non-numeric entries, and performing extensive data cleaning or imputation risked introducing synthetic bias into the unsupervised models. Crucially, the use of specific energy (kW_h/kW_p) serves as a fundamental normalization factor that decouples energy output from the plant's physical capacity. By focusing on normalized performance rather than absolute magnitude, this approach ensures the generalizability of the anomaly detection framework, enabling its application to other PV systems regardless of their scale or installed power [9].

Fig. 1 shows the methodology for this research. An initial exploratory data analysis (EDA) is conducted to understand the PV dataset. The dataset contains 70693 entries comprising numerical and temporal variables. The dataset's structure, column names, and data types are reviewed to ensure compatibility with subsequent analysis procedures. Descriptive statistics are then generated for the numerical features to summarize their central tendencies and dispersion. This summary helps identify

early indications of skewness [10] or data quality issues [11] that could impact the reliability of machine learning models. Missing values are also investigated at this stage. Both the absolute number and percentage of missing values per feature are calculated to assess the extent of incomplete data. A heatmap is used to visually highlight missing entries across the dataset, enabling easier detection of patterns in missingness. These insights are critical for designing appropriate imputation strategies during preprocessing.

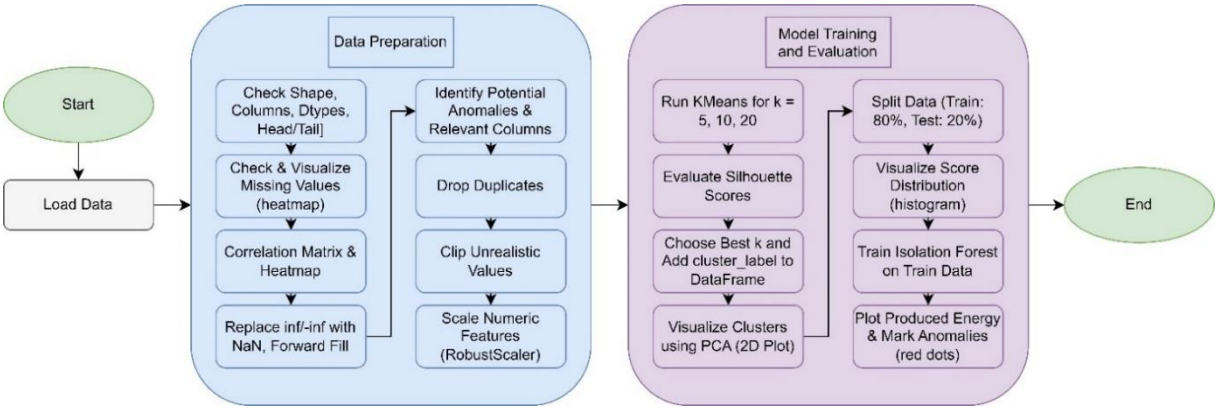

Fig. 1 Flowchart diagram

Subsequently, graphical analyses are conducted to better understand the distributions and anomalies in the numerical attributes [12]. A correlation matrix is computed to quantify the relationships between numerical variables [13]. This analysis provides a preliminary understanding of how different energy metrics relate to one another, which is helpful for feature selection and model development. While "produced energy" and "specific energy" are strongly correlated as expected, the role of "CO_2 avoided" is found to be less central due to its inconsistencies and is therefore excluded in subsequent modeling steps. The data preparation process begins with the identification and removal of duplicate records, a fundamental step to ensure the dataset's integrity and reliability. In total, four duplicate rows are detected within the operational data. These duplicates may arise from data quality degradation [14] or inconsistencies in periodic reporting [15].

In the context of PV energy production anomaly detection, not all available features contribute meaningfully to the modeling objective. This decision is grounded in both data quality concerns and the feature's relevance to detecting system anomalies. Data preparation plays a crucial role in ensuring the quality and effectiveness of subsequent modeling tasks, especially in time series anomaly detection. This stage focuses on transforming raw PV plant data into a structured, model-ready format by time-indexing, feature engineering, handling missing values, and scaling features. To enable temporal analysis, the 'Date' column was converted to a datetime format and used as the dataset's index. This restructuring enables chronological ordering and allows accurate computation of rolling statistics and temporal transformations.

2.3. Model training and evaluation

The initial selection of cluster centers and the distance metric used will influence the results [16]. Descriptive statistics (mean, standard deviation, min, max) are computed for each cluster to characterize PV system behavior within each group. This analysis provides insight into the operational variability and potential outliers or failure modes in the dataset. To visualize the clustering results, Principal Component Analysis (PCA) is used to reduce the feature space to two principal components. A scatter plot was then generated to display a representative subset of 1,000 randomly selected data points, colored by their cluster assignments. This visualization facilitates intuitive interpretation of the cluster distribution and the degree of separation between groups in the reduced feature space.

An anomaly-detection model is developed and evaluated to identify unusual patterns in PV system energy production data. The isolation forest algorithm is selected due to its effectiveness in detecting outliers in high-dimensional datasets without requiring labeled anomaly data [17]. The input feature space for both the K-means clustering and Isolation Forest models

comprises ten variables: produced energy (kWh), specific energy (kWh/kWp), and their respective rolling means, rolling standard deviations, first-order differences, and percentage changes. These features are selected to provide a multi-dimensional representation of system behavior, capturing both absolute performance levels and temporal volatility. Specifically, specific energy serves as a crucial normalization factor to ensure the model's generalizability across different plant capacities, effectively decoupling output from physical scale.

Rolling statistics—calculated over a 7-hour window to align with peak daylight dynamics—and difference metrics are integrated to capture the short-term variability of solar generation. This allows the models to distinguish gradual environmental shifts from abrupt technical anomalies, such as inverter faults or localized shading. The dataset is split into training and test sets at 80/20 ratio. The anomaly threshold is set at -0.05 based on a sensitivity analysis of the decision function scores. As illustrated in Fig. 2, this value corresponds to the critical inflection point (or “elbow”) of the curve. Below this threshold, the detection rate remains relatively flat, indicating potential insensitivity. Conversely, beyond -0.05, the number of detected anomalies exhibits an exponential increase, suggesting the inclusion of significant noise or false positives. Therefore, -0.05 is selected to maximize the detection of true outliers while maintaining model stability.

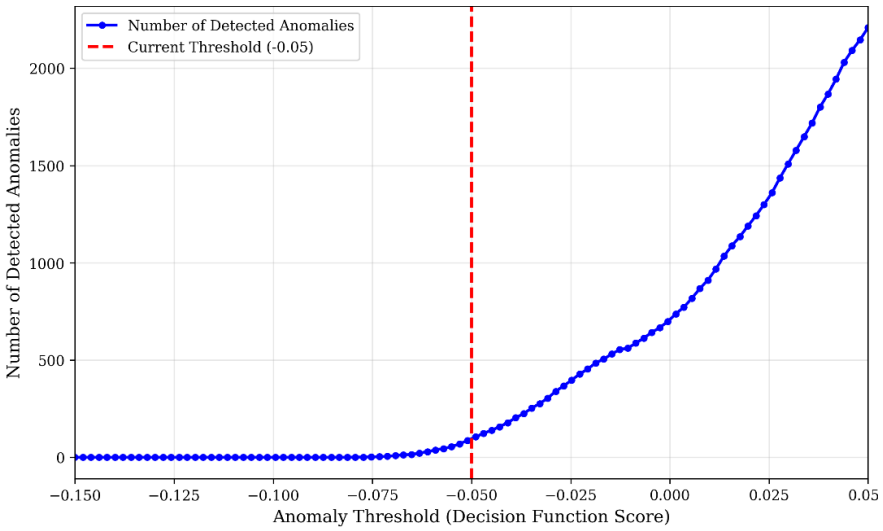

Fig. 2 Sensitivity analysis (threshold)

3. Result and Discussion

The Results and Discussion section employs a three-stage analysis to evaluate PV system performance. Initial correlation analysis established a perfect linear relationship between produced energy and specific energy, while K-means clustering (k=3) categorized system operations into three distinct regimes (low, moderate, high), reflecting varying irradiance conditions. Subsequently, the isolation forest algorithm identifies 17 anomalies. Although many of these occurred during the 2020 COVID-19 period, meteorological validation confirmed they are primarily associated with adverse weather conditions—specifically reduced shortwave radiation—rather than technical malfunctions.

3.1. Correlation Matrix Heatmap

The results serve as a foundation for determining appropriate preprocessing and modeling strategies in the later stages of analysis. Prior to generating the correlation matrix, feature engineering is performed to create rolling statistics and difference metrics, expanding the feature set for deeper analysis [18]. The dataset spans from January 1, 2019, to December 31, 2022, providing multi-year temporal coverage suitable for time-series analysis. The consistent hourly frequency and chronological ordering support rolling statistical analysis and seasonal pattern detection. The features produced energy (kWh) and specific energy (kWh/kWp) are of type float64, making them well-suited for quantitative analysis and machine learning modeling.

However, the column CO2 avoided (tons) is of type object, indicating the presence of non-numeric entries or blank strings. Although no formal missing values are detected, the column format suggests that additional cleaning or exclusion is required unless properly transformed [19].

Descriptive statistics reveal substantial variability in energy production. produced energy (kWh) ranges from 0.0 to 40.0, with a mean of 7.53 and a standard deviation of 11.45; Specific energy (kWh/kWp) ranges from 0.0 to 0.87, with a mean of 0.16 and a standard deviation of 0.25. A significant number of values are equal to zero, particularly in the lower quartiles. This pattern is expected in solar energy systems, where zero production corresponds to nighttime or adverse weather conditions. To avoid misclassifying these extended zero sequences as anomalies, they are retained as valid data. The isolation forest algorithm identifies frequent and high-density cyclical patterns as normal behavior, thereby distinguishing regular nighttime inactivity from isolated system failures.

Inspection of missing values confirms that no columns contain formal null values. However, CO2 Avoided (tons) contains blank strings (" "), which, although not detected as NaN, do not provide helpful information. This requires additional processing, such as type conversion or imputation depending on the intended analysis. The data exhibits strong temporal integrity. The first and last records fell within the defined time frame, and no temporal gaps are observed. The consistent timestamp format and granularity confirm time-series modeling techniques, including trend decomposition, anomaly detection, and rolling window computations.

Fig. 3 shows a correlation matrix heatmap that reveals several important observations. A strong positive correlation is observed between produced energy and specific energy, with a correlation coefficient of 1.00, indicating an almost perfect linear relationship. Both variables show a strong positive correlation ($r = 0.94$) with CO_2 avoided (tons), validating the metric as a meaningful environmental indicator. The Day_of_Week feature shows negligible correlations with all other variables (approx. 0.01), suggesting no linear weekly pattern in production. Additionally, the cluster label demonstrates a moderate positive correlation (0.51) with produced energy, indicating that the clustering structure is associated with variations in energy output levels.

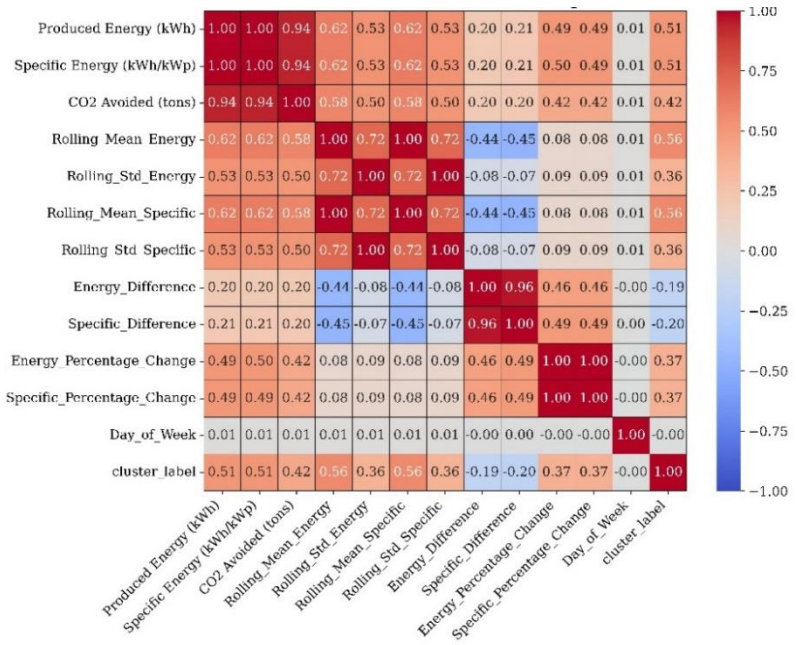

Fig. 3 Correlation matrix heatmap

3.2. Clustering using K-means

During the initial preprocessing phase, the CO_2 avoided (tons) column contains numerous missing values (NaN). Upon review, this feature is deemed non-essential for the objectives of the current analysis, particularly clustering and pattern

recognition. Therefore, no imputation or row removal is performed based on these missing entries. The dataset inspection reveals four duplicate entries, each corresponding to the end of October in 2019–2022. These rows exhibited zero values across all numerical features and missing values in the CO_2 avoided (tons) column. As they contribute no meaningful variance to the data and introduce redundancy, these rows are removed using the drop duplicates function. This step improves data integrity and reduces noise in subsequent analysis.

The next stage focuses on identifying and correcting unrealistic or physically implausible values in the produced energy (kWh) and specific energy (kWh/kWp) columns. Negative values are considered invalid for both metrics, as energy production cannot be negative. Additionally, upper bounds were imposed to eliminate extreme outliers—50 kWh for produced energy and 1.0 kWh/kWp for specific energy. Values outside these thresholds are capped using the clip method to ensure consistency with expected operational ranges. Following the above cleaning procedures, a comprehensive check is conducted to ensure the dataset no longer contains missing (NaN) or infinite values across any of the key numerical features. The results confirm that the data is fully sanitized and meets the criteria for subsequent analytical stages, including scaling, feature generation, and clustering. This quality control step is critical for ensuring robustness and interpretability in the modeling process.

The energy data clustering analysis determines the optimal number of clusters using the silhouette score as shown in Fig. 4. The best silhouette score is achieved at k = 3, with a value of 0.6652, indicating reasonable cluster cohesion and separation [20]. Smaller groups or data points located farther from the cluster centroids are considered potential anomaly candidates.

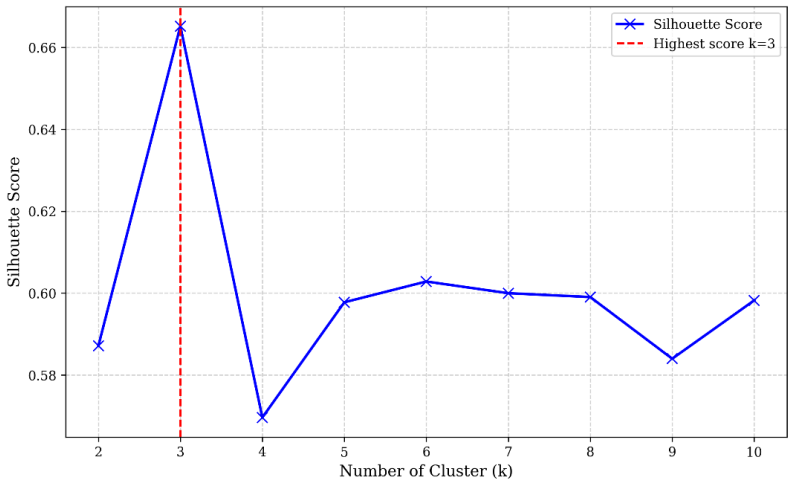

Fig. 4 Silhouette score

K-means clustering was applied to the feature-engineered dataset to identify patterns of operational behavior in the PV energy system. PCA is employed strictly for post-clustering visualization, rather than as a dimensionality reduction preprocessing step, to preserve the full variance and physical interpretability of the original features during the clustering process. Cluster distributions and energy patterns are visualized to provide deeper insights into the data structure and support the identification of underlying behavioral trends across different energy consumption profiles.

Clustering algorithms have become increasingly pivotal in engineering applications for autonomously defining operational boundaries and safety constraints. Their applications range from obstacle avoidance in mobile robotics [21] to performance monitoring in energy systems. To classify the operational states of the PV system without prior labeling, K-means clustering was applied, a method widely recognized for its efficacy in partitioning unlabelled PV monitoring data [22].

The optimal number of clusters is guided by the Elbow Method, which analyzes the relationship between the number of clusters (k) and the within-cluster sum of squares (inertia). As evidenced in Fig. 4, the inertia plot exhibits a distinct inflection point at k=3. Beyond this threshold, the rate of reduction in inertia diminishes significantly. This heuristic approach aligns with standard practices for balancing model complexity and interpretability.

To corroborate the cluster separability established in the full feature space, PCA is utilized to project the high-dimensional data into two principal components. These components explain 89.44% and 7.81% of the total variance (97.25%), respectively. As illustrated in Fig. 5, the scatter plot demonstrates a clear stratification of the three discrete regimes along the first principal component (PC1).

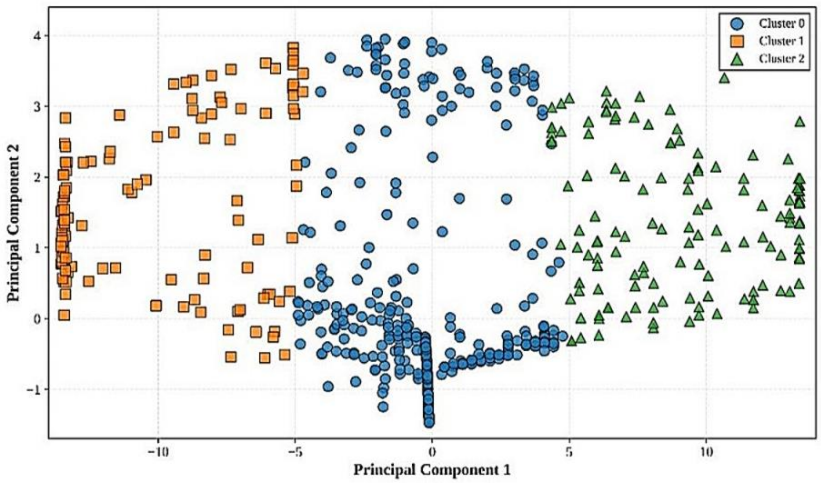

Fig. 5 K-means Clustering Visualization

The scatter plot demonstrates a clear stratification of the data into three discrete regimes along the first principal component (PC1). Cluster 0 (Low Production) aggregates data points associated with the low-production regime. Although meteorological variables are not explicitly used as model inputs, these instances correspond to intervals of minimal solar irradiance, such as heavily overcast conditions or early morning/late afternoon periods, given that energy output acts as a direct proxy for underlying environmental conditions. Previous studies have characterized such low-output clusters as indicators of environmental shading or low irradiance impact [23].

Cluster 1 (Moderate/Transitional) represents a transitional operational state, capturing the variability inherent in fluctuating weather patterns. Energy output in this cluster is consistent but sub-optimal due to intermittent cloud cover [24]. Cluster 2 (High Production) characterizes the high-production regime, signifying optimal system performance during periods of peak solar irradiance and clear-sky conditions.

3.3. Anomaly detection using isolation forest

Fig. 6 shows the distribution of anomaly scores generated by a detection model applied to Solar PV system data. The distribution reveals a significant concentration of instances with low anomaly scores, indicative of typical operational conditions. However, a subset of data points displays elevated scores, approximately between 0 to 0.2, indicating potential anomalous behavior.

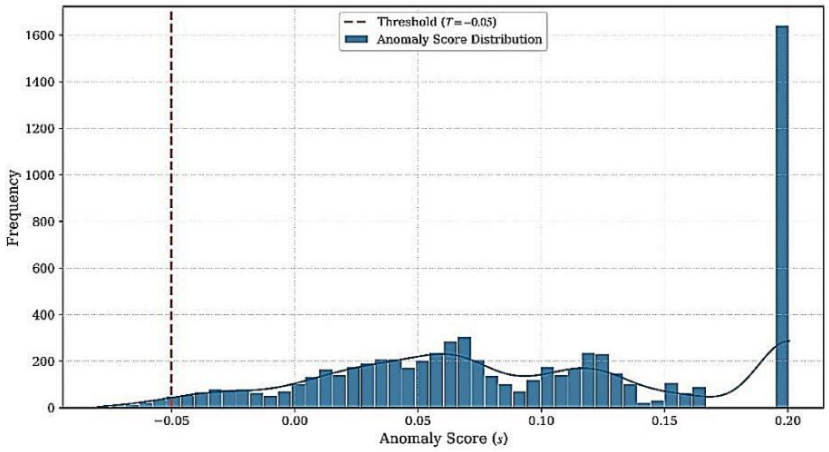

Fig. 6 Distribution of Anomaly Scores

The multimodal nature of the anomaly score distribution, with several smaller peaks beyond the dominant mode near 0.21, suggests the presence of multiple types of anomalies. These could stem from degraded panel performance, inverter malfunctions, tracking system errors, or connection problems.

The distribution of anomaly scores underscores the necessity for further in-depth investigation. Specifically, it is important to examine the data features associated with higher anomaly scores in order to effectively diagnose and address the underlying causes of deviations from regular Solar PV system operation.

Fig. 7 shows the time series chart titled "Detected Anomalies in Produced Energy", presenting daily energy production (kWh) from January 2019 to January 2023. The blue area represents total energy produced each day, while red "X" markers highlight detected anomalies—data points that deviate significantly from typical production levels. During these four years, the produced energy generally fluctuates between approximately 2.0-3.2 kWh, indicating relatively stable performance. Nevertheless, 17 anomalies were detected, occurring unevenly across time, with clusters in mid-2020 and scattered instances in 2019, 2021, and 2022.

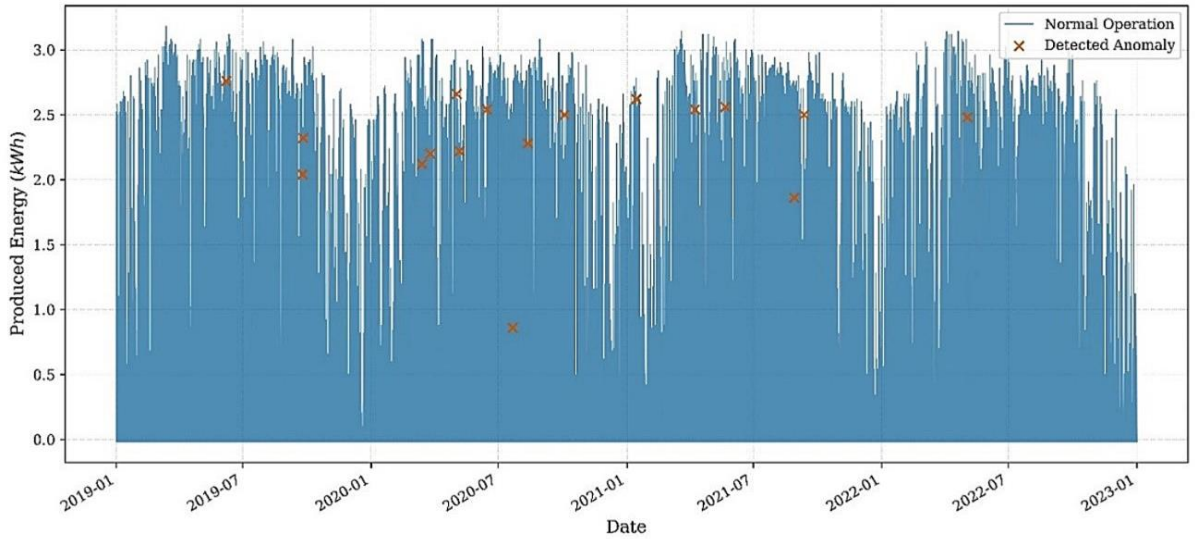

Fig. 7 Detected Anomalies data

The most critical deviation occurred in mid-2020, when the produced dropped below 1.0 kWh—substantially lower than the average daily value. This may be indicative of major disruptions, such as technical failures, sensor errors [25], energy system downtime [26], or environmental conditions [23]. In the absence of ground truth validation, true system faults are distinguished from normal variability by cross-referencing detected anomalies with concurrent meteorological data; specifically, deviations under optimal weather are considered potential technical failures, whereas those under poor weather are attributed to environmental effects.

Some anomalies fall within the broader 2.0–2.8 kWh range but are still flagged as abnormal, showing that the model detects contextual and temporal deviations rather than relying solely on fixed thresholds. The highest anomaly frequency occurs in 2020 (8 out of 17 events). Although this coincided with global disruptions due to the COVID-19 pandemic, which may have affected maintenance schedules [27], this study's meteorological analysis (see Table 1) indicates that most of these deviations were primarily driven by adverse weather conditions rather than systemic failures. Preventive measures, such as the availability of batteries [24] or combining PV generation with traditional energy sources [28], can mitigate intermittency or periods of low production.

Table 1 presents the meteorological conditions coinciding with the detected anomalies. A close examination reveals that the majority of significant energy deviations correspond to periods of reduced solar availability or adverse weather patterns. For instance, anomalies observed on dates such as August 12, 2020, and July 22, 2020, were characterized by high cloud cover

(reaching 96%) and significantly reduced shortwave radiation (recorded as low as 178–275 W/m²). These conditions naturally limit photovoltaic generation, suggesting that the system performed consistently with the available resource rather than suffering from technical malfunction.

Table 1 Weather on all Anomalies Data

Time (PM)	Temp. (Â °C)	Humidity (%)	Dew Pt. (Â °C)	App. Temp (Â °C)	Cloud (%)	Wind Spd (Km/H)	Wind Dir (Â °)	SW Rad. (W/mÂ ²)
01/14/21 03:00	12.5	68	6.7	9.3	19	16	352	387
08/12/20 01:00	21.7	79	17.9	22.5	96	14.2	294	275
09/24/19 04:00	23.6	72	18.2	25.2	14	9.7	297	401
09/25/19 05:00	20.9	63	13.5	19.4	0	17.6	337	309
03/25/20 04:00	18	67	11.8	16	30	17.4	326	434
04/08/21 02:00	18.8	70	13.2	18.9	36	12	229	699
09/11/21 04:00	26.4	56	16.8	27.1	0	11.4	235	542
10/03/20 03:00	19.6	50	8.9	17.1	85	15	294	456
03/13/20 02:00	17.4	63	10.3	15.5	42	18.4	354	679
05/03/22 05:00	19.3	57	10.6	15.8	20	25.3	345	412
05/02/20 01:00	21.1	67	14.8	24.7	46	2.8	220	916
06/14/20 03:00	21.3	54	11.8	20.8	2	14.8	274	760
05/06/20 04:00	20.7	63	13.5	20.3	31	12.5	316	607
08/28/21 05:00	24.1	65	17.2	24.7	40	12.9	306	342
06/07/19 01:00	20.2	36	4.8	18.6	47	14.3	342	881
07/21/20 07:00	29	52	18.1	31.4	1	3.7	349	178
05/21/21 04:00	20.1	63	12.7	18.2	29	20.9	336	616

To quantitatively validate these observations, a pearson correlation analysis is conducted specifically on the anomalous data points listed in Table 1. The analysis shows a strong positive correlation ($r = 0.69$, $p < 0.05$) between produced energy and solar shortwave radiation. The robustness of the detected anomalies is validated through physical consistency analysis rather than algorithmic parameter tuning. These findings demonstrate that detection aligns strongly with meteorological conditions. In particular, the 17 anomalies primarily reflect environmentally driven operational variability rather than stochastic noise or model artifacts.

In the absence of labeled ground truth, this study utilizes internal validation metrics to contrast the efficacy of the two methods. The K-means algorithm demonstrates robust structural definition of operational regimes, as evidenced by a silhouette score of 0.6652, which indicates a high degree of cluster cohesion and separation.

In comparison, the isolation forest model's performance is objectively assessed through the distribution of anomaly scores and a sensitivity analysis of the decision function. An inflection point at -0.05 optimizes the trade-off between detection sensitivity and stability. This comparison indicates that while K-means is superior for minimizing within-cluster variance to define global operational baselines, isolation forest offers greater discriminatory power for isolating sparse, stochastic deviations that do not conform to the dense cluster structures.

Beyond diagnostic accuracy, computational efficiency is a critical factor for real-time in industrial settings. Both K-means and isolation forest are computationally lightweight algorithms. K-means operates with a time complexity of $O(n \cdot k \cdot i)$, where n is the sample size, k is the number of clusters, and i is the number of iterations. Isolation forest exhibits a logarithmic time complexity of $O(n \log n)$ for training and effectively constant time for inference. This low computational overhead implies that the framework does not require high-performance computing infrastructure and is suitable for deployment on resource-constrained edge devices or embedded controllers.

4. Conclusion

Based on four years of operational data from the solar PV system, this study employed a multivariate approach, using K-means clustering (visualized via PCA) and anomaly detection, to characterize both typical and irregular system behaviors. The main conclusions are summarized as follows:

(1) The K-means clustering analysis, visualized via PCA, categorized the system's operational performance into three distinct clusters. This framework confirms the existence of unique production-related patterns shaped by a combination of environmental and performance variables.

(2) The results reinforce that while solar irradiance is the primary driver, PV performance is significantly affected by secondary meteorological parameters—including extreme temperature, high humidity, dense cloud cover, and potentially unmeasured atmospheric pollutants.

(3) A total of 17 operational anomalies were identified across the four years. The highest frequency of anomalies (8 occurrences) was observed in 2020, a finding that may be tentatively linked to external systemic disruptions, such as those associated with the COVID-19 pandemic or operational delays.

(4) The observed reductions in PV output on the selected dates were consistently attributable to non-ideal and variable weather conditions, confirming that these meteorological factors directly impact the solar energy conversion efficiency.

(5) To address system intermittency and abnormal output reductions, preventive approaches are recommended. These include the incorporation of battery energy storage solutions (BESS) and the strategic integration of PV systems with conventional power sources to ensure supply continuity and enhance overall system reliability.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] A. A. Bayod-Rújula, "Solar Photovoltaics (PV)," in Solar Hydrogen Production: Processes, Systems and Technologies, Elsevier, pp. 237-295, 2019.

[2] E. Jiménez-Delgado, C. Meza, A. Méndez-Porras, and J. Alfaro-Velasco, "Data Management Infrastructure from Initiatives on Photovoltaic Solar Energy," Proceedings of International Conference on Information Technology & Systems, pp. 113-121, 2019.

[3] M. A. Koondhar, I. A. Laghari, B. M. Asfaw, R. Reji Kumar, and A. H. Lenin, "Experimental and Simulation-Based Comparative Analysis of Different Parameters of PV Module," Scientific African, vol. 16, article no. e01197, 2022.

[4] R. S. Hansen, A. A. Munaf, H. L. Allasi, S. Endro, J. Leno, and S.K. R. Kanna, "Experimental and Theoretical Optimization of an Inclined Type Solar Still Using PV Sustainable Recirculation Technique," Materials Today: Proceedings, vol. 45, Part 7, pp. 7063-7071, 2021.

[5] E. Koubli, D. Palmer, P. Rowley, and R. Gottschalg, "Inference of Missing Data in Photovoltaic Monitoring Datasets," Institution of Engineering and Technology Renewable Power Generation, vol. 10, no. 4, pp. 434-439, 2016.

[6] S. Touzani, A. K. Prakash, Z. Wang, S. Agarwal, M. Pritoni, M. Kiran, et al., "Controlling Distributed Energy Resources via Deep Reinforcement Learning for Load Flexibility and Energy Efficiency," Applied Energy, vol. 304, article no. 117733, 2021.

[7] T. Park, K. Song, J. Jeong, and H. Kim, "Convolutional Autoencoder-Based Anomaly Detection for Photovoltaic Power Forecasting of Virtual Power Plants," Energies, vol. 16, no. 14, article no. 5293, 2023.

[8] E. Sarmas, N. Matias, C. Pereira, and A. R. Antunes, "Photovoltaic Power Production Dataset," Mendeley Data, vol. 3, 2025.

[9] J. L. de S. Silva, M. V. de Paula, J. De S. G. Barros, and T. A. Dos S. Barros, "Anomaly Detection Workflow Using Random Forest Regressor in Large-Scale Photovoltaic Power Plants," IEEE Access, vol. 13, pp. 54168-54176, 2025.

[10] I. H. Adil, A. Wahid, and E. H. Mantell, "Split Sample Skewness," Communications in Statistics - Theory and Methods,

vol. 50, no. 22, pp. 5171-5188, 2021.
[11] C. M. Chang, D. Cheng, R. E. Smith, S. G. Tan, and A. Hossain, "SMART Quality Control Analysis of Pavement Condition Data for Pavement Management Applications," International Journal of Transportation Science and Technology, vol. 18, pp. 227-244, 2024.
[12] W. Cui and H. Wang, "A New Anomaly Detection System for School Electricity Consumption Data," Information, vol. 8, no. 4, article no. 151, 2017.
[13] P. Schober, C. Boer, and L. A. Schwarte, "Correlation Coefficients: Appropriate Use and Interpretation," Anesthesia & Analgesia, vol. 126, no. 5, pp. 1763-1768, 2018.
[14] O. Azeroual, A. Nikiforova, and K. Sha, "Overlooked Aspects of Data Governance: Workflow Framework for Enterprise Data Deduplication," Proceedings of 2023 International Conference on Intelligent Computing, Communication, Networking and Services (ICCNS), IEEE, pp. 65-73, 2023.
[15] V. B. Sorkhabi, M.-R. F. Derakhshi, and H. Shahamfar, "An Algorithm for Detecting Similar Data in Replicated Databases Using Multi Criteria Decision Making," Proceedings of 2009 Second International Conference on Environmental and Computer Science, IEEE, pp. 199-203, 2009.
[16] H. Qi, X. Di, J. Li, and H. Ma, "Improved K-Means Algorithm and Its Application to Vehicle Steering Identification," Proceedings of International Conference on Advanced Hybrid Information Processing, pp. 378-386, 2018.
[17] S. Buschjager, P.-J. Honysz, and K. Morik, "Generalized Isolation Forest: Some Theory and More Applications Extended Abstract," Proceedings of 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), IEEE, pp. 793-794, 2020.
[18] F. P. Monteiro, S. Monteiro, C. Rodrigues, J. Reis, U. Bezerra, M. E. Tostes, Frederico, et al., "A Hybrid Methodology Using Machine Learning Techniques and Feature Engineering Applied to Time Series for Medium- and Long-Term Energy Market Price Forecasting," Energies, vol. 18, no. 6, article no. 1387, 2025.
[19] M. Usmani, Z. A. Memon, A. Zulfiqar, and R. Qureshi, "Preptimize: Automation of Time Series Data Preprocessing and Forecasting," Algorithms, vol. 17, no. 8, article no. 332, 2024.
[20] A. Sleiman and W. Su, "Combined K-Means Clustering with Neural Networks Methods for PV Short-Term Generation Load Forecasting in Electric Utilities," Energies, vol. 17, no. 6, article no. 1433, 2024.
[21] J. -F. Wu, C. -C. Huang, and M. -Y. Cheng, "DBSCAN-Based Minimum Enclosing Ellipse Using the Control Barrier Function for Safe Navigation of Mobile Robots," Advances in Technology Innovation, vol. 10, no. 4, pp. 358-369, 2025.
[22] G. Liu, L. Zhu, X. Wu, and J. Wang, "Time Series Clustering and Physical Implication for Photovoltaic Array Systems with Unknown Working Conditions," Solar Energy, vol. 180, pp. 401-411, 2019.
[23] M. A. M. Ramli, E. Prasetyono, R. W. Wicaksana, N. A. Windarko, K. Sedraoui, Y. A. Al-Turki, "On the Investigation of Photovoltaic Output Power Reduction Due to Dust Accumulation and Weather Conditions," Renewable Energy, vol. 99, pp. 836-844, 2016.
[24] H. A. H. Al-Hilfi, A. Abu-Siada, and F. Shahnia, "Estimating Generated Power of Photovoltaic Systems During Cloudy Days Using Gene Expression Programming," IEEE Journal of Photovoltaics, vol. 11, no. 1, pp. 185-194, 2021.
[25] S. Saha, M.E. Haque, C.P. Tan, M.A. Mahmud, M.T. Arif, S. Lyden, et al., "Diagnosis and Mitigation of Voltage and Current Sensors Malfunctioning in a Grid Connected PV System," International Journal of Electrical Power & Energy Systems, vol. 115, article no. 105381, 2020.
[26] B. M. Ali, T. J. Al-Musawi, A. Mohammed, H. F. Fakhruldeen, T. M. Hanoon, A. Khurramov, et al., "Sustainable Strategies for Preventive Maintenance and Replacement in Photovoltaic Power Systems: Enhancing Reliability, Efficiency, and System Economy," Unconventional Resources, vol. 6, article no. 100170, 2025.
[27] D. Deshwal, P. Sangwan, and N. Dahiya, "How Will COVID-19 Impact Renewable Energy in India? Exploring Challenges, Lessons and Emerging Opportunities," Energy Research & Social Science, vol. 77, article no. 102097, 2021.
[28] J. Wang, X. Teng, and X. Zhang, "Coordinated Control Strategy Based on Photovoltaic Generation Integration," Proceedings of 11th IET International Conference on Developments in Power Systems Protection (DPSP 2012), IET, pp. 1-4, 2012.