Solar Energy 307 (2026) 114351

Contents lists available at ScienceDirect

Solar Energy

journal homepage: www.elsevier.com/locate/solener

Anomaly detection, filtering and visualization for Outdoor photovoltaic data

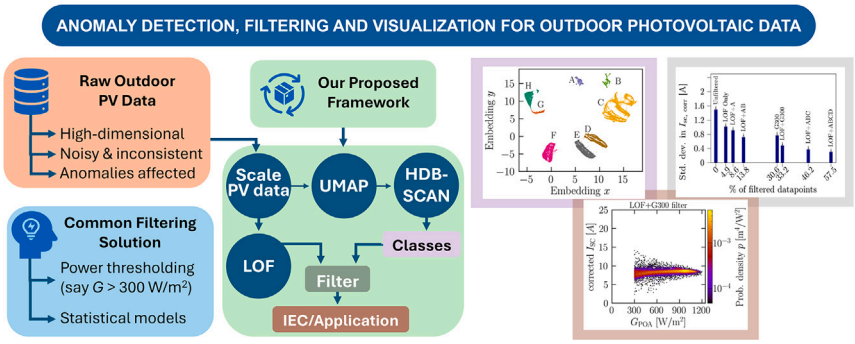
Goodfriend M. Whyte[a,b], Andreas Gerber[a], Uwe Rau[a,b], Bart E. Pieters[a,*]

[a] IMD-3 Photovoltaik, Forschungszentrum Jülich GmbH, 52425 Jülich, Germany
[b] Jülich Aachen Research Alliance (JARA-Energy) and Faculty of Electrical Engineering and Information Technology, RWTH Aachen University, Aachen, Germany

HIGHLIGHTS

- UMAP–HDBSCAN enables unsupervised classification of outdoor PV data.
- Identified data classes have clear physical interpretations.
- LOF filtering supports flexible, application-specific data filtering.
- Method corrects I_{sc} to STC and uses its standard deviation as noise metric.
- LOF + cluster removal outperforms thresholding, reducing noise with less data loss.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Monitoring data
UMAP
LOF
Filtering
Anomalies

ABSTRACT

Long-term outdoor testing of photovoltaic (PV) modules is critical for assessing reliability but is often challenged by inconsistent and noisy data, necessitating robust filtering techniques. Traditional threshold-based filtering methods (e.g., irradiance > 300 W/m^2) are inadequate for capturing complex data patterns inherent to outdoor data. This study presents an advanced multi-dimensional filtering framework combining Uniform Manifold Approximation and Projection (UMAP) and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) for unsupervised classification and visualization. Our approach effectively identifies meaningful data classes, such as conditions of partial shading and fluctuating irradiance levels, enabling the development of targeted filters. Additionally, we employ the Local Outlier Factor (LOF) to detect and remove anomalies. To demonstrate the effectiveness of class-specific filtering—such as the exclusion of partially shaded data—combined with LOF-based anomaly detection, we correct outdoor measurements to Standard Test Conditions (STC) using IEC 60891:2021. Our results show that this methodology achieves higher accuracy than conventional threshold-based filtering while preserving a larger proportion of the data.

1. Introduction

The rapid expansion of large-scale outdoor photovoltaic (PV) installations has underscored the importance of monitoring and analyzing extensive, long-term datasets of PV module performance [1]. Accurate data analysis is crucial for optimizing system efficiency, diagnosing performance issues, and ensuring reliable energy production. However, data collected over long periods is often compromised by anomalies and outliers arising from various sources. These irregularities can stem from sensor malfunctions, shading, sensor drift, calibration errors, and

* Corresponding author.
Email address: b.pieters@fz-juelich.de (B.E. Pieters).

https://doi.org/10.1016/j.solener.2026.114351
Received 8 September 2025; Received in revised form 12 January 2026; Accepted 16 January 2026
Available online 27 January 2026

data retrieval issues such as missing values, data shifts, repetitions, or synchronization problems [2].

Hawkins [3] defines an anomaly as "an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism." His definition thus emphasizes that anomalies are not merely random deviations but are caused by a fundamentally different process or mechanism. Detecting such anomalies is critical, as they can obscure accurate performance assessment and hinder fault detection. Conventional filtering methods, such as power thresholding, are often employed to address this challenge by discarding data points exceeding predefined limits [4–8]. While simple and effective for eliminating extreme outliers, these threshold-based approaches are inherently limited. They fail to capture the complex interdependencies among variables and offer little insight into the root causes of anomalies, thus risking false positives (valid data being discarded) or false negatives (anomalies being overlooked).

To overcome these limitations, more robust techniques such as the Mahalanobis distance-based filtering approach have been introduced [9–11]. This method evaluates the multivariate distance of data points from the mean, considering correlations between variables. By accounting for the covariance structure within the dataset, Mahalanobis filtering is more suited for detecting subtle and context-dependent anomalies, making it a valuable tool for ensuring data integrity in PV system monitoring.

Recent work has suggested that leveraging the correlations between multiple dimensions in outdoor PV datasets can enhance anomaly detection accuracy [12]. This approach relies on constructing models that represent well-established relationships between different variables in the dataset. Data points that deviate from these expected correlations are identified as potential anomalies. The rationale behind this method aligns with the notion that normal data arises from an underlying mechanism, and any departure from this mechanism signals an outlier.

However, this correlation-based approach has a notable drawback: it requires prior knowledge of all the relevant mechanisms governing data generation, along with accurate model descriptions for each. In practice, this is often challenging. For instance, in the work by Vaas et al. [12], the method effectively filtered out partial shading conditions. While this may be desirable in some scenarios, partial shading is not necessarily an anomaly. It represents a common operational condition in outdoor PV systems. Moreover, accurately modeling partial shading effects requires precise information about the shading conditions themselves – data that is typically not monitored. This limitation parallels the previously discussed power thresholding approach, which tends to filter out low irradiance data points – conditions often associated with partial shading. Consequently, if the objective is to analyze the effects of partial shading, both power thresholding and correlation-based filtering methods may prove unsuitable.

Essentially, the drawbacks of these filtering procedures include the unintended removal of usable data, the insufficient filtering of frequent measurement errors, and the often-naive imputation of untrusted data points [13–15]. Additionally, these anomaly detection methods are generally applied on a global scale. For instance, power threshold filtering typically results in the exclusion of large portions of data, raising concerns about whether the remaining dataset accurately represents the system's performance under real operating conditions [2,7].

In this work, we therefore propose the use of the Local Outlier Factor (LOF) approach [16], a density-based anomaly detection method that evaluates data points based on their local neighborhood. LOF compares the density of a point to that of its nearest neighbors and assigns an "outlierness" measure, thereby identifying anomalies as data points that are isolated or lie in sparse regions relative to their surroundings.

Furthermore, we suggest the application of an unsupervised classification technique. Unlike supervised methods, unsupervised classification requires no prior knowledge of the data and can reveal distinct modes of operation within the dataset. Analyzing these identified classes provides insight into the typical behavior of the PV system under different conditions. Such fine-grained classification is particularly valuable for data filtering, as it facilitates the selection of relevant data points tailored to specific analysis objectives. Notably, filtering can be viewed as a classification task in itself – separating useful from non-useful data – underscoring the potential of unsupervised classification as a flexible and robust tool for improving data quality in PV monitoring systems.

For the unsupervised classification of the monitoring data, we employ a combination of Uniform Manifold Approximation and Projection (UMAP) and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). This approach achieves two complementary objectives: it performs unsupervised classification to identify clusters representing different modes of system operation, while simultaneously providing a convenient low-dimensional representation of the data. This low-dimensional projection is particularly valuable for visualizing the dataset and the identified classes, offering an intuitive way to explore system behavior and assess the effectiveness of the classification.

In this work, we demonstrate our proposed filtering and classification approach using a dataset from a monocrystalline silicon PV module located in Arizona, USA. The dataset spans over two years and includes measurements of plane-of-array irradiance, module temperature, and complete current-voltage (I–V) sweeps. As a first step, the extended solar cell parameters [17] are extracted from the I-V data, reducing each sweep to a set of ten well-defined electrical parameters. Combined with the irradiance and temperature measurements, this results in a 12-dimensional dataset.

We apply UMAP and HDBSCAN to perform an unsupervised classification of the data, aiming to determine whether the identified classes correspond to meaningful operational states, such as partial shading, and low- and high-irradiance conditions. Additionally, we use the LOF approach to detect outliers. The low-dimensional UMAP projection is further employed to visualize the classification results and the identified outliers.

To assess the practical utility of our classification scheme for filtering purposes, we apply the fitting procedure 2 of the International Electrotechnical Commission (IEC 60891:2021) standard to the filtered subsets of the data. We evaluate the standard test conditions (STC) short-circuit current (I_{sc}), using the standard deviation of the resulting I_{sc} values and the percentage of filtered data points as performance metrics. We compare the effectiveness of filters based on our unsupervised classification and LOF approach with the commonly applied power thresholding method. The overall anomaly detection pipeline is depicted in Fig. 1.

This paper is organized as follows: Section 2 outlines the methods employed in this work, including the Local Outlier Factor (LOF) algorithm, UMAP, HDBSCAN, and the IEC 60891:2021 Procedure 2. Section 3 describes the dataset used in this study, along with the preprocessing steps applied to the feature vectors, such as normalization and scaling. Section 4 presents the results and discussion of the key findings and their broader implications. Finally, Section 5 concludes the study by summarizing the key findings.

2. Methods

2.1. Anomaly detection using LOF

The LOF algorithm is a widely adopted method for detecting anomalies in data. Unlike global outlier detection methods that identify outliers with respect to the entire data set, LOF focuses on local regions, allowing it to detect outliers that may be masked in a global context. A more comprehensive description of LOF may be found in [16]. In the following we briefly introduce the LOF. Before we introduce an expression for the LOF

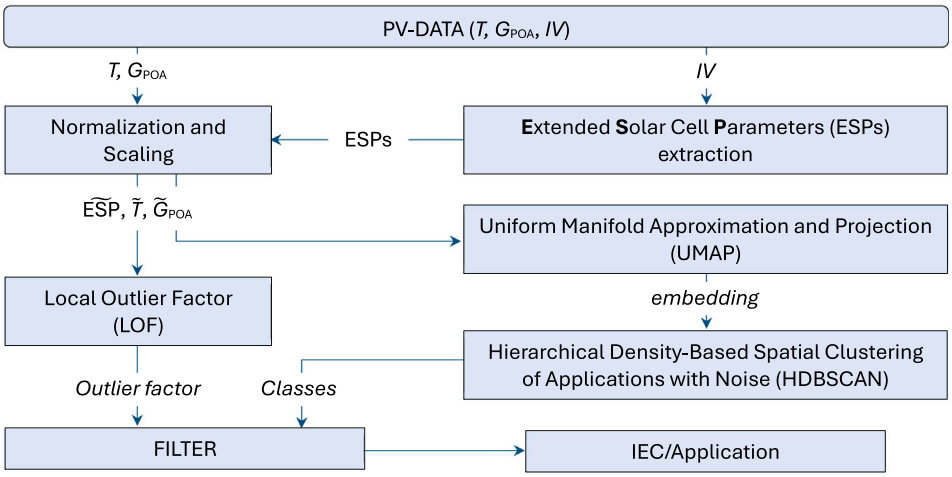


Fig. 1. Illustration of the anomaly detection pipeline, showing data flow from PV module data preprocessing through the anomaly detection module (UMAP → HDBSCAN → LOF) to an exemplary application.

of a point in a dataset, let us first introduce several related terms and quantities.

- ***k*-distance:** The LOF is based on the k-nearest neighbors algorithm (k-NN), which determines for every point, p, in the dataset, the closest k points. To this end, all distances between point p and all other points in the dataset D, are computed and sorted in ascending order. After which the first k points are selected. The k-distance, $d_k(p)$ is defined as the distance between p and its k-th neighbor, i.e., $d_k(p)$ is the maximum distance between p and its k nearest neighbors.
- **Reachability-Distance:** The reachability distance, RD_k is defined for any two points, p and o, as:

$$\mathrm{RD}_k(p,o) = \max\left(d_k(o), d(p,o)\right), \tag{1}$$

  where $d(p,o)$ is the distance between points p and o. If, for example, point p is far away from point o, and is thus not one of o's nearest neighbors, RD_k equals the distance between points p and o. However, if p is one of o's nearest neighbors, RD_k equals the k-distance of o, $d_k(o)$.
- **Local Reachability Density:** The local reachability density (LRD) measures the density of an area by considering the reachability distances of a point to its neighbors. The LRD of a point p is the inverse of the average reachability distance of p to its k-nearest neighbors, and is defined as:

$$\mathrm{LRD}_k(p) = \left(\frac{\sum_{o\in N_k(p)} \mathrm{RD}_k(p,o)}{|N_k(p)|}\right)^{-1}, \tag{2}$$

where $N_k(p)$ is the set of nearest neighbors to point p and $|N_k(p)|$ is the total number of nearest neighbors of p (in general this will be equal to k, however, in some implementations of k-NN there may be situations where the number of neighbors varies under certain conditions, for example when several points have the same distance to point p).

With these terms and definitions, we can now write the LOF for point p:

$$\mathrm{LOF}_k(p) = \frac{\sum_{o\in N_k(p)} \frac{\mathrm{LRD}_k(o)}{\mathrm{LRD}_k(p)}}{|N_k(p)|}. \tag{3}$$

Thus, the LOF for a point p, quantifies how much the local density of p deviates from that of its neighbors. Points with LOF values close to 1 are considered to be in regions of similar density, indicating normal points while significantly larger values indicate outliers, as they lie in sparse regions relative to their neighbors. In this work, we applied the commonly used LOF threshold ($p > 1.5$) as a simple and reliable criterion to identify anomalies in datasets with varying densities [16]. The LOF algorithm's capacity to adapt to the varying densities of different regions in the dataset makes it a powerful tool for anomaly detection across a wide range of applications.

2.2. UMAP dimensionality reduction

To visualize the complex process of anomaly detection in high-dimensional data, our study uses UMAP, an algorithm known for its ability to reduce dimensionality. UMAP achieves this by combining ideas from manifold learning and topological data analysis [18,19]. At its core, UMAP operates on the idea that high-dimensional data often lies on a lower-dimensional surface, or manifold, hidden within the larger space. The goal of UMAP is to map this complex high-dimensional structure into a simpler, lower-dimensional form while preserving as much of the original data's local and global relationships as possible.

The process begins by building a weighted graph, where each data point becomes a node. Connections between nodes represent distances between points, typically calculated using a nearest-neighbor approach. The strength of these connections reflects how similar the points are, often measured using a local metric like Euclidean distance. The number of neighbors considered (k-nearest neighbors) plays a key role in shaping these connections. In brief, UMAP's operation is shaped by its key hyperparameters:

- **n_neighbors:** controls the size of the local neighborhood around each point, affecting how UMAP balances local detail versus global structure.
- **min_dist:** determines how close points can appear in the reduced space, influencing the tightness of clusters and the overall clarity of the visualization.
- **metric:** defines how distances are measured in the original data, offering flexibility to handle different types of data.

In this work, n_neighbors, min_dist, and metric were set to 500, 0.0, and Euclidean, respectively, to maintain a global-local structure balance, facilitate compact clustering, and accommodate dense, continuous data. This careful choice of parameters leverages UMAP's distinguishing ability, compared to prior dimensionality reduction methods like t-SNE, to preserve global structure while remaining computationally efficient and scalable, even for large, high-dimensional datasets [20,21].

In preserving local structure, this means points that are close together in the high-dimensional space tend to stay close in the reduced space, making it easier to explore nuanced patterns within the data. However, this local focus can sometimes come at the cost of accurately representing global relationships. While clusters themselves are well-defined, the overall arrangement of clusters across the dataset may not be as precise.

Another important aspect of UMAP is its tendency to emphasize clustering. Because the algorithm prioritizes local similarities, it can lead to an exaggerated perception of cluster separation and density. While this feature is advantageous for identifying and visualizing discrete groupings within the data, it necessitates a cautious interpretation of clustering results. However, quantitative assessment via silhouette analysis demonstrates that genuine cluster structure is present in the original high-dimensional space (see Appendix A).

2.3. HDBSCAN clustering

HDBSCAN is a density-based clustering algorithm that identifies clusters of varying densities and detects noise points [22–24]. Its operation begins by transforming the data into a density-adaptive space using the mutual-reachability distance ($d_{\mathrm{mr,k}}(p,q)$), defined as:

$$d_{\mathrm{mr,k}}(p,q) = \max\left(d_k(p), d_k(q), d(p,q)\right), \tag{4}$$

where $d(p,q)$ is the Euclidean distance between points p and q, and $d_k(p)$ is the core distance of p, which is the distance from p to its k-th nearest neighbor. This transformation ensures that points in dense regions are closer, while sparse regions are farther apart. HDBSCAN then constructs a weighted graph, where data points are vertices, edges represent connections between points, and weights on the edges are the mutual reachability distances. This weighted graph is used to build a Minimum Spanning Tree (MST), which connects all points with the smallest total edge weight. By iteratively removing edges with the highest weights (weakest links), HDBSCAN constructs a hierarchy of clusters and extracts stable clusters based on their persistence across density thresholds.

An important feature of HDBSCAN is its ability to identify noise points–data that do not belong to any cluster with sufficient density. Rather than forcing all data points into clusters, HDBSCAN labels low-density points as noise, which helps reduce overfitting and improves the interpretability of the results, particularly in noisy or high-dimensional datasets. This makes HDBSCAN especially well-suited for real-world data where outliers or ambiguous points are expected.

An important hyperparameter in HDBSCAN is the *minimum cluster size*, which specifies the smallest size grouping that will be considered a cluster. This parameter has a direct impact on both the number of detected clusters and the proportion of points labeled as noise. Smaller values of *min_cluster_size* allow the algorithm to detect finer-grained structures, potentially leading to a higher number of smaller clusters, but can also increase sensitivity to noise and result in over-clustering. Conversely, larger values favor the formation of more robust, coarse-grained clusters, at the expense of potentially merging distinct groups and labeling more ambiguous points as noise. Selecting an appropriate *min_cluster_size* thus involves a trade-off between cluster resolution and robustness, and should be guided by domain knowledge or validated empirically based on the intended application.

As we already use UMAP for visualization, we also chose to employ UMAP as a dimensionality reduction step prior to applying HDBSCAN. In certain datasets, applying UMAP before clustering has been shown to improve clustering accuracy and reduce noise [25,26]. In our case, the resulting clusters aligned well with the UMAP visualizations, offering an interpretable representation of the data structure. However, it should be noted that using UMAP as a preprocessing step for clustering may introduce discontinuities in the data manifold, potentially leading to under- or over-clustering. Additionally, the observed reduction in noise, when using UMAP as a preprocessor, contradicts, to some extent, HDBSCAN's intentional design of identifying noise points. However, since we also apply the LOF to remove outliers, HDBSCAN's internal noise modeling is not critical for our application.

Preliminary tests, summarized in the supplementary material, demonstrate that clustering quality, as evaluated by metrics such as the silhouette score and cluster stability, is highly sensitive to the number of nearest neighbors, with larger values generally improving performance at the cost of increased computational time. We posit that the minimum cluster size parameter in HDBSCAN should be chosen to be less than or equal to the number of nearest neighbors used in UMAP. This is because UMAP primarily focuses on preserving local distances, and a minimum cluster size significantly larger than the UMAP neighborhood size increases the risk of fragmenting clusters into noise. Specifically, if the minimum cluster size exceeds the number of neighbors considered by UMAP, clusters smaller than this threshold may be incorrectly identified as noise. Initial results corroborate this argument; we occasionally observed an elevated noise fraction in tests when the HDBSCAN minimum cluster size was greater than the number of UMAP nearest neighbors. Furthermore, selecting a smaller minimum cluster size allows HDBSCAN to identify finer-grained clusters potentially lost when a larger value is used.

2.4. IEC 60891:2021 procedure 2

To correct the IV characteristics to Standard Test Conditions (STC), for a given non-standard temperature (T) and irradiance (G), we utilize Procedure 2 from the IEC 60891:2021 standard [27]. This procedure defines a method to translate a single IV curve from one set of temperature and irradiance conditions to another. Specifically, it provides equations to adjust the current and voltage coordinates of each point in the IV curve from condition 1 (T_1, G_1) to condition 2 (T_2, G_2). The translation is performed with respect to STC (T_0 = 25 °C, G_0 = 1000 W/m^2). The current is translated using:

$$I_2 = I_1 \frac{G_2}{G_1} \frac{1+\alpha(T_2 - T_0)}{1+\alpha(T_1 - T_0)}, \tag{5}$$

where $I_{1,2}$ are the currents under conditions 1 and 2, respectively, and α is the temperature coefficient of the short-circuit current.

The voltage is translated according to:

$$\begin{aligned} V_2 =& V_1 - R_{\mathrm{s1}}\left(I_2 - I_1\right) - \kappa I_2 \left(T_2 - T_1\right) \\ &+ V_{\mathrm{oc},0}\left(\beta\left[f(G_2)(T_2 - T_0) - f(G_1)(T_1 - T_0)\right]\right. \\ &\left. + \frac{1}{f(G_2)} - \frac{1}{f(G_1)}\right), \end{aligned} \tag{6}$$

with:

$$R_{\mathrm{s1}} = R_{\mathrm{s}} + \kappa(T_1 - T_0), \tag{7}$$

and:

$$f(G) = \frac{V_{\mathrm{oc},0}}{V_{\mathrm{oc}}(G)} = B_2 \ln^2\left(\frac{G_0}{G}\right) + B_1 \ln\left(\frac{G_0}{G}\right) + 1. \tag{8}$$

The coefficients used for voltage translation include:

- α: temperature coefficient of the short-circuit current
- R_{s}: internal series resistance
- κ: temperature coefficient of R_{s}
- $V_{\mathrm{oc},0}$: open-circuit voltage at STC
- β: temperature coefficient of the open-circuit voltage
- B_1, B_2: irradiance correction coefficients

We apply this procedure to translate the short-circuit current (I_{sc}) from the measured conditions to STC, i.e., T_0 = 25 °C and G_0 = 1000 W/m^2. The standard deviation of the corrected I_{sc} values is used as a metric to assess the performance of our filtering routine. Both the

parameterization of Procedure 2 from IEC 60891:2021 and the I_{sc} translation to STC are implemented using the PV-CRAZE library [17,28]. All other analyses were implemented using Python 3.8.20 on Ubuntu Linux with major open-source libraries including NumPy 1.24.4, Pandas 2.0.3, UMAP (umap-learn) 0.5.7, HDBSCAN 0.8.40, Matplotlib 3.7.3, and Scikit-learn 1.3.2, all available via the Python Package Index.

3. Datasets

3.1. Used data

The dataset utilized in this study comprises outdoor meteorological and photovoltaic (PV) module performance data for a crystalline silicon module, collected in Tempe, Arizona, USA. It includes ambient temperature, Global Horizontal Irradiation (G_{GH}), Diffuse Horizontal Irradiation (G_{DH}), Plane of Array Irradiation (G_{POA}), back-of-module temperature, and current–voltage (IV) characteristics. Spanning from January 1, 2015, to July 1, 2017, it features 60,372 complete records, each with a 15-minute resolution during daylight hours. For an in-depth description of the dataset, we refer to Schweiger et al. [29].

3.2. Data processing: normalization and scaling

For this study, we concentrated on data directly measured from the photovoltaic (PV) modules, specifically the IV data, module temperature T, and G_{POA}. We characterized the IV data using Extended Solar Cell Parameters (ESPs), a 10-parameter feature vector extracted from the IV curves to encapsulate the shape of IV characteristics more comprehensively [17]. The ESPs extend standard solar cell parameters, such as open-circuit voltage (V_{oc}), short-circuit current (I_{sc}), voltage at maximum power point (V_{mpp}), and current at maximum power point (I_{mpp}), by including six additional parameters for a more detailed description of the IV data. The added parameters are the conductance at short circuit (G_{sc}), resistance at open circuit (R_{oc}), and two sets of "quasi maximum power points" (qmpp- with V_{qmp-} and I_{qmp-}, and qmpp+ with V_{qmp+} and I_{qmp+}).

Along with irradiance (G_{POA}) and temperature (T), the ESPs amount to a total of 12 dimensions. Given the heterogeneity of measurement units across these dimensions, scaling the data prior to analysis with algorithms such as LOF and UMAP is required. A standard scaling approach subtracts the mean and normalizes against the standard deviation. However, this approach can obscure meaningful data relationships when dimensions share units, such as voltage or current, or exhibit strong dependencies, like G_{POA} with I_{sc}. Moreover, we observe that the ESP voltages are strongly correlated with V_{oc} and the currents with I_{sc}, suggesting that much of the variance across the 12 dimensions reflects information predominantly about two dimensions: T and G_{POA}.

To mitigate these issues, we adjusted our scaling strategy by first aligning dimensions with shared units and strong dependencies. Specifically, all ESP voltages, except V_{oc}, were scaled relative to their own V_{oc}, and all currents, excluding I_{sc}, were adjusted according to their own I_{sc}. This scheme also applied to G_{sc} and R_{oc}, which were normalized against I_{sc}/V_{oc} and V_{oc}/I_{sc}, respectively. Note that this scaling makes the IV representation consist of the V_{oc}, the I_{sc}, and 8 dimensionless parameters describing the shape of the IV. After this, we normalized I_{sc} relative to G_{POA} to further disentangle these dependencies. Following this targeted scaling, the dataset exhibited reduced inter-parameter correlations, allowing us to employ a standard scaling protocol across all dimensions, thereby ensuring a more meaningful and less biased analytical outcome.

4. Results

4.1. UMAP and clustering of data

Before we present our results on filtering, we first demonstrate the use of UMAP to visualize the dataset and how it embeds the data in a low-dimensional space while keeping similar points close to each other.

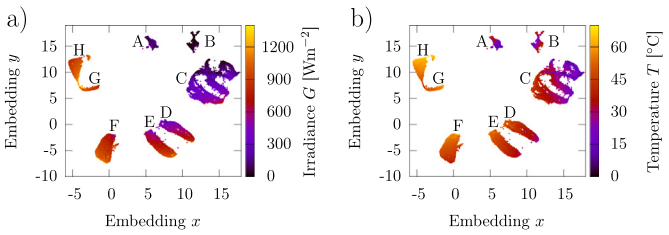


Fig. 2. UMAP embedding (with 500 neighbors per sample and minimum distance of 0.0 as hyperparameters) showing the map of our data colored according to (a) Irradiance and (b) Temperature dimensions. UMAP performed an unsupervised classification of our data into various clusters, A – H, of low, mid and high irradiance and temperatures.

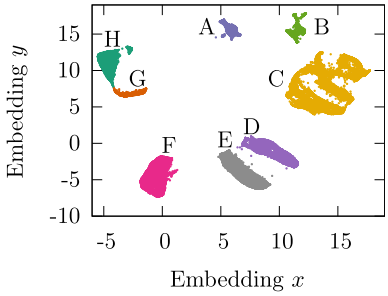


Fig. 3. UMAP embedding clustered with HDBSCAN with minimum cluster size of 500 data points as a hyperparameter (where smaller cluster sizes are considered as noise). The different clusters are colored and labeled A – H for the ease of identification and analysis.

In Fig. 2, we show the computed UMAP embedding of the complete dataset where we colored the datapoints according to the irradiance (Fig. 2(a)) and temperatures (Fig. 2(b)). Note that for the projection UMAP uses all 12 dimensions, however, we use the irradiance and temperature values to illustrate how UMAP sorts the data in a low dimensional space, such that similar points are close to each other. The UMAP embedding emphasizes the clustering of datapoints, resulting in several disjoint clusters of points. The overall irradiance and temperature vary between the different clusters. As one would expect, the highest module temperatures are observed in the cluster with the highest irradiance.

Interestingly, one can observe that within the clusters, the irradiance and temperatures are also sorted, with some clusters having opposite trends in irradiance and temperature (i.e., lower temperatures at higher irradiance levels). This demonstrates UMAP's ability to effectively cluster data by similarity, capturing the principal trend of rising temperatures with increased irradiance while also preserving the range of temperatures observed at similar irradiance levels.

In Fig. 3, we show the UMAP embedding in the two-dimensional reduced space, where the data have been grouped into eight classes, labeled A–H, using the HDBSCAN clustering algorithm. For clustering, we applied a minimum cluster size of 500 data points. As discussed in Section 2.3, the use of UMAP prior to clustering results in clearly separated clusters and minimal noise—only about 1% of the data points were classified as noise. This increased separability is due to UMAP's tendency to introduce discontinuities in the data manifold, which can artificially enhance cluster boundaries even in otherwise continuous data. Nevertheless, the outcome is a fully unsupervised classification of the data points, which provides a practical basis for downstream analysis.

Having obtained clusters from the dataset, we now examine the properties of the individual classes. Fig. 4(a–d) shows representative IV curves randomly selected from clusters A, B, E, and H. As illustrated

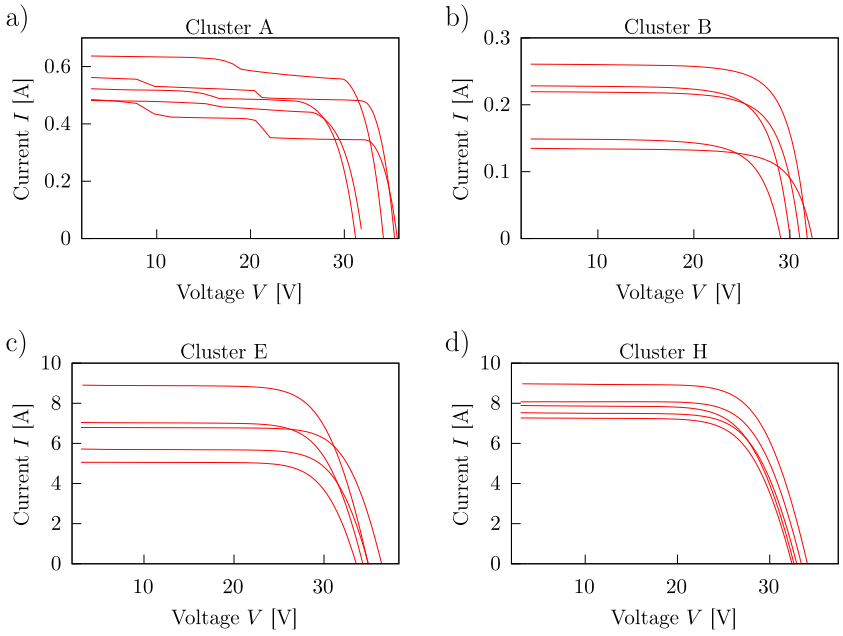


Fig. 4. (a–d) Exemplary IV curves corresponding to the clusters A, B, E and H in Fig. 2. Cluster A consists of partially shaded IV curves. Cluster E and H have a similar property of high currents. Cluster B consists of low light IV curves, exhibiting lower currents and a few series resistance effects.

Table 1
Cluster characteristics.

Cluster	Characteristics
A	Low irradiance and partial shading cluster obtained during mostly sunny days (see Fig. 4a).
B	Very low irradiance cluster on cloudy days.
C	A mix of low and mid irradiance cluster during partly cloudy and sunny days.
D–H	Consecutively increasing irradiance and temperature clusters.

in Fig. 4(a), cluster A primarily consists of IV curves affected by partial shading conditions, characterized by step-like features due to the activation of bypass diodes. In general, the irradiance is low for this cluster, which is expected since shading typically occurs when the solar elevation is low and shadows are long.

Cluster B also represents low-light conditions, however, without partial shading conditions. These points are predominantly associated with overcast skies. The cloudiness index for these instances—defined as the ratio of diffuse horizontal irradiance (DHI) to global horizontal irradiance (GHI)—is approximately equal to unity. This indicates that DHI ≈ GHI, meaning the incoming radiation is entirely diffuse and there is no direct sunlight.

Fig. 4(c) corresponds to cluster E, which represents higher irradiance conditions, as indicated by the increased current levels compared to clusters A and B. Fig. 4(d) shows representative IV characteristics from cluster H. Compared to cluster E, the curves in cluster H exhibit a lower open-circuit voltage (V_{oc}), which is consistent with the higher module temperatures observed for this cluster (see also Fig. 2b).

Overall, the unsupervised clustering resulted in physically meaningful groupings, successfully identifying distinct modes of operation within the dataset. A summary of the clusters and their typical characteristics in terms of irradiance and temperature is provided in Table 1.

At this point, we wish to highlight the significance of the achieved classification, particularly in the context of data filtering. Cluster A, which captures partial shading conditions, serves as a compelling example of the UMAP/HDBSCAN combination's ability to identify data points that, in line with Hawkins's definition of an anomaly, appear to be generated by "a different mechanism." Importantly, this does not imply that entire clusters—such as cluster A—should be labeled as anomalous, as the cluster represents a common mode of operation and is not characterized by noise or faults. Rather, the value of the unsupervised classification lies in its ability to differentiate between distinct operational modes within the dataset.

To validate our overall unsupervised classification methodology, we applied the same approach to an independent dataset from a module in Ancona, Italy. The results of this classification, demonstrating robust performance across datasets, are detailed in the supplementary material.

In subsequent filtering steps, this knowledge enables targeted exclusion of specific clusters, depending on the filtering objective. However, it should be noted that the classification is not well-suited for detecting rare anomalies or faults, as the use of UMAP significantly reduces the proportion of data points classified as noise.

4.2. Data filtering

To detect rare anomalies, we apply the LOF algorithm in the normalized 12-dimensional feature space. LOF identifies data points whose (locally approximated) density differs significantly from that of their neighbors, making it well-suited for spotting isolated or unusual patterns.

Fig. 5(a) shows the anomalies detected by LOF (in red) overlaid on the UMAP embedding (in gray). The anomalies span a broad range of irradiance (G_{POA}) and temperature (T), affecting multiple clusters. Notably, some clusters are less susceptible to outliers—for example, cluster B contains very few anomalous points.

As we benchmark our approach compared to a conventional thresholding method, Fig. 5(b) illustrates the effect of filtering data with $G_{POA} < 300\,W/m^2$. This simple thresholding effectively removes cluster B and significantly reduces the size of clusters A and C. However, it is worth noting that portions of cluster A—associated with partial shading—still remain after thresholding. This highlights a limitation of basic irradiance thresholds—they may fail to capture more nuanced or context-dependent data patterns.

To illustrate the impact of LOF-based filtering, we apply IEC 60,891 Procedure 2 to correct all measured data to standard test conditions (STC). The standard deviation of the corrected short-circuit current is

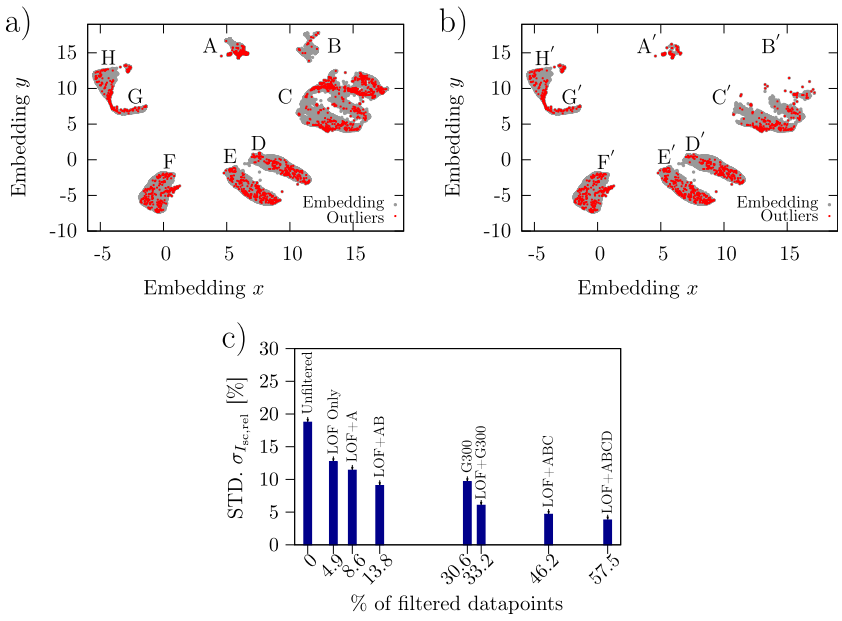


Fig. 5. (a) UMAP embedding (in gray) with anomalies detected using LOF highlighted in red. The anomalies are distributed across a wide range of G_{POA} and T, with some clusters more prone to outliers than others. (b) Data filtered using a threshold of $G_{POA} > 300\,W/m^2$. (c) Bar plot showing the normalized standard deviation of the corrected short-circuit current ($\sigma_{I_{sc,rel}}$) expressed as a percentage of $I_{sc,STC}$, plotted against the percentage of filtered data points. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

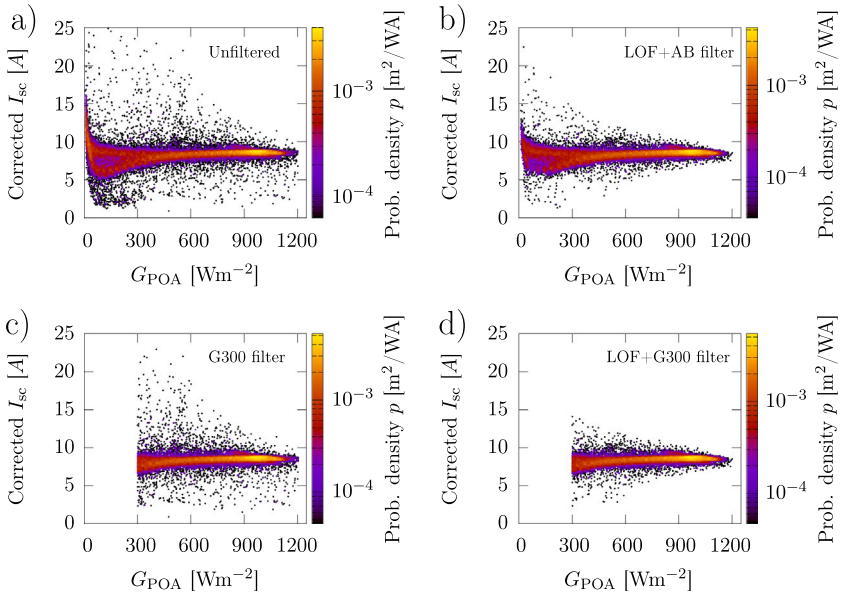


Fig. 6. Scatterdensity plots of the corrected I_{sc} against the G_{POA} for the (a) unfiltered data (b) LOF + AB filter (c) G300 filter and (d) LOF + G300 filter. The color of each scatter point represents the estimated probability density on a logarithmic scale, indicating the likelihood of a point occupying a specific location in the plot. The density integrates to 1 over the entire dataset, with units of $1/(W/m^2 \cdot A)$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

normalized to the nominal $I_{sc,STC}$ ($\sigma_{I_{sc,rel}}$) and expressed as a percentage, providing a dimensionless metric that reflects the variability of the data relative to its expected STC value. This normalized percentage is plotted against the percentage of filtered data points as shown in Fig. 5(c), enabling its use as a metric for data quality. Higher values indicate greater dispersion (noise) and therefore reduced measurement consistency, while lower values reflect a more stable and cleaner dataset. Thus, the plot in Fig. 5(c) captures the trade-off between data retention and residual noise in the filtered dataset.

Furthermore, Fig. 5(c) compares several filtering strategies, including simple thresholding at $G_{POA} > 300\,W/m^2$, LOF-based filtering, and the removal of specific clusters. By combining these approaches, a Pareto front of optimal filtering strategies can be constructed to balance noise suppression with data retention. In this context, clustering enables a modular approach to filtering, allowing different components (e.g., LOF, cluster exclusions, thresholding) to be mixed and matched according to the needs of downstream analysis.

Compared to pure thresholding, LOF-based filtering combined with the removal of specific clusters proves highly effective. For instance, the "LOF + AB" filter — applying LOF alongside the exclusion of clusters A and B—achieves a lower standard deviation than thresholding, while discarding only 13.8% (8,251) data points. In contrast, the thresholding approach removes about 30.6% (18,290) which is more than twice as many points. To further assess the operational implications of the proposed filtering strategy, its effect is examined in the time domain using a representative one-week segment of PV data. A comparative illustration against the traditional irradiance threshold filtering is provided in Fig. B.8 in Appendix B.

In Figs. 6(a–d), we present scatter density plots of the corrected I_{sc} under standard test conditions as a function of irradiance. Fig. 6(a) shows the unfiltered dataset. Note that the color scale reflects the estimated density of points on a logarithmic scale. At lower irradiance levels, the spread in the corrected I_{sc} increases significantly. Below 300 W/m^2, the data appear to bifurcate into two branches: one that remains near the standard I_{sc} value of approximately 9 A, and a second, lower branch extending down to around 6 A. The lower branch is attributed to partial shading effects [12].

Fig. 6(b) shows the results after applying the "LOF + AB" filter. The most notable changes include a clear reduction in overall scatter and the removal of many points associated with the lower branch. Fig. 6(c) displays the outcome of applying the simple thresholding filter, while Fig. 6(d) shows thresholding combined with LOF. A comparison between Figs. 6(c) and (d) reveals that LOF effectively eliminates isolated data points that deviate significantly from the expected I_{sc} under standard test conditions.

5. Conclusions

We presented a method based on UMAP and HDBSCAN to achieve an unsupervised classification of outdoor PV performance data. The method was demonstrated on a dataset collected from a crystalline silicon module operating under outdoor conditions in Arizona. The unsupervised classification resulted in physically meaningful groupings, including distinct clusters corresponding to partial shading, low irradiance, and high irradiance conditions. A key advantage of this approach is its ability to automatically identify and separate different modes of operation present in the data.

For anomaly removal, we employed LOF-based filtering. We demonstrated that the targeted removal of specific classes—such as those associated with partial shading or low irradiance—enables the construction of flexible, application-specific filters. This was illustrated by correcting the short-circuit current to standard test conditions and using the standard deviation of the corrected current as a metric for noise.

Our results show that combining LOF with the removal of selected clusters outperforms traditional threshold-based filtering (e.g., $G_{POA} > 300$ W/m^2) in terms of noise reduction, while discarding less than half as many data points. This demonstrates the potential of our approach to create more accurate, efficient, and targeted data filtering strategies for downstream PV data analysis.

CRediT authorship contribution statement

Goodfriend M. Whyte: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Conceptualization. **Andreas Gerber:** Resources, Project administration, Funding acquisition. **Uwe Rau:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition. **Bart E. Pieters:** Writing – review & editing, Validation, Supervision, Methodology, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work received sponsorship and funding from the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF) under the YESPV–NIGBEN–CLIENT II project (BMBF Förderkennzeichen 03SF0576A-B) and H2ATLAS-AFRICA project (03EW0001C). Open Access funding was enabled and organized by Projekt DEAL.

Appendix A. Cluster validation via silhouette analysis

A.1. Methodology

Silhouette analysis was employed to quantitatively evaluate clustering quality across different feature representations. The silhouette coefficient measures both cluster cohesion (intra-cluster similarity) and separation (inter-cluster dissimilarity) for each data point, computed as in Eq. (A.1):

$$s_i = \frac{q_i - p_i}{\max(p_i, q_i)} \tag{A.1}$$

where p_i represents the mean distance to points within the same cluster and q_i represents the mean distance to points in the nearest neighboring cluster. The global silhouette score ranges from −1 (poor clustering) to +1 (excellent clustering).

A.2. Comparative results

Silhouette scores were computed for clustering results in both the original high-dimensional space and the UMAP space.

A.3. Visual analysis

Fig. A.7 illustrates the distribution of silhouette scores across clusters in both representations. The UMAP-embedded space demonstrates superior clustering performance, with most clusters exhibiting positive silhouette scores and clearer separation.

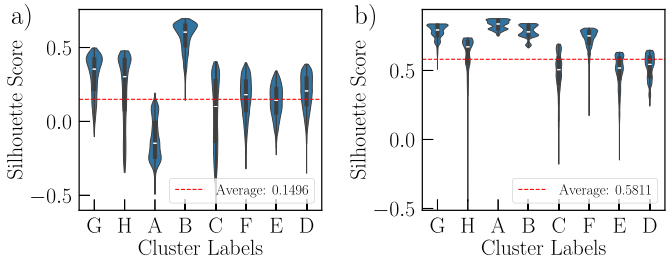


Fig. A.7. Distribution of silhouette scores across clusters in (a) original feature space and (b) UMAP-embedded space. The red broken line represents the mean silhouette score for all the clusters. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

A.4. Discussion

The substantial improvement in silhouette score from 0.1496 to 0.5811 as shown in Table A.2, indicates that while UMAP may enhance apparent cluster separation, the embedded representation captures meaningful underlying structure that exists in the original data, albeit less distinctly.

Table A.2
Comparative silhouette analysis across feature spaces.

Feature space	Silhouette score	Interpretation
Original space	0.1496	Weak cluster structure
UMAP space	0.5811	Reasonable cluster structure

Appendix B. Timeseries analysis

B.1. Time-series illustration of filtering strategies

Fig. B.8 illustrates the effect of different data filtering strategies on a representative one-week segment of PV module measurements. The week was selected to include periods of moderate-to-high irradiance where partial shading and irregular behavior are known to occur.

As earlier noted, the traditional threshold-based filter discards all data points with $G_{\mathrm{POA}} < 300\ \mathrm{W/m^2}$, regardless of the electrical response. In contrast, the proposed approach combines Local Outlier Factor (LOF) anomaly detection with cluster information obtained from HDBSCAN, allowing selective removal of locally inconsistent operating points and cluster-specific abnormal behavior.

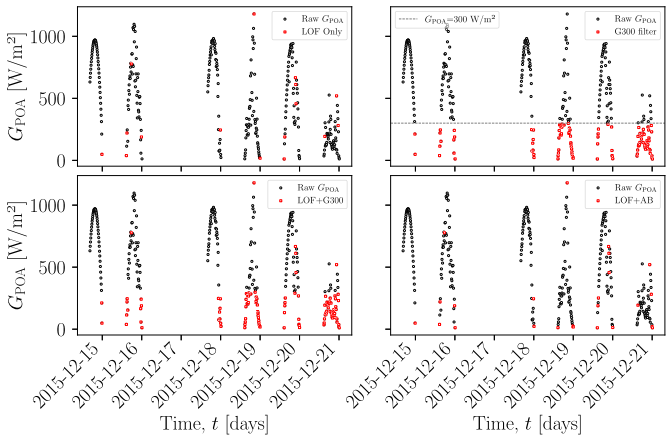


Fig. B.8. Comparison of traditional irradiance threshold filtering and the proposed hybrid filtering approach on a representative one-week time series. Black markers indicate the raw plane-of-array irradiance G_{POA}. Red markers denote data points discarded by each filtering strategy. Top row: (left) LOF-based anomaly detection only; (right) traditional irradiance threshold filtering ($G_{\mathrm{POA}} < 300\ \mathrm{W/m^2}$). Bottom row: (left) combined LOF and irradiance threshold filtering; (right) proposed hybrid filter combining LOF-based anomalies with selected HDBSCAN clusters associated with partial shading and abnormal operation. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

By visualizing discarded and retained data directly in the time domain, the figure demonstrates that our proposed framework removes anomalous and shaded operating points while retaining physically consistent measurements that would otherwise be discarded by a simple irradiance threshold.

Appendix C. Supplementary data

Supplementary data for this article can be found online at doi:10.1016/j.solener.2026.114351.

Data availability

The authors do not have permission to share data.

References

[1] L. Bommes, M. Hoffmann, C. Buerhop-Lutz, T. Pickel, J. Hauch, C. Brabec, A. Maier, I. Marius Peters, Anomaly detection in IR images of PV modules using supervised contrastive learning, Prog. Photovolt. Res. Appl. 30 (6) (2022) 597–614.
[2] S. Lindig, A. Louwen, D. Moser, M. Topic, Outdoor PV system monitoring—input data quality, data imputation and filtering approaches, Energies 13 (19) (2020) 5099.
[3] D.M. Hawkins, Identification of Outliers, vol. 11, Springer, 1980.
[4] M.B. Øgaard, H.N. Riise, H. Haug, S. Sartori, J.H. Selj, Photovoltaic system monitoring for high latitude locations, Sol. Energy 207 (2020) 1045–1054.
[5] D.C. Jordan, C. Deline, S.R. Kurtz, G.M. Kimball, M. Anderson, Robust PV degradation methodology and application, IEEE J. Photovolt. 8 (2) (2017) 525–531.
[6] S. Silvestre, L. Mora-López, S. Kichou, F. Sánchez-Pacheco, M. Dominguez-Pumar, Remote supervision and fault detection on OPC monitored PV systems, Solar Energy 137 (2016) 424–433.
[7] G. Belluardo, P. Ingenhoven, W. Sparber, J. Wagner, P. Weihs, D. Moser, Novel method for the improvement in the evaluation of outdoor performance loss rate in different PV technologies and comparison with two other methods, Sol. Energy 117 (2015) 139–152.
[8] Photovoltaic system performance–part 1: monitoring, 2017. [Online]. Available: International Electrotechnical Commission, Geneva, CH, Standard IEC 61724-1, https://webstore.iec.ch/en/publication/65561
[9] R. Todeschini, D. Ballabio, V. Consonni, F. Sahigara, P. Filzmoser, Locally centred mahalanobis distance: a new distance measure with salient features towards outlier detection, Anal. Chim. Acta 787 (2013) 1–9.
[10] R. Turudija, D. Stojiljković, M. Zdravković, M. Ignjatović, Towards an approach to multivariate outlier detection for district heating system data, in: Conference on Information Technology and Its Applications, Springer, 2024, pp. 49–61.
[11] F. Harrou, A. Dairi, B. Taghezouit, B. Khaldi, Y. Sun, Automatic fault detection in grid-connected photovoltaic systems via variational autoencoder-based monitoring, Energy Convers. Manag. 314 (2024) 118665.
[12] T.S. Vaas, J. Körtgen, E. Sovetkin, U. Rau, B.E. Pieters, Plausibility filtering of PV outdoor data, in: 2023 IEEE 50th Photovoltaic Specialists Conference (PVSC), IEEE, 2023, pp. 1–4.
[13] H. Demirhan, Z. Renwick, Missing value imputation for short to mid-term horizontal solar irradiance data, Appl. Energy. 225 (2018) 998–1012.
[14] C.C. Turrado, M.D.C.M. López, F.S. Lasheras, B.A.R. Gómez, J.L.C. Rollé, F.J. de Cos Juez, Missing data imputation of solar radiation data under different atmospheric conditions, Sensors 14 (11) (2014) 20382–20399.
[15] S.K. Kwak, J.H. Kim, Statistical data preparation: management of missing values and outliers, Korean J. Anesthesiol. 70 (4) (2017) 407.
[16] M.M. Breunig, H.-P. Kriegel, R.T. Ng, J. Sander, LOF: identifying density-based local outliers, in: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, 2000, pp. 93–104.
[17] B. Pieters, Extended solar cell parameters-general purpose descriptive I/V parameters for solar cells, Authorea Preprints, 2023.
[18] L. McInnes, J. Healy, J. Melville, UMAP: Uniform manifold approximation and projection for dimension reduction, arXiv preprint arXiv:1802.03426, 2018.
[19] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I.W.H. Kwok, L.G. Ng, F. Ginhoux, E.W. Newell, Dimensionality reduction for visualizing single-cell data using UMAP, Nat. Biotechnol. 37 (1) (2019) 38–44.
[20] L. Van Der Maaten, Accelerating t-SNE using tree-based algorithms, J. Mach. Learn. Res. 15 (1) (2014) 3221–3245.
[21] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (11) (2008).
[22] R.J.G.B. Campello, D. Moulavi, J. Sander, Density-based clustering based on hierarchical density estimates, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2013, pp. 160–172.
[23] L. McInnes, J. Healy, Accelerated hierarchical density based clustering, in: 2017 IEEE International Conference on Data Mining Workshops (ICDMW), IEEE, 2017, pp. 33–42.
[24] A. Starczewski, P. Goetzen, M.J. Er, A new method for automatic determining of the DBSCAN parameters, J. Artif. Intell. Soft Comput. Res. 10 (2020).
[25] Using UMAP for clustering, https://umap-learn.readthedocs.io/en/latest/clustering.html (accessed: 26 March 2026).
[26] M. Allaoui, M.L. Kherfi, A. Cheriet, Considerably improving clustering algorithms using UMAP dimensionality reduction technique: a comparative study, in: International Conference on Image and Signal Processing, Springer, 2020, pp. 317–325.
[27] Photovoltaic devices - procedures for temperature and irradiance corrections to measured I-V characteristics, 2021. [Online]. Available: International Electrotechnical Commission, Geneva, CH, Standard IEC 60891, https://webstore.iec.ch/en/publication/61766
[28] B.E. Pieters, PV-CRAZE, 2022, https://github.com/IEK-5/PV-CRAZE
[29] M. Schweiger, J. Bonilla, W. Herrmann, A. Gerber, U. Rau, Performance stability of photovoltaic modules in different climates, Prog. Photovolt. Res. Appl. 25 (12) (2017) 968–981.