PVeSight: Dimensionality Reduction-Based Anomaly Detection and Visual Analysis of Photovoltaic Strings

Yurun Yang[a], Xinjing Yi[a], Yingqiang Jin[b], Sen Li[b], Kang Ma[c], Shuhan Liu[d], Dazhen Deng[a], Di Weng[a,*] and Yingcai Wu[d]

[a]School of Software Technology, Zhejiang University, Ningbo, Zhejiang, China
[b]DaTang Zhejiang Branch, Hangzhou, Zhejiang, China
[c]China Datang Corporation Technology Innovation Co., Ltd., Xiong'an New Area, Hebei, China
[d]State Key Lab of CAD & CG, Zhejiang University, Hangzhou, Zhejiang, China

ARTICLE INFO

Keywords:
Photovoltaic string
Anomaly detection
Visual analytics
Dimensionality reduction pattern

ABSTRACT

Accurate identification and diagnosis of inefficient photovoltaic (PV) strings are essential for ensuring the stable operation of PV power stations. Existing studies primarily focus on automated anomaly detection models based on temporal abnormalities. However, since anomaly analysis often relies on domain knowledge, current methods have significant limitations in assisting experts in understanding the causes and impacts of anomalies. In close collaboration with domain experts, this study systematically identifies the specific user requirements for PV string anomaly detection and proposes an innovative workflow for identifying and diagnosing inefficient PV strings. Furthermore, we develop an interactive visual analytics system, PVeSight, to support experts in efficiently analyzing and diagnosing PV string anomalies. The system leverages dimensionality reduction techniques and weather clustering methods to generate string pattern maps, which are then utilized for anomaly detection, anomaly classification, comparative analysis between PV strings, and degradation rate assessment. This approach enables experts to accurately identify inefficient PV strings, trace the root causes of anomalies, and gain valuable insights. Through case studies and evaluation experiments, we validate the usability and effectiveness of PVeSight in PV string anomaly detection.

1. Introduction

In the contemporary era, with the continuous depletion of fossil fuels and the increasing severity of environmental pollution, the utilization of renewable energy has received unprecedented attention. Solar energy, known for its cleanliness, inexhaustibility, and economic advantages, is becoming a crucial component of the global energy structure. PV power generation technology is an effective means of converting solar energy into electrical energy and plays a vital role in the renewable energy sector.

However, with the commissioning and operation of a large number of PV power stations, anomalies in core equipment like PV strings are emerging incessantly. Anomalies in PV systems can reduce power generation efficiency and pose safety risks, including fire hazards. Moreover, as installed capacity significantly increases, the method of assigning staff to inspect equipment device by device becomes impractical due to manpower shortages. Therefore, developing effective intelligent anomaly detection technologies for PV strings is crucial.

In recent years, to enhance the performance of PV power stations, researchers have proposed various advanced methods for detecting anomalies in PV strings. In traditional methods, Simon and Meyer (2010) used infrared thermography to map the surface temperature distribution of solar cells under reverse bias. They proposed a fault diagnosis method based on infrared image analysis. Platon et al. (2015) investigated the operational status of PV strings under various illumination conditions by simulating different surface temperatures and irradiance levels, and comparing theoretical normal values with actual measurements to identify anomalies. However, the accuracy and determination of anomalies heavily depend on expert-set thresholds and involve some randomness. With the development of artificial intelligence technologies, traditional PV string anomaly detection techniques have seen new advancements. Chine et al. (2016) proposed an anomaly diagnosis method for PV strings using artificial neural networks (ANNs). This method involves training neural networks using characteristic data such as current, voltage, and peak values from the I-V curve of PV strings under given irradiance and temperature conditions, thus diagnosing string faults. Although this approach has improved the accuracy of anomaly detection, it is highly dependent on the quality of the sample data.

The integration of artificial intelligence with PV strings anomaly detection improves the model's accuracy but suffers from a lack of interpretability due to its black-box structure (Castelvecchi, 2016). Data visualization encodes the attributes of data in intuitive visual charts (Liu et al., 2014), incorporating human perception into data analysis through human-machine interaction, thereby enhancing accuracy and interpretability. Therefore, this paper applies visual analytics technology to the anomaly detection of PV

*Corresponding author
yurunyang@zju.edu.cn (Y. Yang); yixinjing@zju.edu.cn (X. Yi); 24039221@qq.com (Y. Jin); lsien323@163.com (S. Li); 13436362144@139.com (K. Ma); shliu@zju.edu.cn (S. Liu); dengdazhen@zju.edu.cn (D. Deng); dweng@zju.edu.cn (D. Weng); ycwu@zju.edu.cn (Y. Wu)
ORCID(s): 0009-0009-8118-0827 (Y. Yang)

strings to delve deeper into the anomalies and their patterns in the strings.

We propose a hierarchical anomaly detection method based on dimensionality-reduced pattern maps. This method utilizes the dimensionality reduction algorithm UMAP to generate reduced pattern maps from time-series electrical data of PV strings. Additionally, environmental data are incorporated through K-Means clustering labels to further enhance the dataset. The anomaly detection model is trained hierarchically, considering the structural levels of the PV station as well as weather classifications. Unlike common time-series data anomaly detection methods, our approach does not require drone inspection images or I-V scan results. We employ an unsupervised model to address challenges such as the lack of labeled data, diverse anomaly types, and large-scale data. After discussing with PV field experts and summarizing actual needs, a visual analytics system comprising five views was designed to assist users in uncovering and analyzing anomalies and patterns in PV strings. The visual analytics system was applied to four months of electrical and environmental data from a large PV station in a real-world scenario. The system demonstrated its effectiveness in aiding users to detect and verify anomalies, analyze situations, and explore string operation conditions. The system's usability and effectiveness were evaluated through multiple case studies.

Our contributions are summarized as follows:

- We propose a novel anomaly detection method for PV strings, utilizing dimensionality reduction and hierarchical methods to improve the model's accuracy in identifying anomalous strings.
- We developed a visual analytics system that provides carefully designed visualizations and rich interactions to explain the model's anomaly detection results, facilitating efficient analysis and annotation of anomalous strings by users.
- We conducted three case studies and evaluations to verify the usability and effectiveness of the proposed model and system.

2. Related Work

2.1. Photovoltaic string anomaly detection

PV power generation systems are an critical component of the development of new energy systems. The energy flow hierarchy of the system includes PV power stations, combiner boxes, inverters, PV arrays, PV strings, and PV modules. PV strings are a crucial component, and their operational stability determines the overall efficiency of the power generation system. To ensure stable operation, intelligent PV string anomaly detection technology is widely applied in PV operation and maintenance (Tsanakas et al., 2016). Since PV power stations generate a significant amount of time-series data, including current, voltage, and I-V curves, which PV string anomaly detection primarily relies on. Features are extracted from these data to improve the accuracy of detection, considering cost and feasibility (Eskandari et al., 2023).

Common approaches to detecting anomalies in PV string systems involve several traditional methods, including ground capacitance measurement (Takashima et al., 2008), infrared imaging detection (Gallardo-Saavedra et al., 2018), electroluminescence imaging (Otamendi et al., 2021), and time-domain reflectometry (Roy et al., 2018). Although such methods can diagnose and locate faulty strings, they require significant investment in sensor equipment in terms of performance and quantity. The practical application costs increase with the scale of the power station, and there are limitations on the types of faults covered (Zhu et al., 2019).

To reduce costs and improve detection accuracy, researchers have proposed combining expert knowledge with mathematical models. By comparing the differences between the outputs of theoretical models and actual observations, judgment thresholds are set based on expert experience and reference data, and these thresholds are applied for anomaly detection of strings (Silvestre et al., 2013; Drews et al., 2007). These efforts have somewhat reduced costs and improved fault detection rates. However, threshold selection heavily depends on the model, introducing randomness in performance. Additionally, the mapping from operational status to threshold variables, based on data and expert experience, lacks generalizability across different systems, limiting its practical applicability.

With improvements in computer hardware performance and rapid developments in artificial intelligence technologies, numerous anomaly detection models and algorithms such as Support Vector Machines (SVM) (Cortes and Vapnik, 1995), Random Forest (Breiman, 2001), deep learning models like Convolutional Neural Networks (CNN) (Kramer, 1992), and AutoEncoders (Krizhevsky et al., 2017) have been proposed. Supervised methods (Manno et al., 2021) require a large labeled dataset, which is challenging to satisfy in PV scenarios due to the scarcity of fault samples. Similarly, semi-supervised methods (Zhao et al., 2013) also rely on labeled data. Unsupervised methods (Zhu et al., 2018), although not requiring labeled samples, still lag in accuracy compared to the former two. Overall, methods incorporating artificial intelligence have shown certain improvements in performance and generalizability, but challenges still exist when applied to large-scale time-series PV datasets.

2.2. Time series visual analysis

Time series data consists of sequences of data points arranged in chronological order, with each point containing the state occurring at that moment. Based on the number of state attributes, time series can be classified into univariate and multivariate time series. Common time series data in the real world, such as log data, stock and financial data, and energy data, are usually sampled at fixed time frequencies (e.g., yearly, monthly, daily). Effective analysis of time series helps to uncover the underlying structures and patterns, and

is widely applied in scientific and engineering fields. For example, in the fields of economics and finance, models like autoregressive integrated moving average (ARIMA) and autoregressive conditional heteroskedasticity (ARCH) are used to forecast the price fluctuations of financial assets; in stock analysis, methods such as template matching and Dynamic Time Warping (DTW) identify recurring patterns or shapes in time series; and anomalies in time series are identified using statistical methods like the 3-sigma rule and Grubbs's Test, similarity-based methods like DBSCAN, LOF, Isolation Forest, as well as machine learning models including generative AutoEncoders and discriminative LSTMs.

As the scale and dimensions of time-series data increase, they present challenges with varied structures and complex content for data analysis tasks. Interactive visual analysis techniques combine visual chart information with human perception (Kohlhammer et al., 2011), aiding tasks such as pattern recognition, comparative analysis, and result traceability. Fan et al. (2017) developed a visual analysis system based on smartphone data, introducing a multi-view interactive approach to explore users' behavior patterns with their phones and analyze the causes of specific events. Malik et al. (2016) employed an efficient hypothesis testing method to systematically explore and compare relationships between large-scale event sequences, also conducting tests on medical datasets. Franklin et al. (2016) assisted healthcare professionals and patients in discussing and choosing the most suitable treatment options. By analyzing the event sequences related to different drug efficacies and side effects, it summarizes suitable plans, aiding users in decision-making.

Anomaly detection is also a major focus in time-series visual data analysis research. Thom et al. (2012) proposed a visual analysis method for spatio-temporal anomaly detection using geolocated Twitter messages. This method extracts keywords, time, and coordinates from tweets, applies clustering to obtain grouped tag clouds, and employs statistical analysis to identify clusters with significantly different patterns. Riveiro et al. (2008) performed anomaly detection on ships based on sensor data, integrating multiple sources such as radar and satellite images. They detected attributes such as ship speed and historical behaviors of ships docking, meeting the needs for processing large volumes of information and rapidly identifying anomalous behaviors. Cao et al. (2016) addressed large-scale data management and user security on social platforms by extracting behavioral characteristics from users' text, login, and interaction information. Anomaly detection algorithms were employed to identify suspicious users, and their behavioral traits were visually encoded for display. These studies offer valuable insights that we used to design our PV data-based visual anomaly detection system.

3. Background

3.1. Problem Statement

During the operation and maintenance of PV power stations, certain PV strings may experience a significant reduction in power generation efficiency due to environmental factors, equipment aging, or operational anomalies, resulting in inefficient strings. Timely identification of these inefficient strings and accurate diagnosis of their root causes are crucial for ensuring the long-term stable operation of PV power stations. However, this task poses several challenges, primarily in the following three aspects:

Q1. How to effectively utilize massive power station operation data for inefficient string identification in the absence of labeled data? PV power stations are typically equipped with monitoring systems that collect multi-dimensional operational data in real-time, including direct current (I_{dc}), voltage (V_{dc}), alternating current power (P_{ac}), ambient temperature (T), and irradiance (G). The state of a PV string at time t is represented as $X_t^i = \{I_{dc}, V_{dc}, P_{ac}, T, G, \dots\} \in \mathbb{R}^d$, while the overall state of all N strings in the power station is given by $X_t \in \mathbb{R}^{N\times d}$. Although this data contains rich information, the complexity of inefficient string manifestations makes traditional threshold-based methods ineffective, while existing intelligent approaches rely heavily on labeled data. However, in practical applications, labeled data is extremely scarce, and existing methods often depend on additional monitoring techniques (e.g., infrared imaging, I-V curve scanning), increasing operational costs.

Q2. How to mitigate the interference of non-operational factors and improve the accuracy of inefficient string identification? The power generation performance of PV strings is influenced not only by operational factors (e.g., short circuits, dust accumulation, hot spots) but also by non-operational factors such as installation position, orientation, and tilt angle, as well as transient shading (e.g., clouds, birds). These factors may lead to increased false positive rates, affecting maintenance decisions. Given a PV string i at time t, its theoretical power output is denoted as $P_i^{th}(t)$ and its actual power output as $P_i^{act}(t)$. The relative inefficiency metric is defined as $D_i(t) = (P_i^{th}(t) - P_i^{act}(t))/P_i^{th}(t)$. If inefficient strings are identified solely based on a fixed threshold $D_i(t) > \tau$, transient fluctuations may be mistakenly classified as anomalies.

Q3. How to accurately differentiate various types of inefficient strings and enhance diagnostic reliability? Inefficient strings may result from multiple causes, including dust accumulation, shading, hot spots, bypass diode failures, and microcracks, with potential interdependencies among these factors. For example, dust accumulation can exacerbate hot spots, eventually leading to short circuits. Let the state vector of a PV string be X_t^i, and let the set of inefficiency categories be $C = \{c_1, c_2, ..., c_K\}$. The objective is to compute $P(C = c_k|X_t^i)$ to determine the inefficiency category of the string. However, the coupling effects among

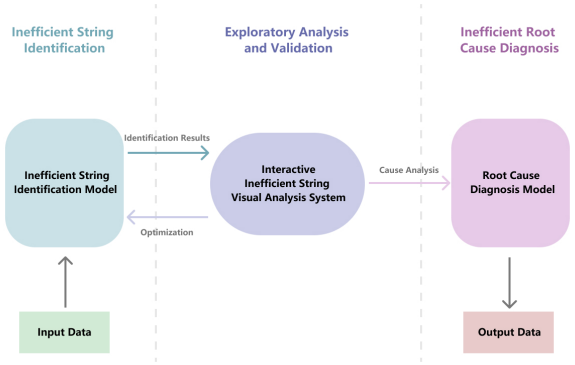

Figure 1: Workflow of inefficient PV string identification and root cause diagnosis: An interactive visual analysis system as the core, integrating upstream identification models and downstream diagnosis models

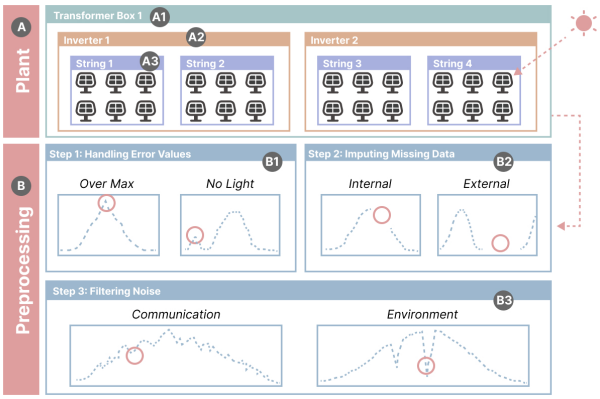

Figure 2: Pipeline for obtaining PV data set. (a) The structural components of a PV power station, including transformer boxes, inverters, and strings. (b) The preprocessing of the original data set, including Handling Error Values, Imputing Missing Data, and Filtering Noise.

different failure types make traditional classification methods less effective in precise modeling.

To address these challenges, we propose an inefficient PV string identification and root cause diagnosis workflow, as illustrated in Fig. 1. This workflow takes power station operational data as input, initially identifying inefficient strings through a detection model, followed by an interactive visual analytics system for diagnosis, determining both the inefficiency status and category of each string. Additionally, the diagnostic results contribute to the development of an intelligent and reliable root cause diagnosis model, enabling precise differentiation of complex inefficiency types. Moreover, these results serve as feedback to iteratively refine the inefficient string identification model, improving overall recognition accuracy and robustness.

3.2. Dataset

In this work, we collected real operational data from five PV stations located in different geographical areas in Zhejiang Province, China. Due to the similarity in data structures across these stations, we use one station as an example, which has an installed capacity of approximately 25 MWp. Its architectural configuration consists of string inverters, local voltage boost, and centralized grid integration. The facility utilizes over 50,000 uniform high-efficiency monocrystalline panels, organized in a hierarchical structure from modules to strings to inverters to transformer boxes. Data collection was facilitated by a dedicated PV station NCS backend monitoring system.

The dataset encompasses both time-series electrical measurements and environmental data. The electrical dataset captures readings from 8 transformer boxes (Fig. 2 a1), 75 inverters (Fig. 2 a2), and 1350 strings (Fig. 2 a3), recorded every minute from December 9, 2022, to March 26, 2023. It includes parameters such as DC side *current* and *voltage*, AC side three-phase *current* and *voltage*, active and reactive *power*, *IGBT temperature*, and daily and cumulative *power generation*. The PV power station follows a top-down hierarchical structure from transformer boxes to inverters to PV strings. Accordingly, each string's id follows the format BT[*number*]-I[*number*]-PV[*number*], representing the serial number at each level. The environmental data, collected at the same one-minute intervals, include *global irradiance*, *ambient temperature*, and *wind speed*, which are critical to assessing the performance of the PV strings.

The initial dataset contained numerous anomalies caused by equipment malfunctions and communication breakdowns, necessitating the design of an appropriate data preprocessing workflow (Fig. 2). This involved 1) **handling error values** due to sampling equipment defects, which resulted in monitoring data that violated real conditions, such as current values under no-light conditions, exceeding rated current, or negative values (Fig. 2 b1); 2) **imputing missing data** caused by communication failures, which led to extensive gaps in the dataset and severely affected the training and validation of subsequent models (Fig. 2 b2), addressed through linear or nonlinear interpolation methods; and 3) **filtering noise** introduced by unstable communications or factors like cloud cover, affecting the time-series data (Fig. 2 b3), mitigated by using a moving average method (Chou, 1969) to smooth the data and reduce interference.

3.3. Requirement analysis

To facilitate the development of a visual analytics system, we engaged in close collaboration with three experts (E1, E2, and E3) in the field of PV. Experts E1 and E2 are senior researchers with over a decade of experience in energy research, while E3 has been involved in PV research for more than five years. Through regular meetings with these experts and on-site field research at PV power stations, we identified the main operational challenge as the inefficiency of PV strings. Previous studies have applied machine learning or deep learning techniques to detect anomalies in PV strings, identifying faulty strings for maintenance personnel to repair. However, most methods (Kellil et al., 2023; Korkmaz

and Acikgoz, 2022; Fonseca Alves et al., 2021) rely on high-quality annotated data, which is costly to obtain in practical scenarios. Inherent factors (e.g., equipment defects and installation angles) and transient disturbances (like cloud cover) can lead to severe false positives in the models (Zhao et al., 2019), increasing processing costs. Additionally, the inherent black-box nature of these model methods (Castelvecchi, 2016) results in outputs lacking interpretability, making it difficult for users to understand the impact of different input features and derive effective analytical conclusions. The experts emphasized that existing anomaly detection methods still fail to balance usability, accuracy, and interpretability. Combining PV string anomaly detection with visualization is crucial for effectively identifying inefficient strings and enhancing model interpretability. Through prolonged and in-depth exchanges with the three experts, we summarized the following list of requirements for the visual analytics system:

R1. Display of anomalous distribution across all PV strings. The anomaly detection model outputs the probability of anomalies for all input strings. Textual or tabular presentations of this data neglect the hierarchical structure of the strings within the station, making it less intuitive for users to understand the overall condition of the station. A fundamental feature of the proposed system is a hierarchical display of anomalies across the transformer box, inverter, and string levels.

R2. Enhanced interpretability of the anomaly detection model. Models built on machine learning or deep learning algorithms inherently possess a "black box" nature, which hinders users from interpreting the relationship between inputs and outputs. The system must provide intuitive visualizations that map these relationships, aiding users in analyzing and verifying patterns within the data and algorithms.

R3. Guided exploration and analysis of string patterns. Faced with a vast amount of string data, the system should guide users through the analysis process. By using coding forms such as size, shape, and color, along with comparison, filtering, and recommendation operations, the system assists users in accurately analyzing the model results.

R4. Hierarchical interactive comparative analysis among different strings. PV strings have a multi-layered structure. Strings under the same inverter share similarities in aspects such as irradiance, temperature, and orientation. The system should provide hierarchical interactive methods, allowing users to compare and analyze strings from different perspectives—location, time, and environmental conditions—to unearth meaningful patterns within the string data.

R5. Facilitation of data labeling for users. During operation, PV stations generate large volumes of unlabeled time-series electrical data, and downstream tasks like anomaly detection and root cause analysis require extensive labeled data to improve accuracy. Much of this time-series data is underutilized due to the lack of labels. By integrating human perception and expert experience, the visual analytics system can rapidly produce valuable labeled data, supporting more complex downstream tasks.

4. Anomaly Detection

Driven by the actual requirements identified in our research, we designed an interactive PV string anomaly detection framework to identify inefficient strings in solar power stations. The workflow of the entire anomaly detection framework is shown in Fig. 3, starting with loading the time-series electrical and environmental data collected from the solar power station, and then performing data cleaning through a series of well-designed preprocessing operations. Next, the dimensional transformation operations and the UMAP (McInnes et al., 2018) are used to calculate the reduced-dimension pattern maps corresponding to the time-series electrical quantities of the strings. Simultaneously, the environmental data of the power station are clustered using K-Means (Macqueen, 1967) to compute cluster labels for each time point, and the clustering results are mapped back to the generated string dimensionality-reduced pattern maps. In the anomaly detection model, the dimensionality-reduced pattern maps labeled with weather tags are used as input. The maps are grouped hierarchically according to inverters and weather categories, and corresponding anomaly detection models are trained individually for each group. The results from each model are aggregated to generate the final output. Finally, the developed visual analysis system presents the model results in a visual form to the users and provides interactive operations to guide users in exploring and analyzing the results.

4.1. Dimensionality reduction pattern

Time-series electrical data of the strings is transformed into dimensionality-reduced pattern maps using algorithms, capturing potentially important feature information hidden in the data (Fig. 3 A). The process of generating dimensionality-reduced pattern maps for the strings consists of three parts. The first part involves downsampling the time-series data (Fig. 3 A1). The original time-series electrical data has a time granularity of 1 minute. To reduce computational costs without losing precision, the granularity is downsampled to 1 hour using an averaging method. Assuming the total length of the time series is n and here is one feature (mainly focusing on real-time string current), the shape of the input vector after downsampling changes from $(n, 1)$ to $(n/60, 1)$. The second step is the dimensional transformation (Fig. 3 A2). After downsampling, the number of points per day is 24, so every 24 points are transformed into one row, changing the input vector's shape $(n/60, 1)$ to $(n/1440, 24)$, where each row represents the current data of the string for one day. Finally, we use a dimensionality reduction algorithm to transform the dimensionally transformed data into a pattern map (Fig. 3 A3). Using the UMAP algorithm, the 24-dimensional vector is converted into a 2-dimensional vector, and the shape of the input vector changes from $(n/1440, 24)$ to $(n/1440, 2)$, projecting

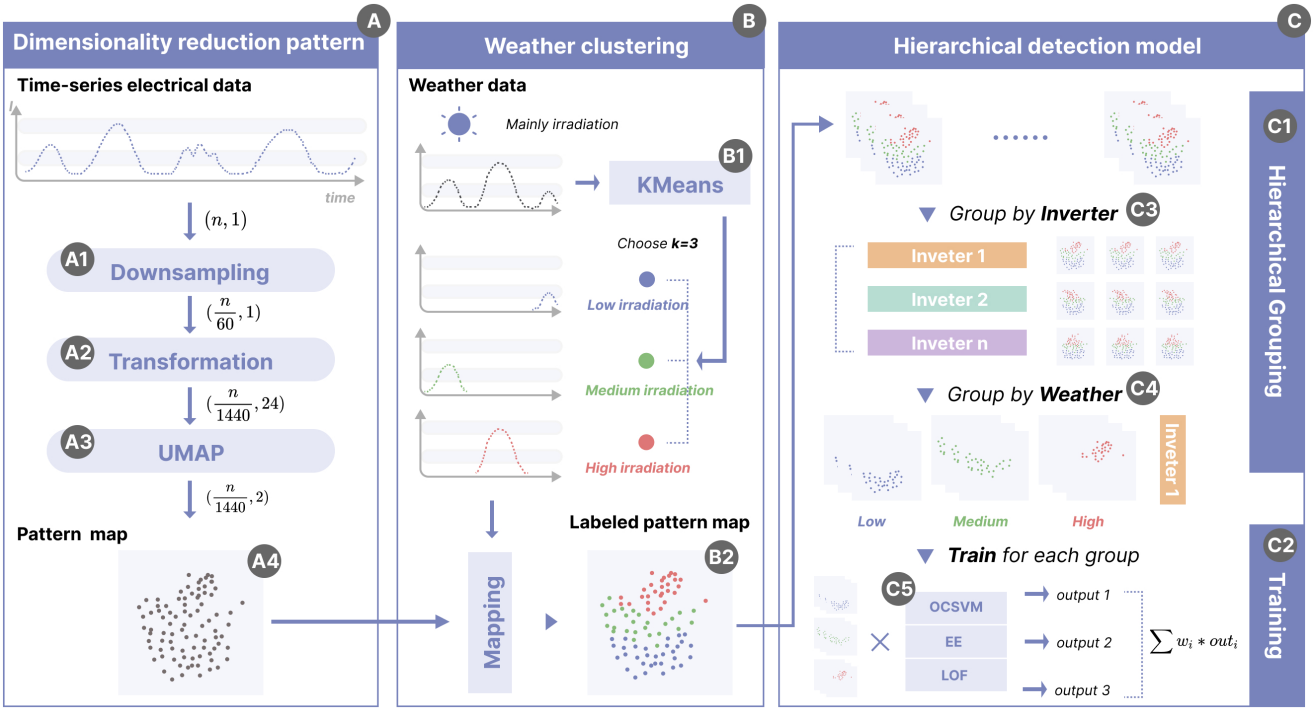

Figure 3: The PV string anomaly detection framework comprises three parts: (A) *Dimensionality Reduction Pattern*, which transforms time-series data into 2D pattern map; (B) *Weather Clustering*, which clusters weather data to enrich the maps' information; (C) *Hierarchical Detection Model*, which trains the anomaly detection model hierarchically based on power station structure and weather conditions.

it onto a 2-dimensional plane (Fig. 3 A4). Each point in the dimensionality-reduced pattern map represents the power generation condition of a string on a specific day.

The time-series electrical data is downsampled for two main purposes. 1) to reduce computational costs and accelerate the training of the subsequent anomaly detection model. 2) to filter out short-term disturbances, such as cloud cover, thereby reducing data volatility. Next, we transform the 1-dimensional feature into 24 dimensions and change the time granularity from 1 hour to 1 day, which further reduces the data size. Finally, we use the UMAP to project the high-dimensional data onto a 2-dimensional plane, facilitating the transformation from time-series data to a pattern map. Each point on the pattern map represents a string's power generation status on a specific day. The spatial distance between points reveals the similarity of the string's status over different time periods: closer distances indicate higher similarity, while farther distances indicate lower similarity. Through the spatial characteristics of the dimensionality-reduced pattern map, similar points cluster together, while outliers represent the time points of string anomalies, i.e., the inefficient modes of the strings. By translating the time-series anomalies of string electrical data into more intuitive two-dimensional spatial distance anomalies, the interpretability of the anomaly detection model results is enhanced. Furthermore, each point represents the data characteristics of a string for one day, rather than one hour or one minute, achieving a good balance between accuracy and computational complexity.

4.2. *Weather clustering*

Varying weather conditions significantly affect the power generation performance of strings. When using the spatial distance characteristics of dimensionality-reduced pattern maps to identify outliers, different irradiation conditions can interfere with the judgment of outliers. For example, strings operate better under high irradiation conditions with higher total power generation, whereas under low irradiation conditions, their performance tends to deteriorate. If there are outlier points under high irradiation conditions, it becomes difficult to distinguish these from points under low irradiation conditions in terms of spatial distance, reducing the accuracy of the model in identifying anomalous strings. Therefore, it is necessary to group and cluster the strings based on weather conditions at different times, aligning with the time dimension of individual points in the dimensionality-reduced pattern map. Using one day as the smallest unit, the K-Means is employed to cluster the time series. The main clustering factor chosen is the irradiance level, which has the most significant impact on the power generation performance of PV strings. The number of clusters is set to three for representativeness and simplicity, representing *low*, *medium*, and *high* irradiance levels.

The specific implementation process for weather clustering is shown in Fig. 3(B). The shape of input vector is $(n, 1)$, where n represents the total time length, and

the 1 corresponds to the feature of irradiance size. The collection frequency of irradiance data is also 1 minute; thus, downsampling is used to convert the time granularity to hourly, changing the vector shape to $(n/60, 1)$. This reduces the computational load of the model and mitigates data fluctuations caused by communication failures and other disturbances. The next step involves segmenting the irradiance data by day, with individual vector shapes becoming (24, 1), which serves as the subject for the K-Means. Days with similar weather conditions are grouped into the same category, resulting in three clusters representing *low*, *medium*, and *high* irradiance levels (Fig. 3 B1), color-coded blue, green, and red. The daily weather condition labels calculated are then mapped back to the strings' dimensionality-reduced pattern maps to provide the input for the anomaly detection model, as shown in Fig. 3(B2) with weather-labeled string dimensionality-reduced pattern maps.

4.3. Hierarchical detection model

When detecting faults in PV strings, the most straightforward approach is to compare their power output or current with other strings. PV strings with the same specifications should produce equal electricity under identical environmental conditions (e.g., irradiance, tilt angle) when functioning normally. However, for large-scale solar power stations, the geographical location differences between different strings can be significant, and the variations in spatial distribution lead to differences in orientation, irradiance, temperature, and other factors among the strings. As a result, strings with lower power output may still operate normally,, leading to a high rate of false positives in detection results (Zhao et al., 2019). A viable solution is to consider the hierarchical structure of PV strings. Strings connected to the same inverter are geographically close, sharing similar environmental conditions like orientation, irradiance, and temperature, making the results comparable and reducing false positives. Therefore, when constructing a model for anomaly detection of PV strings, hierarchical grouping and training of fit models are used to simulate the comparison process mentioned above, aiming to enhance the model's accuracy in identifying anomalies (Fig. 3 C1).

After obtaining the dimensionality-reduced pattern maps of the PV strings labeled with weather conditions through clustering algorithms, a hierarchical structure can also be used to train the anomaly detection model. When fitting models to the dimensionality-reduced pattern maps of all strings under a single inverter, the distribution of points in the pattern maps is significantly influenced by different weather conditions such as irradiance. The spatial distance between points under high and low irradiance conditions varies greatly, and the outlier characteristics of anomaly points can easily confuse the model's recognition ability. Therefore, by utilizing the introduced weather clustering labels to group the dimensionality-reduced pattern maps at the inverter level, a separate anomaly detection model is trained for each weather condition to eliminate data distribution biases caused by irradiance factors, further optimizing the model's performance (Fig. 3 C2).

Based on the ideas mentioned above, we propose a hierarchical anomaly detection method for PV strings based on dimensionality-reduced pattern maps of time-series electrical data, with the overall architecture illustrated in Fig. 3(C). First, the labeled dimensionality-reduced pattern maps of the PV strings serve as the input for model training. Adopting a hierarchical approach, the data is grouped according to the corresponding inverters into $Inv_1, Inv_2, \ldots, Inv_n$ where n represents the total number of inverters in the dataset, and group Inv_i represents the dimensionality-reduced pattern maps of all branch strings under inverter i (Fig. 3 C3). In the second step, all dimensionality-reduced pattern maps within each Inv_i are combined, and each point is grouped according to its corresponding weather clustering label. Each group Inv_i is further divided into $Inv_{i,low}$, $Inv_{i,med}$ and $Inv_{i,high}$, corresponding to *low*, *medium*, and *high* irradiance levels, respectively (Fig. 3 C4). Finally, after two stages of grouping, a separate anomaly detection model is trained for each $Inv_{i,label}(i = 1, 2, \ldots, n; label = low|med|high)$, as shown in Fig. 3(C5). Inspired by the concept of ensemble learning (Opitz and Maclin, 1999), we chose to combine three unsupervised models: One Class SVM, Elliptic Envelope (Rousseeuw and Driessen, 1999), and Local Outlier Factor (Breunig et al., 2000). The weighted average of these three models serves as the output of the anomaly detection model.

The structure of the composite model fully leverages the advantages of each model across different data distributions. For instance, One Class SVM does not rely on a probability distribution model of the data, making it suitable for datasets with unknown or complex distributions. The introduction of kernel tricks enables it to capture non-linear features of the data, enhancing detection accuracy. Elliptic Envelope uses the covariance matrix to comprehensively consider the interrelations between dimensions, providing a simple and efficient calculation, suitable for large datasets that follow a multivariate normal distribution. Local Outlier Factor evaluates the anomaly level of a data point by comparing the density of its local neighborhood, independent of the global data distribution, and can handle varying density patterns. In the experiments, we compared the combinations of seven models: Isolation Forest, One Class SVM, Elliptic Envelope, Local Outlier Factor, Gaussian Mixture Model (GMM), DBSCAN, and OPTICS. We calculated the identification rate of anomalous strings in their Top-K data (detailed content is provided in section 7) and ultimately selected the combination of One Class SVM, Elliptic Envelope, and Local Outlier Factor as the basic model for PV string anomaly detection due to their superior performance and computational efficiency. Using more advanced machine learning or neural network models could achieve even higher accuracy.

The output of the model consists of all outlier points from the dimensionality-reduced pattern maps after hierarchical grouping. To quantify the anomaly level of each string, the anomaly value R_a for each string is defined as eq. (1). This means that each string's anomaly value is obtained by a weighted sum of the number of its corresponding

outlier points. The larger the R_a, the higher the degree of anomaly in the string. A threshold *arr_thr* is set in the model to differentiate whether a string is anomalous. The *arr_thr* value is automatically calculated using the common 3-sigma statistical method. If a string's $R_a > arr\_thr$, the model identifies it as an anomalous string, and vice versa.

$$R_a = \frac{\frac{1}{n}\sum_{i=1}^{n} Out_{m=OCSVM|EE|LOF} - Min(\frac{1}{n}\sum_{i=1}^{n} Out)}{Max(\frac{1}{n}\sum_{i=1}^{n} Out)} \tag{1}$$

Beyond identifying inefficient PV strings in a power station, further evaluating the extent of performance degradation caused by inefficiency is crucial for making effective maintenance decisions. In practical applications, string failures do not always directly lead to performance degradation; instead, photovoltaic strings may experience gradual deterioration due to aging or environmental factors. Therefore, if the degree of string performance degradation can be scientifically quantified, maintenance personnel can prioritize addressing inefficient strings that are significantly affected by external factors, thereby improving operational efficiency. To this end, we propose a novel string degradation rate evaluation algorithm. The core idea is to first filter out normal strings using an inefficiency identification model and then fit the distribution boundary of the normal data cluster in a two-dimensional space based on a dimensionality-reduced pattern map. Subsequently, the proportion of each string's data points deviating from the normal distribution boundary in the pattern map is calculated as an approximate measure of the string performance degradation rate.

$$\begin{aligned}
&\mathcal{L}(\theta \mid x) = p(x) = MinMax(\sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)) \\
&R_d(t) = \max\left(0, 1 - \frac{\mathcal{L}_t(\theta \mid x)}{\mathcal{L}_t(\theta \mid \arg\min_x \mathcal{L}_t(\theta \mid x \in C_N))}\right) \\
&R_d = \sum_{t=1}^{n} R_d(t)
\end{aligned} \tag{2}$$

Specifically, the identification model is first used to filter out inefficient strings. For the remaining normal strings C_N, a hierarchical grouping method is applied, categorizing them based on their associated inverters and weather conditions. Instead of adopting a composite structure of multiple unsupervised models, the Gaussian Mixture Model (GMM) is employed to fit the distribution characteristics of the normal strings' dimensionality-reduced pattern maps in a two-dimensional space and to estimate the corresponding normal distribution boundary. After fitting the GMM to the normal string data, the likelihood value $\mathcal{L}(\theta)$ is computed for each data point in the dimensionality-reduced pattern maps of all photovoltaic strings. This value reflects the consistency of the data point with the normal string group. A higher likelihood value indicates that the point is closer to the normal string distribution, whereas a lower value suggests a higher degree of abnormality. Based on this property, the string degradation rate R_d is further defined, as shown in eq. (2). First, the $MinMax$ normalization method is applied to map the likelihood values of each data point in the dimensionality-reduced pattern map, computed under the fitted GMM, to the range [0, 1]. The data is then grouped based on time. Within the same time period, the minimum likelihood value $\mathcal{L}(\theta)$ among all normal strings is taken as the baseline for evaluating the deviation degree of each data point in the time group t. The deviation degree is calculated as the ratio between the likelihood value of a data point, $\mathcal{L}_t(\theta \mid x)$, and the baseline value, $\mathcal{L}_t(\theta \mid \arg\min_x \mathcal{L}_t(\theta \mid x \in C_N))$. If the likelihood value $\mathcal{L}_t(\theta \mid x)$ exceeds the baseline, it is set to 0, indicating that the string performs normally at time t. Finally, the deviation degrees $R_d(t)$ of each string over all time periods t are aggregated using a weighted sum to obtain the approximate string degradation rate R_d. This value quantitatively reflects the trend and extent of string performance degradation, providing a scientific basis for maintenance decision-making.

5. Visual Design

Based on the established design requirements, we developed the interactive visual analysis system PVeSight, as shown in Fig. 4. PVeSight consists of five collaborative views designed to facilitate intuitive and seamless user interaction, while the visual design incorporates unified color coding and spatial layout adjustments to reduce the visual burden on users. In this section, we will detail the design approach and interactive features of each view.

5.1. Global View

The Global View (Fig. 4 A) displays the recognition results of the PV string anomaly detection model on the dataset through frequency distribution histograms and a treemap. This view helps users explore and understand the operational status of the PV station (**R1, R2**).

The frequency distribution histogram (Fig. 4 A1) displays the anomaly value distribution for all strings in the station calculated by the PV string anomaly detection model. The horizontal axis shows the magnitude of anomaly values, and a red vertical line in the middle represents the anomaly discrimination threshold automatically calculated by the model using statistical methods. The slider above the histogram adjusts the Anomaly Threshold value. By normalizing, the anomaly values calculated by the anomaly detection model are mapped to the interval [0, 1]. Users can modify this by dragging left or right or by clicking the '▲' or '▼' buttons on the right side to increase or decrease the value in increments of 0.01. After users modify the Anomaly Threshold, the associated results in the Hierarchy View and Pattern View will also be regenerated.

The treemap (Fig. 4 A2) hierarchically displays the anomaly values of all strings in the station as calculated by the PV string anomaly detection model. Initially, it displays the box transformer level, where each rectangle represents a box transformer in the station. The upper left corner of the rectangle displays the corresponding box transformer

Figure 4: The PVeSight interface consists of five views. Global View (A) displays the overall abnormal conditions of the power station, guiding users to explore and detect string anomalies. Hierarchy View (B) provides a dimensionality reduction pattern map of the selected inverter along with the model's detection performance. Top View (C) shows information on all strings under the selected inverter. Pattern View (D) is used to analyze potential patterns in the time series and dimensionality reduction of the specified string. Analysis View (E) explains and annotates possible causes of model anomalies.

number, formatted as BT[*number*] (such as BT001). The upper left corner of the outermost layer represents the parent element, which is the overall PV station. On the right side is a color scale for the rectangles, ranging from left to right, representing anomaly values from 0 to 1, with deeper colors indicating higher anomaly values. Clicking on a rectangle, such as BT001, the treemap will display the next level, which is the inverter level, now with BT001 as the parent element. Each rectangle represents an inverter within the station, with the inverter number in the upper left corner, formatted as BT[*number*]-I[*number*] (such as BT001-I001). If a rectangle representing an individual inverter is clicked, the treemap will not proceed to the next level, but the content in the associated four other views will be regenerated. If users wish to return to the previous level, they should click the back arrow in the upper right corner of the outermost layer. Since the anomaly detection model calculates anomaly values for individual strings, each rectangle in the treemap represents the average anomaly value of its child elements. For an inverter, it is the average of its PV strings, and for a box transformer, it is the average of its inverters.

5.2. *Hierarchy View*

The Hierarchy View (Fig. 4 B) shows the dimensionality-reduced pattern maps and the model's recognition results for all strings under a selected inverter in the Global View, using scatter plots and contour maps. This view assists users in analyzing and interpreting the detection results of a single inverter and the model (**R2, R3**).

After selecting the inverter ID, the system retrieves the dimensionality-reduced pattern maps for all associated strings and combines them into a scatter plot for display, aiding in the analysis of string operations under that inverter. To ensure comparability and consistency, the x and y axes of the scatter plot are set to the global minimum and maximum values of the dimensionality-reduced pattern maps for all strings in the station, adjusted by ±5% of the range. This allows for the detection of anomalous patterns when switching between different inverters, aiding in user exploration and analysis. Visually, the three weather labels in the string's dimensionality-reduced pattern maps are color-coded as blue (Fig. 4 B3), green (Fig. 4 B2), and red (Fig. 4 B1) to represent low, medium, and high conditions, respectively, enhancing the informativeness of the scatter plot. This helps users distinguish the distribution of different types of points and understand the impact of weather factors on the dimensionality-reduced pattern maps. Additionally, the scatter plot includes zooming and panning functionalities, allowing users to analyze both the overall and local features of the inverter. When a user hovers the mouse over a point, the corresponding attribute information is displayed, including the string number, time, and anomaly value.

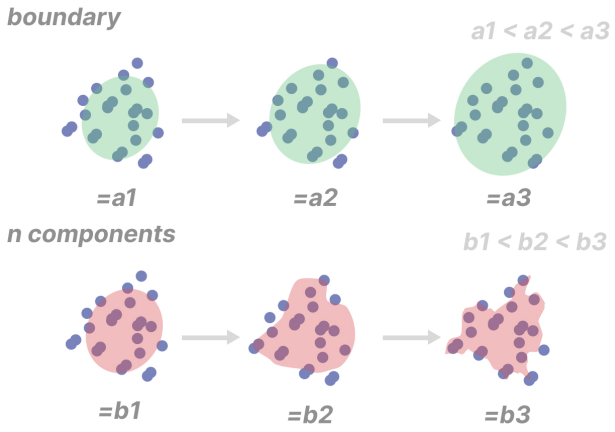

Figure 5: The impact of different *boundary* and *n components* on the distribution of the contour map.

To further enhance the expressiveness of the combined inverter's dimensionality-reduced pattern map, we used the contour map. We define the contour map as the boundary of normal point distribution obtained by fitting a GMM model trained on normal PV strings identified by the anomaly detection model. The contour line boundary values are calculated using the k-sigma method, computing the $\mu - k\sigma$ of all points' likelihood values (by default, k=3). Smaller likelihood values indicate a lower probability of belonging to the distribution. Additionally, two sliders are placed above the scatter plot. 1) *boundary* adjusts k for the contour map boundary calculation, with a range of $[1, 5]$. 2) *n_components* controls the n_components hyperparameter of the GMM model, setting the number of sub-distributions present in the GMM model. The larger the number, the more complex the fitted distribution, with a range of $[1, m]$, where m is the total number of strings under that inverter. The default value is obtained by calculation using the Bayesian Information Criterion (Schwarz, 1978). By adjusting these two parameters (Fig. 5), the system can fully leverage human perceptual advantages, making the model's fit of the normal point distribution of strings more accurate.

5.3. Top View

Top View (Fig. 4 C) displays the raw current data of all strings under the inverter selected in the Global View, as well as the output results calculated by the model. Users can filter the strings of interest based on the displayed data and metrics for analysis in the Pattern View and Analysis View (**R3, R4**).

The view is presented in a tabular format, containing four columns from left to right: *Pid*, *Distribution*, *Ra*, and *Rd*. The *Pid* column shows the string number under the inverter, formatted as BT[*number*]-I[*number*]-PV[*number*]. For example, if the inverter is BT001-I001 and the *Pid* is PV1, then the string number format is BT001-I001-PV1. The *Distribution* column shows a line chart thumbnail of the string's time-series current data, with the x-axis representing time (December 9, 2022, to March 26, 2023) and the y-axis representing current values (0 to $(1 + 5\%) \times I_{max}$), ensuring visual comparability. The *Ra* and *Rd* columns correspond to the anomaly values and degradation rates of the strings, respectively, both ranging from $[0, 1]$ and displayed as horizontal bar graphs, with longer bars indicating higher values in the corresponding column. Additionally, a sorting function has been added to the table header, allowing users to reorder the results by *Pid*, R_a, and R_d values in ascending or descending order to select strings of interest for further analysis. Users can select strings by dragging the corresponding time-series graph into the Pattern View. Top View displays data for all strings under one inverter, and when the data volume exceeds the page length, users can switch pages using the pagination feature at the bottom to view results for all strings.

5.4. Pattern View

After selecting PV strings in the Top View, users can explore the corresponding time-series current graphs and dimensionality-reduced pattern maps in the Pattern View (Fig. 4 D). By comparing horizontally and vertically, users can investigate the differences and similarities between the anomaly and normal PV strings' pattern maps and derive insights (**R2, R4**).

The Pattern View consists of line charts and scatter plots. The line chart (Fig. 4 D1) represents the time-series current data of the PV string selected through the drag-and-drop interaction in the Top View, with the x and y axes settings identical to those in the thumbnail line chart of the Top View. There are two time-series current graphs for the strings in the line chart. The string marked in red serves as the analysis string, typically selected from strings with higher *Ra* or *Rd* values in the Top View. The string marked in blue serves as the reference string, usually selected from strings with lower *Ra* and *Rd* values in the Top View. Users obtain insights into the reasons for the higher *Ra* or *Rd* values of the analysis string by comparing the current deviations over the same period with the reference string, and interpret the results reasonably with domain knowledge. Additionally, the line chart also includes horizontal zooming and dragging features, facilitating detailed analysis of specific sections by users.

The scatter plot consists of two parts. The left side represents the dimensionality-reduced pattern map of the analysis string (Fig. 4 D2), while the right side represents that of the reference string (Fig. 4 D3). Both sides display the string's identifier in the upper left corner, formatted as BT001-I001-PV1, with colors consistent with those in the legends of the line chart. The dimensionality-reduced pattern maps of the strings employ the same presentation method as the combined dimensionality-reduced pattern maps of the inverters in the Hierarchy View, namely, scatter plots with weather clustering encoded colors and the corresponding contour maps of normal point distribution. The x and y axes settings are also consistent, with the difference being that the Pattern View's pattern map only displays all points of the corresponding string, not all points under the entire inverter. Additionally, to further highlight the anomalous patterns

present in the string, points identified as anomalies by the model are coded as stars in shape and are larger than normal points.

The toolbar on the right provides a related mapping function to help interpret the model results. The first button resets the layout to its default state, while the second button is a lasso tool that, when clicked, allows users to select points within a specific area of the dimensionality-reduced pattern map. Points at the same time on both sides will be highlighted simultaneously. Hovering the mouse over an individual point displays detailed attributes of that point, such as time and anomaly value, and the corresponding time segment in the line graph will also be highlighted. This design facilitates providing users with the means to analyze and unearth potential associations between the anomalous patterns and original current data, offering a reasonable interpretation of the model recognition results.

5.5. Analysis View

The Analysis View (Fig. 4 E) is used to analyze the causes of anomalous patterns in strings and provides a clear and concise annotation interface for obtaining labeled data for downstream complex tasks such as root cause diagnosis (**R2, R5**).

In the left part of the Analysis View (Fig. 4 E1), the statistical characteristic indicators of the analysis string and the reference string are displayed. These indicators provide detailed explanations for the causes of anomalous patterns within the strings. To rationally interpret these anomalous patterns, three important metrics for result analysis have been designed and introduced: Normal Rate (R_n), Relative Power Generation Rate (R_{rpg}), and Irradiance Correlation Coefficient (C_{Irr}).

- **Normal Rate** represents the proportion of normal points in the dimensionality-reduced pattern map of the string.
- **Relative Power Generation Rate** is defined as the average ratio of the daily power generation P of a string S relative to the best-performing string S_{max} under the same inverter on the same day, as seen in eq. (3).
- **Irradiance Correlation Coefficient** is the Pearson coefficient (Pearson, 1895) between the time-series current data I and irradiance data R of a string, with a range between $[-1, 1]$. For a normal string S, the current data I is typically proportional to the irradiance data R. Negative values should be set to 0. Therefore, the definition of the irradiance correlation coefficient is given by eq. (4).

This design ensures that all metrics have a uniform range $[0, 1]$ and the same directionality; that is, the closer an indicator is to 1, the more normal the string is considered, and vice versa. This helps enhance the visual expressiveness of these metrics when mapped to visual forms.

$$R_{rpg} = \frac{1}{N}\sum_{d=1}^{N}\left(\frac{P_{s,d}}{P_{max,d}}\right) \tag{3}$$

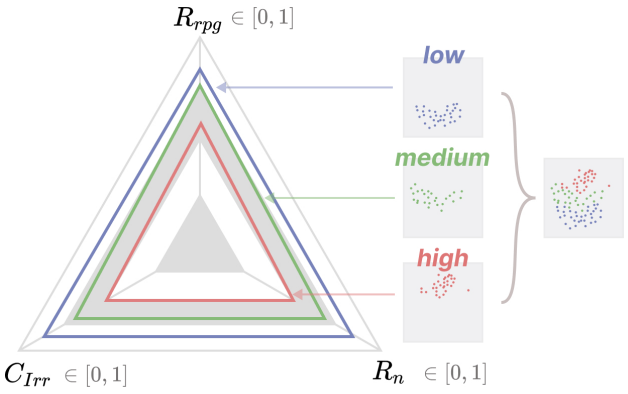

Figure 6: Radar chart visualization design for string metrics.

$$C_{Irr} = max\left(\frac{\sum_{t=1}^{N}\left(I_{s,t}-\overline{I}_s\right)\left(R_t-\overline{R}\right)}{\sqrt{\sum_{t=1}^{N}\left(I_{s,t}-\overline{I}_s\right)^2}\sqrt{\sum_{t=1}^{N}\left(R_t-\overline{R}\right)^2}}, 0\right) \tag{4}$$

To enhance the clarity and intuitiveness of the metrics, we use a radar chart for visualization, as shown in Fig. 6. Since there are three metrics involved, the background shape of the radar chart is a triangle. The space is divided into four equal parts in alternating gray and white patterns to facilitate a rough visual estimation of the indicator values. The center point represents a value of 0, while the outer vertices represent a value of 1. The smaller the triangular radar area formed by the metrics, the more anomalous the corresponding string is considered, allowing users to quickly analyze the potential causes of string anomalies visually. The radar charts are arranged in a 2×3 format, with the top row representing the analysis string and the bottom row representing the reference string from the Pattern View. Each row contains three columns, representing the metrics for low, medium, and high irradiance clusters, using the same clustering color codes as the dimensionality-reduced pattern maps in section 5.5. When users hover the mouse over the boundary lines of an area, further numeric information for each feature is displayed.

The right section of the Analysis View (Fig. 4 E2) is used for data annotation of strings, correcting any erroneous results identified by the model to facilitate subsequent iterative optimizations. The annotation includes four parts: *String ID*, *Anomaly Label*, *Reasoning*, and *Comment*. *String ID* specifies the number of the string being annotated. *Anomaly Label* indicates whether an anomaly exists in the string, with the initial value being the result identified by the model, which users can modify if the analysis results do not match the actual situation. *Reasoning* is a multiple-choice section that explores possible causes of string inefficiency based on time-series current data. A single string may have multiple causes. Based on statistical analysis of the dataset and in-depth discussions with photovoltaic experts, we propose four possible reasons for string inefficiency:

1. **Zero Current.** The string has a continuous zero current for an extended period, such as a day or a week, despite adequate irradiation.
2. **Dust Accumulation or Shading.** The string is obstructed by dust accumulation or objects like plants and shadows, causing consistently low power generation that is minimally affected by weather conditions and does not change with weather variations in extreme cases.
3. **Internal Faults.** Some components within the string have faults like short circuits, fractures, or hot spots, leading to consistently low power generation, but the power output proportionally decreases in correlation with weather changes.
4. **Dual Connection per Port.** Due to engineering reasons, some inverters have two strings connected to a single port, resulting in the string's current being generally half that of a normal string.

Lastly, *Comment* is for notes on the analysis of the string, which can be used for training downstream expert large models. After filling out the form, clicking the "Label" button completes the annotation task for a string. Finally, clicking the "Export" button in the system will obtain the annotated results of the PV string dataset.

6. Case Study

We designed a series of experiments using the PVeSight system based on real operational data from a PV power station to explore and analyze inefficient strings within the station. The aim is to uncover potential relationships between anomalous patterns and influencing factors in the data, and to discuss the causes and differences of the anomalies in detail.

6.1. Exploring the inefficiency of PV strings

After the system has loaded the power station dataset, we first analyze the Global View to obtain the anomaly detection model's results for identifying the inefficient states of strings across the entire PV power station. The distribution of string anomaly values displayed in the histogram shows that most strings have anomaly values around 0.1, while the anomaly discrimination threshold automatically calculated by the model based on statistical methods is 0.20 (Fig. 7 A1). Observing the position of the red threshold line in the histogram confirms that the discrimination threshold reasonably differentiates the strings' anomalies. Next, we explore the physical distribution of anomalous strings using a treemap. Referring to the scale where color correlates with the degree of string anomalies, darker rectangle colors indicate more anomalous box transformers. Comparisons reveal that box transformers BT002, BT004, and BT005 exhibit more significant anomalies. Clicking to enter the next level, it is found that inverters BT002-I009 and BT002-I012, inverter BT004-I004, and inverter BT005-I017 show anomalies. To further explore the current conditions of strings under severely anomalous inverters, selecting the

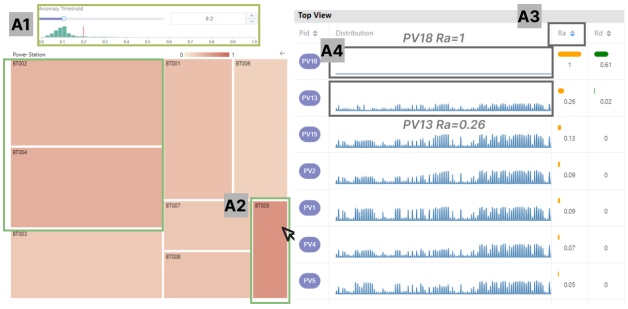

Figure 7: Exploring and identifying inefficient PV strings within the power station.

darkest-colored inverter (Fig. 7 A2), BT005-I017, displays the real-time current of all strings from PV1 to PV18 in the Top View. Sorting the strings by descending anomaly value R_a (Fig. 7 A3), it is found that two strings, PV18 and PV13, have R_a values greater than the discrimination threshold of 0.2, with respective anomaly values of 1 and 0.26. Examining the real-time current time-series graphs reveals that PV18 has zero current while PV13 shows significantly reduced overall current magnitude, clearly indicating anomalies (Fig. 7 A4). This demonstrates the model's ability to identify inefficient strings within the dataset.

6.2. Mining string dimension reduction patterns

The anomaly detection model for strings proposed in this work is based on dimensionality-reduced pattern maps of strings and implemented through a hierarchical training approach. Due to the black-box nature of such methods, the results of string identification should be made more interpretable. We selected the inverter BT005-I017 in the Global View as the object of analysis. The Hierarchy View displays the combined dimensionality-reduced pattern maps of all strings within this inverter, categorizing the points in the combined pattern map into three classes based on their color coding: high, medium, and low irradiance, along with their respective normal cluster distributions (Fig. 8 A). Observing the boundaries of these distributions, the anomaly detection model fitted for this inverter effectively segments and combines strings under the three environmental conditions in two-dimensional space. Vertically, from bottom to top, the strings are arranged as low, medium, and high irradiance, with some overlap at the boundaries of each section, particularly between high and medium irradiance where the overlap is significant (Fig. 8 A1). Analyzing the arrangement of dimensionality-reduced points, the higher the points in the space, the greater the irradiance and the better the power generation performance, meaning the height in the space is directly proportional to both the irradiance level and the string's daily performance (Fig. 8 A2). The same observations hold for different inverters. Additionally, the overlap in string BT005-I017 suggests errors at the model's weather clustering boundaries, with more overlap between high and medium irradiance. This is likely because low

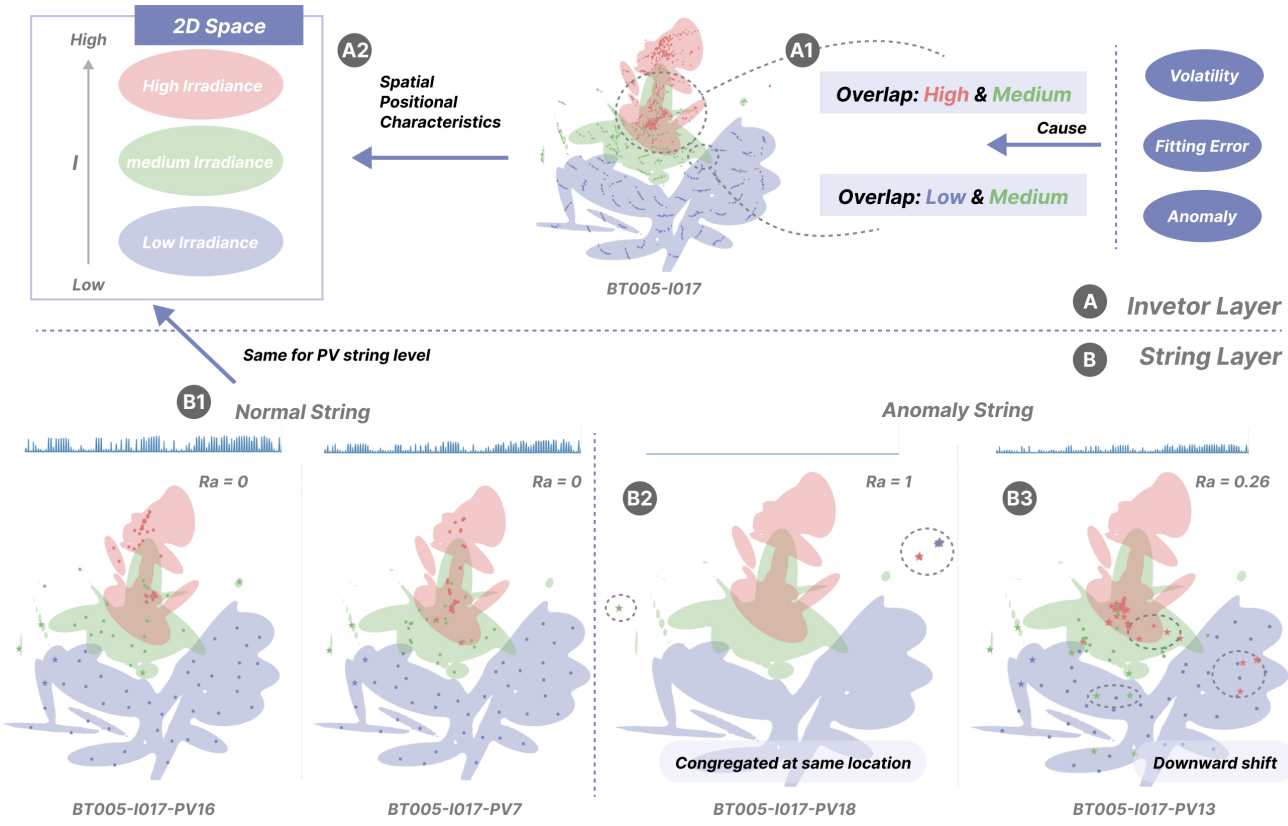

Figure 8: Analyzing the dimensionality-reduced pattern map from the inverter and string levels. (A) Investigation of the spatial distribution at the inverter level. (B) Analysis of differences between normal and abnormal string patterns, compared with the time-series graph.

irradiance fluctuates less than medium irradiance, making the distinction clearer.

We further explored and analyzed each string under the inverter in the Top View (Fig. 8 B), sorting them in descending order by the anomaly value R_a. The two strings with the lowest R_a values, PV16 ($R_a = 0$) and PV7 ($R_a = 0$), were selected as the analysis string (left side) and the reference string (right side), respectively. Our observation of the distribution of points under the three weather conditions in the dimensionality-reduced pattern map for normal PV strings also shows that the height in the dimensionality-reduced space is directly proportional to the string current (Fig. 8 B1). Star-shaped data points identified as anomalies by the model deviate significantly from their data clusters in spatial distance, indirectly validating the reliability of the anomaly detection model. Changing the analysis string to PV18 under the inverter, which has the highest R_a, the time-series current graph of this string shows nearly zero current, with $R_a = 1$ indicating that the model considers the daily current performance of the string to be anomalous every day (Fig. 8 B2). Moreover, in the dimensionality-reduced pattern map, all points are nearly gathered in the same local area, which aligns with the actual situation.

When we switched the analysis string to another anomalous string, PV13 ($R_a = 0.26$), and comparing the time-series current graphs of PV13 and PV7 in the Pattern View, we found that PV13 consistently recorded lower current than the normal string PV7 at certain time points, indicating overall lower power generation performance (Fig. 8 B3). In the dimensionality-reduced pattern map for PV13, we observed that the red section (high irradiance) and the green section (medium irradiance) shifted downward compared to the reference string on the right. For example, high irradiance points appeared in the medium range and medium irradiance points appeared in the low range. Since spatial height is directly proportional to string current, this indicates that the string's power generation performance under high and medium irradiance conditions is lower than normal, while it remains normal under low irradiance conditions. Mapping this back to the time-series current chart for verification confirmed that the results align with our conclusions.

6.3. *Comparative analysis of PV string anomalies*

Root cause analysis helps maintenance personnel verify the correctness and repair faults after identifying anomalous PV strings. Different types of anomalies require specific handling methods. For example, removing obstructions for grass or shrub blockages, cleaning solar panels for dust accumulation, and replacing panels for short or open circuits

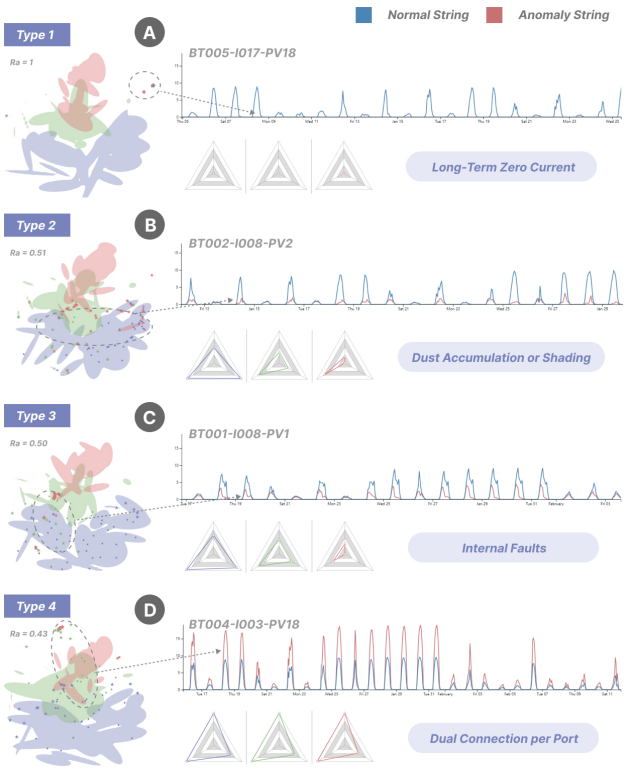

Figure 9: Analysis and annotation results of different types of anomaly strings.

require particular consideration. To cover as many anomalies as possible and verify the accuracy of the anomaly detection model, we sorted the PV strings by the global anomaly value R_a in descending order, selected strings with high anomaly values combined with time-series electrical data for detailed analysis, and made simple annotations (Fig. 9). The following are the specifics of various anomalies identified during the analysis.

Long-Term Zero Current. This anomaly occurs when the current remains zero or near-zero for several days or more, despite sufficient environmental factors such as irradiance (Fig. 9 A). In the Pattern View's dimensionality-reduced pattern map, this anomaly appears as data points clustered within a narrow range on both sides of the space. Correspondingly, the radar chart in the Analysis View shows a point for each of the three weather conditions, with all metrics at zero. When annotating, select the "Long-Term Zero Current" option.

Dust Accumulation or Shading. Analyzing this type of anomaly through the radar chart on the left side of the Analysis View, we find that these inefficient PV strings have a reduced irradiance correlation coefficient. For example, BT002-I008-PV2's coefficients are 0.94, 0.75, and 0.73 under low, medium, and high irradiance conditions (Fig. 9 B), respectively, indicating a reduced impact of irradiance on the string's real-time current due to obstacles like vegetation, buildings, or accumulation of dust and droppings. The greater the irradiance, the more pronounced the effect, but there remains a roughly proportional relationship overall. This anomaly appears in the dimensionality-reduced pattern map as a significant downward shift of points in medium and high irradiance, identified by the model as anomalies. Additionally, the relative power generation rates are 0.47, 0.33, and 0.2, indicating that the string's current is lower than normal, particularly under higher irradiance. The time-series current graphs in the Pattern View confirm an overall reduction in current magnitude, with greater deviation under high irradiance. When annotating, select the "Dust Accumulation or Shading" option.

Internal Faults. Unlike the second type, this anomaly arises from issues such as diode blowouts, surface fractures, or hot spots, causing PV strings to malfunction. This anomaly not only reduces the string's current magnitude but also weakens the correlation between current and irradiance. For BT001-I008-PV1, the irradiance correlation coefficients are 0.95, 0.73, and 0.41, decreasing as irradiance increases (Fig. 9 C). Internal defects prevent some components from working properly even under sunlight, reducing the impact of irradiance on the string's current, especially at high irradiance. Similarly, the relative power generation rates drop to 0.58, 0.45, and 0.33. In the dimensionality-reduced pattern map, points under medium and high irradiance shift downward and are marked as anomalies. In contrast to cases where time-series current values approach a uniformly low level regardless of irradiance, the current in strings with internal defects decreases proportionally, as these defects effectively reduce the number of functioning components. When annotating, select the "Internal Faults" option.

Dual Connection per Port. This anomaly is relatively special and was identified during our field research. It occurs when strings are installed in a constrained real environment, causing a single inverter port to connect two strings. PV strings with dual connections exhibit a 2:1 ratio in time-series current compared to normal strings, and their points in the dimensionality-reduced pattern map shift upwards. Additionally, the relative power generation rates in the Pattern View for dual-connected strings are close to 1, while those for normal strings are around 0.5. When annotating this data, select the "Dual Connection per Port" option.

7. Evaluation

To accurately evaluate the feasibility and usability of our proposed inefficient PV string recognition method and the visual analysis system PVeSight, we conduct validation from two aspects: model performance and system application.

7.1. Model Performance

The accuracy of the inefficient PV string recognition model is a key indicator for evaluating its effectiveness. Since our proposed method utilizes an unsupervised learning model and the dataset does not contain ground truth labels, verifying the accuracy of the model's output can only be achieved through on-site manual inspection. To address this challenge, we designed a model performance evaluation scheme based on the Top K strategy.

Table 1
Model performance evaluation results: Comparison of Top K scores for five PV power stations.

Station	Top 25	Top 50	Top 100	Top 200
A	0.85	0.80	0.77	0.72
B	0.70	0.64	0.61	0.67
C	0.75	0.72	0.68	0.65
D	0.60	0.60	0.62	0.595
E	0.65	0.60	0.51	0.555
(mean)	0.710	0.672	0.638	0.638

To effectively validate the accuracy and generalizability of the proposed inefficient PV string identification model, we collected real-time operational data from PV power stations located in five different geographic regions, including mountainous areas, plains, and tidal flats, for evaluation experiments. These power stations are labeled as A, B, C, D, and E. Following the same methodology described in section 4, we constructed dimension-reduction pattern maps using real-time electrical data from PV strings along with site environmental information. Based on these maps, independent models were trained to detect inefficient strings in each power station dataset. For this unsupervised inefficient string identification model, the experiment introduced a Top K evaluation method. The identified anomalies R_a were ranked in descending order, and the top K strings with the highest anomaly scores were selected. The actual inefficient string ratio among these selected strings was then verified through on-site inspection, serving as a quantitative measure of the model's performance.

To ensure the rationality of the experimental results, each power station provided data from 2000 PV strings for testing, and four evaluation metrics were set: Top 25, Top 50, Top 100, and Top 200. The overall results of the model performance evaluation experiment are presented in Table 1. The results indicate that the average Top K scores across all power stations decrease as K increases, with values of 0.71, 0.672, 0.638, and 0.638 for K=25, 50, 100, and 200, respectively. Although the recognition accuracy declines with increasing K, it remains consistently above 0.6, demonstrating that the proposed method maintains strong performance even in the absence of labeled data. Among the power stations, Station A achieved the highest Top K scores of 0.85, 0.80, 0.77, and 0.72, respectively, while Station E exhibited relatively poorer performance, with Top K scores of 0.65, 0.60, 0.51, and 0.555. A possible reason for this discrepancy is that Station A is located in a plain area, where external environmental factors have minimal impact, leading to higher-quality sensor data collection. In contrast, Station E is situated in an intertidal zone, where humidity and dust accumulation may degrade sensor data quality. Given these observations, incorporating more environmental data could enable more sophisticated modeling of environmental factors and improvements in the structure of the identification model, thereby further enhancing the performance of the low-efficiency string identification model.

Table 2
Experimental results of users diagnosing various types of inefficient PV strings using the PVeSight system

Anomaly Type	Proportion	Accuracy
Long-Term Zero Current	20%	100%
Dust Accumulation or Shading	12%	83.3%
Internal Faults	30%	90.0%
Dual Connection per Port	12%	91.7%
Long-Term Zero Current, Internal Faults	7%	85.7%
Dust Accumulation or Shading, Internal Faults	13%	69.2%
Dual Connection per Port, Internal Faults	6%	66.7%

7.2. System Application

In this work, we developed an interactive visual analysis system, PVeSight, for inefficient PV string detection. This system integrates visual perception with expert knowledge to diagnose the causes of inefficiency and optimize model performance. Specifically, the system first utilizes the developed inefficient PV string recognition model to identify potential inefficient PV strings. Then, through the visual analysis system, experts analyze whether the identified PV strings are indeed inefficient and determine possible causes based on domain knowledge. Finally, the corrected results and labeled data are fed back into the model to further enhance its performance.

To validate the effectiveness of the PVeSight system, we selected as many different types of inefficient PV strings as possible from the power station dataset. These included four fundamental types—long-term zero current, dust accumulation or shading, internal defects, and single-port connection of two strings—as well as three composite types formed by combining two fundamental types, resulting in a total of 100 PV strings as evaluation samples. We then invited five users to conduct an evaluation experiment using our system. During the experiment, users were asked to explore and analyze inefficient PV strings through the visual analysis system, utilizing the annotation function within the analysis view to determine whether a PV string was inefficient and to identify the potential causes of inefficiency. We recorded the average diagnostic accuracy across all users for each type of inefficiency, and the final results are presented in Table 2. The results indicate that, with the assistance of PVeSight, the recognition accuracy of different types of inefficient PV strings improved to a certain extent, demonstrating the practical value of integrating the model with an interactive system. However, there remain challenges in accurately identifying composite-type inefficient PV strings, highlighting the need for more effective interactive visualizations to better distinguish composite types from fundamental types.

In addition, we invited three experts to test our system. During the testing process, we provided each expert with a detailed introduction to all system functionalities and usage procedures. They were then asked to complete two tasks using PVeSight: free exploration and specified PV

string analysis. After the test, we conducted semi-structured interviews with each expert to collect feedback and suggestions for improvement. First, the experts affirmed the effectiveness of the proposed hierarchical anomaly detection method. By grouping data based on the three-tier structure of transformer-inverter-PV string in photovoltaic power stations, the method eliminates deviations in PV string data caused by inherent factors. Additionally, the introduction of dimensionality reduction pattern maps and irradiance clustering labels reduced computational costs while enhancing both the accuracy and interpretability of anomaly detection. From a system design perspective, the Global View, utilizing a rectangular treemap, effectively visualizes the overall status of the power station. This design enables users to quickly identify potential abnormal patterns and systematically analyze them step by step. The dimensionality reduction pattern map, with color encoding based on clustering labels, effectively reveals the influence of irradiance factors on PV string power generation performance in the reduced-dimensional space. This transformation converts temporal shape anomalies into spatial distance anomalies, leveraging human visual perception capabilities for anomaly detection. Moreover, the inclusion of fitted normal PV string contours enhances the expressiveness of the dimensionality reduction pattern map, allowing users to intuitively assess the degree of PV string degradation.

To further improve the system, experts provided the following modification suggestions: **(1)** Introduce actual geographic information in the Global View when displaying the overall situation of the station, facilitating users in discovering interesting patterns. **(2)** Enhance the accuracy and interpretability of the dimensionality-reduced pattern maps. The proposed method relies on the accuracy of the dimensionality reduction method, and the system lacks a mechanism for assessing the correctness of dimensionality reduction. **(3)** The system should integrate existing expert knowledge databases to automatically provide diagnostic results after model recognition of anomalous strings, introducing domain knowledge to enhance the system's analytical and interpretive capabilities.

8. Discussion

We propose a PV string anomaly detection system that demonstrates good performance in detection and analysis. The system features low computational costs and a straightforward operational process, effectively helping maintenance personnel identify and analyze inefficient strings. However, the system has certain limitations, which we plan to address in future work.

The system's limitations include: **(1) Insufficient diversity of the experimental dataset.**PV power stations vary across different types, such as flat, mountainous, and tidal flat stations, with each type affected differently by environmental factors. Future work will need to verify the applicability of the detection methods on different types of stations to enhance the method's general performance. **(2) Lack of automatic explanation capability.** Although the system uses dimensionality-reduced pattern maps and radar charts to interpret and analyze the time-series current results, users must still analyze and check them one by one. We can improve detection efficiency and enhance the accuracy and interpretability of user analyses by incorporating automatic diagnostic capabilities into the system using the existing expert knowledge base.

9. Conclusion

This study summarizes the domain requirements for detecting and analyzing anomalous PV strings through long-term collaboration with experts in the PV field. Based on real-world needs, we propose a hierarchical anomaly detection method and an accompanying visual analysis system, PVeSight. Dimensionality reduction algorithms are employed to transform temporal anomalies in PV string data into spatial anomalies, while environmental information is incorporated to hierarchically build the anomaly detection model. Without labeled data, good recognition results were achieved using only electrical and irradiance data, combined with a method for calculating the degradation rate of anomalous strings. We designed five views to display the model's detection results and analyze the causes of anomalies in each string, with a labeling feature to support data for downstream complex tasks. Finally, the usability of the system was evaluated through a case study and expert interviews. In future work, we will further evaluate and validate the method and system using more power station data.

CRediT authorship contribution statement

Yurun Yang: Conceptualization, Data curation, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Xinjing Yi:** Conceptualization, Data curation, Visualization, Writing – review & editing. **Yingqiang Jin:** Conceptualization, Validation, Supervision. **Sen Li:** Data curation, Validation. **Kang Ma:** Data curation, Validation, Formal analysis. **Shuhan Liu:** Conceptualization, Validation, Writing – review & editing. **Dazhen Deng:** Conceptualization, Validation, Writing – review & editing. **Di Weng:** Conceptualization, Validation, Writing – review & editing. **Yingcai Wu:** Supervision.

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

The work was supported by National Key R&D Program of China (2022YFE0137800), Key "Pioneer" R&D Projects of Zhejiang Province (2023C01120) , and NSFC (U22A2032).

References

Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32.

Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J., 2000. LOF: Identifying Density-Based Local Outliers. ACM SIGMOD Record 29, 93–104.

Cao, N., Shi, C., Lin, S., Lu, J., Lin, Y.R., Lin, C.Y., 2016. TargetVue: Visual Analysis of Anomalous User Behaviors in Online Communication Systems. IEEE Transactions on Visualization and Computer Graphics 22, 280–289.

Castelvecchi, D., 2016. Can we open the black box of AI? Nature 538, 20–23.

Chine, W., Mellit, A., Lughi, V., Malek, A., Sulligoi, G., Massi Pavan, A., 2016. A novel fault diagnosis technique for photovoltaic systems based on artificial neural networks. Renewable Energy 90, 501–512.

Chou, Y.L., 1969. Statistical Analysis. Holt, Rinehart and Winston.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learning 20, 273–297.

Drews, A., De Keizer, A., Beyer, H., Lorenz, E., Betcke, J., Van Sark, W., Heydenreich, W., Wiemken, E., Stettler, S., Toggweiler, P., Bofinger, S., Schneider, M., Heilscher, G., Heinemann, D., 2007. Monitoring and remote failure detection of grid-connected PV systems based on satellite observations. Solar Energy 81, 548–564.

Eskandari, A., Aghaei, M., Milimonfared, J., Nedaei, A., 2023. A weighted ensemble learning-based autonomous fault diagnosis method for photovoltaic systems using genetic algorithm. International Journal of Electrical Power & Energy Systems 144, 108591.

Fan, X., Peng, Y., Zhao, Y., Li, Y., Meng, D., Zhong, Z., Zhou, F., Lu, M., 2017. A personal visual analytics on smartphone usage data. Journal of Visual Languages & Computing 41, 111–120.

Fonseca Alves, R.H., de Deus Júnior, G.A., Marra, E.G., Lemos, R.P., 2021. Automatic fault classification in photovoltaic modules using convolutional neural networks. Renewable Energy 179, 502–516.

Franklin, L., Plaisant, C., Minhazur Rahman, K., Shneiderman, B., 2016. TreatmentExplorer: An Interactive Decision Aid for Medical Risk Communication and Treatment Exploration. Interacting with Computers 28, 238–252.

Gallardo-Saavedra, S., Hernández-Callejo, L., Duque-Perez, O., 2018. Technological review of the instrumentation used in aerial thermographic inspection of photovoltaic plants. Renewable and Sustainable Energy Reviews 93, 566–579.

Kellil, N., Aissat, A., Mellit, A., 2023. Fault diagnosis of photovoltaic modules using deep neural networks and infrared images under Algerian climatic conditions. Energy 263, 125902.

Kohlhammer, J., Keim, D., Pohl, M., Santucci, G., Andrienko, G., 2011. Solving Problems with Visual Analytics. Procedia Computer Science 7, 117–120.

Korkmaz, D., Acikgoz, H., 2022. An efficient fault classification method in solar photovoltaic modules using transfer learning and multi-scale convolutional neural network. Engineering Applications of Artificial Intelligence 113, 104959.

Kramer, M., 1992. Autoassociative neural networks. Computers & Chemical Engineering 16, 313–328.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2017. ImageNet classification with deep convolutional neural networks. Communications of the ACM 60, 84–90.

Liu, S., Cui, W., Wu, Y., Liu, M., 2014. A survey on information visualization: Recent advances and challenges. The Visual Computer 30, 1373–1393.

Macqueen, J., 1967. Some methods for classification and analysis of multivariate observations. Multivariate Observations 5, 281–297.

Malik, S., Shneiderman, B., Du, F., Plaisant, C., Bjarnadottir, M., 2016. High-Volume Hypothesis Testing: Systematic Exploration of Event Sequence Comparisons. ACM Transactions on Interactive Intelligent Systems 6, 1–23.

Manno, D., Cipriani, G., Ciulla, G., Di Dio, V., Guarino, S., Lo Brano, V., 2021. Deep learning strategies for automatic fault diagnosis in photovoltaic systems by thermographic images. Energy Conversion and Management 241, 114315.

McInnes, L., Healy, J., Saul, N., Großberger, L., 2018. UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software 3, 861.

Opitz, D., Maclin, R., 1999. Popular Ensemble Methods: An Empirical Study. Journal of Artificial Intelligence Research 11, 169–198.

Otamendi, U., Martinez, I., Quartulli, M., Olaizola, I.G., Viles, E., Cambarau, W., 2021. Segmentation of cell-level anomalies in electroluminescence images of photovoltaic modules. Solar Energy 220, 914–926.

Pearson, K., 1895. Note on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London 58, 240–242.

Platon, R., Martel, J., Woodruff, N., Chau, T.Y., 2015. Online Fault Detection in PV Systems. IEEE Transactions on Sustainable Energy 6, 1200–1207. doi:10.1109/TSTE.2015.2421447.

Riveiro, M., Falkman, G., Ziemke, T., 2008. Improving maritime anomaly detection and situation awareness through interactive visualization, in: 2008 11th International Conference on Information Fusion, IEEE. pp. 1–8.

Rousseeuw, P.J., Driessen, K.V., 1999. A Fast Algorithm for the Minimum Covariance Determinant Estimator. Technometrics 41, 212–223.

Roy, S., Alam, M.K., Khan, F., Johnson, J., Flicker, J., 2018. An Irradiance-Independent, Robust Ground-Fault Detection Scheme for PV Arrays Based on Spread Spectrum Time-Domain Reflectometry (SSTDR). IEEE Transactions on Power Electronics 33, 7046–7057.

Schwarz, G., 1978. Estimating the Dimension of a Model. The Annals of Statistics 6.

Silvestre, S., Chouder, A., Karatepe, E., 2013. Automatic fault detection in grid connected PV systems. Solar Energy 94, 119–127.

Simon, M., Meyer, E.L., 2010. Detection and analysis of hot-spot formation in solar cells. Solar Energy Materials and Solar Cells 94, 106–113.

Takashima, T., Yamaguchi, J., Ishida, M., 2008. Disconnection detection using earth capacitance measurement in photovoltaic module string. Progress in Photovoltaics: Research and Applications 16, 669–677.

Thom, D., Bosch, H., Koch, S., Worner, M., Ertl, T., 2012. Spatiotemporal anomaly detection through visual analysis of geolocated Twitter messages, in: 2012 IEEE Pacific Visualization Symposium, IEEE. pp. 41–48.

Tsanakas, J.A., Ha, L., Buerhop, C., 2016. Faults and infrared thermographic diagnosis in operating c-Si photovoltaic modules: A review of research and future challenges. Renewable and Sustainable Energy Reviews 62, 695–709.

Zhao, Y., Lehman, B., Ball, R., De Palma, J.F., 2013. Graph-based semi-supervised learning for fault detection and classification in solar photovoltaic arrays, in: 2013 IEEE Energy Conversion Congress and Exposition, IEEE. pp. 1628–1634.

Zhao, Y., Liu, Q., Li, D., Kang, D., Lv, Q., Shang, L., 2019. Hierarchical Anomaly Detection and Multimodal Classification in Large-Scale Photovoltaic Systems. IEEE Transactions on Sustainable Energy 10, 1351–1361.

Zhu, H., Lu, L., Yao, J., Dai, S., Hu, Y., 2018. Fault diagnosis approach for photovoltaic arrays based on unsupervised sample clustering and probabilistic neural network model. Solar Energy 176, 395–405.

Zhu, H., Wang, H., Kang, D., Zhang, L., Lu, L., Yao, J., Hu, Y., 2019. Study of joint temporal-spatial distribution of array output for large-scale photovoltaic plant and its fault diagnosis application. Solar Energy 181, 137–147.