This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSTE.2018.2867009, IEEE Transactions on Sustainable Energy

Hierarchical Anomaly Detection and Multimodal Classification in Large-Scale Photovoltaic Systems

Yingying Zhao, *Student Member, IEEE,* Qi Liu*, *Student Member, IEEE,* Dongsheng Li*, *Member, IEEE,* Dahai Kang, Qin Lv, *Member, IEEE,* and Li Shang, *Member, IEEE*

***Abstract*—Operation anomalies are common phenomena in large-scale solar farms. Effective anomaly detection and classification is essential for improving operation reliability and electricity generation. However, this is a challenging task due to the high complexity and wide variety of frequently occurring anomalies. Furthermore, existing pre-installed supervisory control and data acquisition systems (SCADA) can only provide a limited amount of information regarding the healthy condition of solar farms, making accurate anomaly detection and classification difficult. This paper presents a data-driven anomaly detection and classification solution, which can accurately detect and classify diverse photovoltaic system anomalies. The proposed solution does not require additional equipment or non-SCADA data collection. More specifically, the proposed work consists of two methods: (1) a hierarchical context-aware anomaly detection method using unsupervised learning, and (2) a multimodal anomaly classification method. The proposed solution has been deployed in two large-scale solar farms (39.36 MWp and 21.62 MWp). Multi-month operation demonstrates the effectiveness, robustness, as well as cost- and computation-efficiency of the proposed solution.**

***Index Terms*—Anomaly detection, anomaly classification, photovoltaic system, machine learning**

I. Introduction

THE installation of photovoltaic (PV) systems has experienced rapid growth over the past decade [1]. Such aggressive deployment of solar farms raises serious challenges to system operation and maintenance (O&M) [2], [3]. A large-scale PV system may consist of well over 100,000 PV panels, spanning across a wide ground surface area. Its operation is affected by various environmental effects, and anomalies are common phenomena during daily operation. For instance, surface soiling and partial shading are common concerns caused by the ambient environment. These anomalies, if not detected in a timely manner, may degrade PV system performance and further cause serious system hazards and failures [4].

Recent research has focused on developing anomaly detection and classification (ADC) methods to improve the performance and safety of PV systems [4], [5]. An effective ADC solution can help capture PV system anomalies and make it possible to schedule timely system O&M activities. Furthermore, it helps expedite PV system fault recovery and prevents further system performance deterioration. Recent studies have demonstrated that PV system performance and reliability can be effectively improved by adopting proper ADC solutions [5].

The complexity of large-scale PV systems and the diversity of system anomalies are the primary challenges for ADC. As summarized in Table I, a wide range of anomalies may occur during daily operation. The occurrence of each type of anomalies is further affected by various factors, such as seasonality, PV panel location, and PV system installation time. For instance, one of the solar farms used in this study suffers from severe grass shading in July as weeds grow fast during summer. In addition, different types of anomalies are inter-related. For instance, long-term partial shading may cause hot spots. Since different types of anomalies require different treatments, an effective ADC solution must be able to capture a wide range of anomalies. However, existing ADC methods mostly focus on tackling specific anomaly types, hence with limited application scope [4]. The primary focus of this work is to tackle the aforementioned challenges and develop a solution capable of capturing and classifying commonly occurring anomaly types.

Data collection poses another challenge to ADC solution design. Although supervisory control and data acquisition (SCADA) systems have been widely installed in solar farms to support PV system O&M, the information collected by the SCADA system is limited. More specifically, as shown in Fig. 1 [10], in a solar farm, the large number of PV panels are connected hierarchically – multiple PV panels are connected into a PV string, and multiple PV strings are connected together to a combiner box. Existing SCADA systems can only provide voltage and current information at individual PV string level and temperature information at the combiner box level. The operation status of individual PV panels is unknown. Such limited information poses restrictions to existing ADC designs. For instance, some prior work only provides combiner box level or system level anomaly detection capability [5], making it a challenge for utility operators to locate individual anomalies. Recent work tried to perform anomaly detection at PV string level. Often, additional sensing and monitoring hardware installation are needed [1], [11], [12], introducing extra installation and maintenance effort. Compared with anomaly detection, accurate classification of the diverse types of anomalies (shown in Table I) is more challenging due to the limited amount of information provided by the SCADA system.

There have been continued research developments in visual & thermal methods, given their high detection accuracy

*Corresponding author.

Y. Zhao is with Tongji University, Shanghai, P. R. China e-mail: yingying.zhao@colorado.edu.

Q. Liu, Q. Lv, and L. Shang are with University of Colorado Boulder, Boulder, CO, USA e-mail: Qi.Liu@colorado.edu.

D. Li is with Fudan University, Shanghai, P. R. China e-mail: dongshengli@fudan.edu.cn.

D. Kang is with Concord New Energy Group Limited - China, Beijing, P. R. China.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSTE.2018.2867009, IEEE Transactions on Sustainable Energy

and exact fault localization [9], [13], [14]. However, these methods are difficult to deploy in large-scale PV systems due to the following practical limitations. First, to perform fault detection in a large-scale PV system, a very large power supply is needed when implementing the orthophoto infrared thermography by light unmanned aerial vehicle and a thermal imaging system, resulting in costly and time-consuming experimental set-up [14]. Second, these methods cannot detect anomalies caused by optical degradation and failure, such as glass breakage [9]. Third, these methods are inefficient for newly-installed systems where the proportion of defective PV modules is relatively low [13].

This paper presents a data-driven approach to perform high-accuracy PV string-level anomaly detection and classification, using information solely provided by the de facto installed SCADA system. The proposed anomaly detection method consists of two stages, namely local context-aware detection (LCAD) and global context-aware anomaly detection (GCAD). LCAD aims to identify all potential anomalous PV strings with current characteristics that are distinct from adjacent PV strings under similar environmental conditions. GCAD is designed to minimize false alarms across the whole solar farm. Together, LCAD and GCAD can provide accurate string-level anomaly detection for solar farms. Furthermore, since it is difficult and expensive to obtain labeled anomaly data, the proposed anomaly detection method uses unsupervised machine learning techniques. The proposed anomaly classification method uses multimodal features. High-quality features are the first step towards efficient and accurate anomaly classification [15]. Therefore, the proposed method pays special attention to multimodal feature engineering. In our work, domain-specific features are firstly created. Then, to reduce computation complexity and improve classification performance, multimodal features are carefully designed and extracted. Next, a multimodal model training process is established, aiming to produce an accurate classification model tailored to specific classification scenarios.

The proposed ADC solution has been adopted by two large-scale solar farms with DC nominal capacity of 39.36 MW and 21.62 MW, respectively. Multi-month operation demonstrates the effectiveness, robustness, as well as cost- and computation-efficiency of the proposed solution. The contributions of this work are summarized as follows:

1) This work proposes a hierarchical context-aware anomaly detection method using unsupervised learning, which can accurately detect diverse anomalies without pre-labeling, and is more robust against irradiance and weather variations than previous methods.
2) While some anomalies are nearly undetectable under low irradiance [16] or weather with high cloud cover, the proposed method achieves 90.2% detection accuracy for the top-100 anomalies, while existing methods achieve 78.8% accuracy or lower.
3) To the best of our knowledge, this is the first work that uses SCADA data to classify commonly occurring anomalies at the PV string level in large-scale PV systems. The proposed classification method achieves 93.0% precision for the 5 types of anomalies that occur most commonly.
4) The proposed solution is cost- and computation-efficient, as it utilizes readily-available measurements in existing PV systems, without requiring additional equipment or non-SCADA data collection.

The rest of this paper is organized as follows: Section II surveys the related works. Section III presents data and solution overview. Section IV, Section V describes the proposed anomaly detection and classification methods, respectively. Section VI presents experimental results. Finally, Section VII concludes this work.

II. Related Work

A. Anomaly Detection

Recent anomaly detection approaches in PV systems can be categorized into two types: model-based approaches and data-driven approaches.

1) Model-based Approaches: Model-based methods often require a-prior (physical) model based on domain knowledge to model specific types of anomalies [17].

Platon et al. proposed an online fault detection model to estimate the AC power production, in which solar irradiance and PV module temperature measurements are used [1]. Garoudija et al. proposed a model-based fault detection method, in which temperatures and irradiance are used to detect faulty PV panels by predicting the healthy PV panels' maximum power [18]. Chouder et al. built a model to estimate the overall performance of PV systems by analyzing power loss, and detect faulty strings and partial shading anomalies [11]. Chen et al. used multiple online meters to monitor the voltage and power signals, which are then used for fault

TABLE I: Anomalies in PV Systems

Anomaly Type	Anomalies
Visual	partial shading [6] (e.g., building shading, grass shading), surface soiling [7]
Thermal	hot spot [8]
Others	sensor bias, aging [5], glass (front-cover) breakage [9]

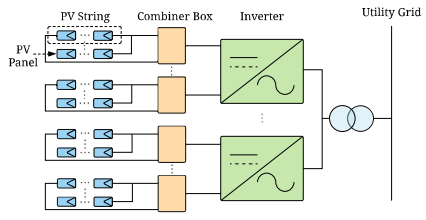


Fig. 1: Diagram of a grid-connected large-scale PV system.

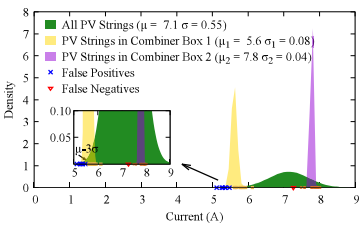


Fig. 2: Gaussian distributions of PV strings at the same timestamp for a 39.36 MWp PV system.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSTE.2018.2867009, IEEE Transactions on Sustainable Energy

detection [19]. Dhimish et al. detected faulty PV modules and strings using two metrics: power ratio and voltage ratio [20]. Some recent work performs fault detection by analyzing the PV string electrical characteristics [21]. In these methods, extra monitoring equipment besides the de facto installed SCADA system is often required for model construction. The overall system maintenance cost thus increases.

2) Data-driven Approaches: Different from model-based approaches, data-driven methods mainly rely on the information provided by SCADA systems with a limited requirement of prior domain knowledge [22].

Mekki et al. used an artificial neural network (ANN) to estimate the output photovoltaic current and voltage to detect partially shaded conditions in a PV module [23]. Similar works, described in [24], [25], used ANN to detect faulty PV modules. In these methods, a large amount of labeled data are needed to train an accurate model. Yi et al. developed a method for line-to-line fault detection based on multi-resolution signal decomposition. A two-stage support vector machine classifier is used to support decision making [26]. Other methods, such as the Bayesian Neural Network [27] and decision tree [22], were also used. Dhimish et al. presented an automatic fault detection and diagnosis solution using statistical methods. Their solution first uses voltage and power measurements to evaluate PV system performance. Then, a fault is detected by comparing the theoretical and measured performance [28]. Other statistical methods were also proposed in [29]. In summary, it is difficult and expensive to collect labeling data from real-world solar farms to build an accurate model using machine learning based methods. In addition, statistical methods suffer from high false alarm issues since they ignore the spatially variant ambient environment in large-scale PV systems.

B. Anomaly Classification

Compared with anomaly detection in PV systems, anomaly classification is under studied [4], [11], [30], [31]. There are only a few studies that tackle the classification problem.

Omran et al. presented an unsupervised learning based method to cluster similar segments of the output PV power [30]. Their method is built at system level, which provides an overall performance evaluation, but is incapable of providing the cause of an anomaly. Chouder et al. introduced an automatic supervised method to classify several types of faults in a laboratory environment [11]. The method provided the cause of faults according to the energy loss. For instance, a string defect fault causes constant energy loss, and a shading fault causes short-term energy loss. Zhao et al. proposed a supervised learning based model to detect and classify fault types in PV arrays [31]. These fault types included line-line faults, open circuit faults, and shading faults. They later proposed a semi-supervised learning based method to classify the same types of fault while reducing the demand for labeled data [4]. The proposed anomaly classification method is different from the above methods in two aspects: (1) the proposed method is capable of classifying anomalies into five types at PV string level based on SCADA systems; and (2) the design of multimodal features improves the classification performance and reduces computational efficiency.

III. Data and Solution Overview

A. PV System Configuration and Data Collection

In this work, we utilize SCADA data collected from two real-world solar farms (site A and site B) located in China. Site A has a DC nominal capacity of 39.36 MW, generated by 131,184 300 W PV panels connected to 8,199 PV strings, and 553 combiner boxes. The DC nominal capacity, 21.62 MW, of site B is generated by 72,080 300 W PV panels connected to 4,240 PV strings, and 294 combiner boxes. Measurements are recorded every minute, and individual strings' current values are used to develop the solution.

B. Data Preprocessing

To achieve accurate ADC, a well-designed data preprocessing procedure for the raw SCADA data is essential.

1) Data Cleaning: SCADA data are usually contaminated by errors and imprecise values due to malfunctions of sensors and the data management system. These errors may include unavailable values (e.g., dummy values), misfielded values [32], duplicates, or out-of-range values. These errors are first removed from modeling the data set. In addition, this study applies data from 8AM to 5PM because this period corresponds to relatively high solar irradiance, usually greater than 50 W/m^2, which corresponds with high measurement accuracy [1]. Also, observations corresponding to lower than zero or greater than short circuit current output are removed.

2) Data Filtering: The cleaned data may still contain random noises, which are handled with a median filter [33]. Since the 1 hour filtering interval has been widely adopted and its accuracy has been verified in an existing fault detection study [1], we adopted the same interval in this work. The impact of data filtering on anomaly detection accuracy is further investigated in Section VI.

3) Data Downsampling: To reduce computation cost without decreasing accuracy, downsampling is applied. In prior works, 1-minute [11], 5-minute [5], and 10-minute downsampling intervals [1] have been widely used. This study implements the proposed solution under all three downsampling intervals. The efficiency and effectiveness of these widely-used downsampling intervals are compared in Section VI.

C. Design Motivations

First, given similar irradiance, healthy PV panels should produce similar amount of power. PV strings connected to the same combiner box are closely located. Therefore, a malfunctioning PV panel/string can potentially be detected by comparing its power production against that of neighboring PV panels/strings connected to the same combiner box. However, PV strings located farther away, e.g., connected to different combiner boxes, may exhibit distinct power production profiles due to spatially variant ambient environments. As a result, direct comparison between PV strings connected to different combiner boxes may draw incorrect conclusions (e.g., high

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSTE.2018.2867009, IEEE Transactions on Sustainable Energy

false negatives and false positives). As shown in Fig. 2, all PV strings connected to combiner box No. 1 operate properly, and one faulty PV string exists in combiner box No. 2. Using direct comparison of the power production of all the PV strings, if the 3-Sigma rule is used for anomaly detection, normal strings in combiner box No. 1 will be detected as false positives, while the faulty string in combiner box No. 2 will be ignored as a false negative.

Second, the number of PV strings connected to each combiner box is limited. Due to sensor noises and environmental variations, anomaly detection based solely on local comparison with a limited number of samples may introduce a high false positive rate. To address this issue, locally detected anomaly candidates need to be further examined at the system level. Statistically, the majority of PV strings of a large-scale PV system are expected to be fault-free most of the time. Such information can be leveraged to identify true faulty PV strings and minimize false alarms. These observations motivate the proposed hierarchical context-aware anomaly detection method.

Furthermore, different types of anomalies require different maintenance treatments. Therefore, anomaly detection must be combined with classification to support maintenance activities. Anomaly classification is generally a challenging problem as the operation of a PV system is affected by a wide range of environmental variables. For instance, site A suffers from grass shading and hot spot anomalies during the summer, while site B suffers from sensor bias anomalies. Thus, how to identify the right anomaly features and design an accurate anomaly classifier is an unsolved research challenge. Fortunately, different anomalies exhibit distinct temporal and spectral characteristics, which motivated us to develop a multimodal anomaly classification method.

D. Solution Overview

Fig. 3 illustrates the flow of the proposed solution, which consists of two layers: anomaly detection and anomaly classification. There are two stages in the anomaly detection layer. First, an unsupervised machine learning technique is applied to find all possible PV strings that may be contaminated by anomalies. Then, in the second stage, those anomaly candidates are further confirmed as true anomalies at the system level. The anomaly classification layer further classifies the detected anomalous strings into multiple categories.

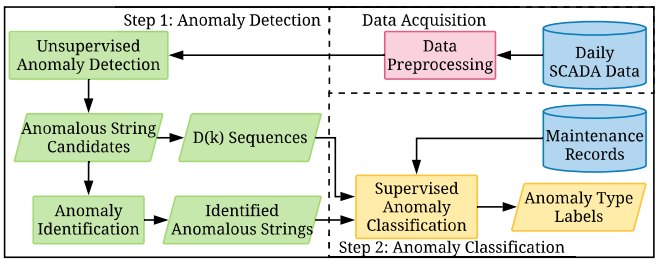


Fig. 3: Overview of the proposed ADC solution for PV systems.

IV. Anomaly Detection

This section details the proposed hierarchical context-aware anomaly detection method. The fundamental idea of the proposed method resides in its ability to learn a normal operation status for all PV strings inside a combiner box in LCAD. The anomalies are then perceived as a long-period deviation from the normal operation status in GCAD. The illustration of the proposed method is shown as Fig. 4.

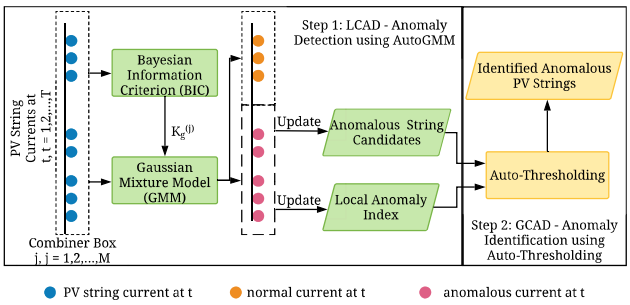


Fig. 4: Diagram of the anomaly detection process.

A. Local Context-Aware Anomaly Detection

As illustrated in Fig. 4, the goal of LCAD is to capture anomalous PV string candidates from each combiner box, leveraging the fact that PV strings in the same combiner box behave similarly except anomalous ones. To achieve this goal, an *AutoGMM* algorithm, which applies the Gaussian Mixture Model (GMM) [34] to represent the behaviors of normal and anomalous PV strings at the combiner box level is proposed with the assumption that the currents measured from normal PV strings and anomalous PV strings follow different Gaussian distributions.

Let us consider a PV system composed by s sensors collecting PV string currents and monitoring a time period n (n is the number of timestamps). A data set $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_l, \mathbf{x}_{l+1}, \ldots, \mathbf{x}_s\}$ is represented a $n \times s$ matrix, in which each column vector $\mathbf{x}_i = [\mathbf{x}_{0,i}^{(j)}, \mathbf{x}_{1,i}^{(j)}, \ldots, \mathbf{x}_{n-1,i}^{(j)}]_{1\times n}^T$ denotes the current values generated from the ith PV string in the jth combiner box. At each timestamp, a mixture of $K_g^{(j)}$ Gaussian distributions $p^{(j)} = \sum_{i=1}^{K_g^{(j)}} \phi_i N(\mu_i, \sigma_i^2)$ is used to represent PV string distributions in the jth combiner box, where $N(\mu_i, \sigma_i^2)$ is the Gaussian component to describe the distribution of currents inside the combiner box, while μ_i and σ_i^2 are the mean and variance of the ith Gaussian component, respectively. The value of $K_g^{(j)}$ is limited by the number of PV strings inside the jth combiner box. As $K_g^{(j)}$ is an unknown parameter to be estimated, this study uses the Bayesian Information Criterion (BIC) [35], [36] to determine the optimal value of $K_g^{(j)}$ automatically. The BIC value increases with the increasing of unexplained variations and the number of explanatory parameters in GMM, hence, the model with the lowest BIC is selected in this study. Eq. (1) shows the estimation of $K_g^{(j)}$.

$$\hat{K_g^{(j)}} = \arg\min_{K_g^{(j)}} m \cdot \ln(\sigma_e^2) + k \cdot \ln(m), \tag{1}$$

where m is the number of data samples, k is the number of free parameters to be estimated, and σ_e^2 is the model error variance. The expectation Maximization (EM) algorithm [37] is adopted to learn the parameters (e.g., means, variances) in GMM. In summary, Algorithm 1 describes the *AutoGMM* algorithm.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSTE.2018.2867009, IEEE Transactions on Sustainable Energy

Algorithm 1 AutoGMM($\mathbf{x}$)

Require: $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ is a set of m PV string currents in a combiner box.
1: Initialize $ModelsNum \leftarrow m$
2: **while** $ModelsNum > 0$ **do**
3: $clusters \leftarrow$ GMM($\mathbf{x}$, $ModelsNum$)
4: $BIC_{ModelsNum} \leftarrow$BIC($clusters$)
5: $ModelsNum \leftarrow ModelsNum - 1$
6: **end while**
7: $OC \leftarrow clusters$ with minimum $BIC_{ModelsNum}$
8: $NC \leftarrow$ the cluster with the maximal centroid in OC
9: $Cen \leftarrow$ the centroid of NC
10: **return** NC, Cen

The *AutoGMM* algorithm generates multiple clusters that include all PV string currents in the same combiner box at a timestamp. Then, the cluster with the maximal centroid current is identified as the normal cluster (NC), and the rest as potential abnormal clusters. This is because the normal PV string currents are greater than the anomalous ones in the same combiner box at a timestamp. To quantify the anomalous level of the ith PV string, a local anomaly index (LAI) is proposed and defined as:

$$LAI_i = \sum_{k=0}^{n-1} f(k)/n, \tag{2}$$

where $f(k)$ is defined as:

$$f(k) = \begin{cases} 1 & \text{if } \mathbf{x}_{k,i}^{(j)} \notin NC \text{ at timestamp } k. \\ 0 & \text{otherwise.} \end{cases}$$

Here, LAI represents the percentage of time that a PV string current is considered abnormal. Theoretically, the higher the LAI is, the higher possibility the PV string is abnormal. Afterwards, $LAI = \{LAI_i, \ldots, LAI_s\}$ are passed to the GCAD stage for further analysis.

B. Global Context-Aware Anomaly Detection

Due to temporal environmental conditions (e.g., cloud drift) and sensor noises, not all PV strings with positive LAIs are true anomalies. To reduce false alarms, a threshold is needed, and PV strings whose LAIs are less than this threshold will be filtered as normal PV strings. However, it is difficult to determine a proper threshold from day-to-day fault detection as the sensing conditions and external environment change over time. To address this issue, this subsection proposes a data-driven auto-thresholding algorithm.

Algorithm 2 AutoThresholding(LAI, K)

Require: A set of $LAI =< LAI_1, LAI_2, \ldots, LAI_s >$.
1: $Kclusters \leftarrow$ K-Means(K)
2: $< c'_1, c'_2, \ldots, c'_K > \leftarrow$ the ascendingly sorted centroids of the $Kclusters$
3: $thr \leftarrow 0$
4: Generating $c^* =< c^*_3, c^*_4, \ldots, c^*_K >$, with each $c^*_j \in c^*$ and $c^*_j = c'_j - 2c'_{j-1} + c'_{j-2}$
5: $thr \leftarrow$ the corresponding LAI value of the first peak in c^*
6: **return** thr

Algorithm 2 presents the auto-thresholding method. Firstly, K-Means clustering is used to partition all LAIs into K clusters. In Section VI, $K = 20$ is empirically set. Let c_i be the centroid LAI value of the ith cluster. The set of $c = \{c_1, c_2, \ldots, c_K\}$ is then sorted in ascending order into $c\prime = \{c'_1, c'_2, \ldots, c'_K\}$. Here, this study assumes that the centroid LAI values from abnormal clusters are significantly larger than those from normal clusters. To capture this significant "divergence" from the sequence c', the second order difference (SOD) of this sequence is computed as SOD mathematically describes the rate of changes. Then, the centroid LAI corresponding to the first peak in the SOD sequence is used as the threshold thr.

V. Anomaly Classification

In this section, multimodal features from both time and frequency domains are first designed and extracted. Then, a classification model is produced for specific classification scenarios.

A. Feature Extraction

As described in Section III, currents of different anomalies exhibit distinct temporal, spatial, and spectral characteristics. The characteristics, originating from the long-term deviations from normal PV string currents, provide helpful information for classifying types of anomalies. However, as discussed in Section III, the normal status of PV strings has a spatial variance, hence when deriving the deviations that characterize anomalies, spatial variance needs to be minimized. Specifically, in this study, the centroid of the normal cluster detected from the LCAD stage during the proposed anomaly detection process can be viewed as the expected current of normal PV strings. Thus, for the ith PV string in the jth combiner box, the deviation $D_i^{(j)}(k)$ as a function of discrete time k is defined in Eq. (3).

$$D_i^{(j)}(k) = Cen_k^{(j)} - \mathbf{x}_{k,i}^{(j)},\ k = 0, 1, \cdots, n-1, \tag{3}$$

where $Cen_k^{(j)}$ is the centroid of a normal cluster. It is necessary to mention that the n can be identified according to the real-time applicability. Since a daily alarm report is sufficient for O&M activities, a daily $D(k)$ sequence is used to describe the characteristics of an anomaly.

However, daily $D(k)$ sequence is a high-dimension feature vector, which is not effective and computationally-efficient for classification. To reduce computation complexity and improve classification performance, lower-dimension feature space extracted from time and frequency domain of $D(k)$ is designed and presented in the following subsection.

1) Aggregation Features: Aggregation features are extracted from a temporal perspective and defined in Eq. (4).

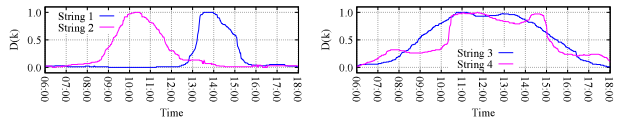
$$\mathcal{F}_a = \{Mean(D(k)), Median(D(k)), Std(D(k)), Max(D(k))\}. \tag{4}$$

As shown in Eq. (4), $\mathcal{F}_a$ consists of the mean, median, standard deviation, and maximum of the $D(k)$ sequence. The aggregation features capture the unique temporal characteristics of different anomalies and invariant characteristics of the same anomalies under spatially variant ambient environments. For instance, two anomalies of the same type $D_i^{(j)}(k)$ and $D_p^{(q)}(k)$ $(j \neq q)$ may be different as the two anomalies are located under two combiner boxes. However, statistical values such as the mean, median, standard deviation, or maximum of daily $D_i^{(j)}(k)$ and $D_p^{(q)}(k)$ sequences are similar. Fig. 5a shows such a case. For two building shading anomalies, the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSTE.2018.2867009, IEEE Transactions on Sustainable Energy

highest scaled $D(k)$ can occur either in the morning (10AM for PV string No. 2) or the afternoon (2PM for PV string No. 1), which depends on both the PV string's location and the dynamic solar incidence angle.

2) Spectrum Features: Spectrum features represent the frequency properties of a $D(k)$ sequence. The intuition behind spectrum features is that the spectral energy of daily $D(k)$ sequences may be composed of different frequency components, depending on the anomaly types.



(a) (b)

Fig. 5: Scaled $D(k)$ sequence examples for two building shading anomalies (string No. 1 and string No. 2), a hot spot anomaly (string No. 3), and a grassing shading anomaly (string No. 4).

For example, the $D(k)$ sequence of a grass shading anomaly (PV string No. 4 in Fig. 5b) may have fluctuations caused by environmental conditions, while daily $D(k)$ sequence of a hot spot anomaly (PV string No. 3 in Fig. 5b) is more stable. In this study, Fast Fourier Transform (FFT) is used to extract the spectrum features, which are defined as Eq. (5).

$$\mathcal{F}_s=\{g(u),u=0,1,\cdots,n-1\},\ \{g(u)\}_{u=0}^{n-1}. \tag{5}$$

$$g(u)=\sum_{k=0}^{n-1} D(k)e^{\frac{-j2\pi}{n}ku}. \tag{6}$$

Since the Fourier spectrum for $D(k)$ sequence is symmetric, this study only considers spectral values for $n/2$ frequencies.

B. Feature Selection

After the feature extraction, the feature dimension is reduced from n of $D(k)$ sequence to $n/2+4$ of extracted multimodal features. The dimension can be further reduced by selection, as the FFT spectrum of $D(k)$ sequence is dominated by a subset of frequency components. The remaining frequency components are of little importance for distinguishing anomaly types. To assess the importance of each feature and select the most important ones, this study first computes features' importance scores using the ranking function of XGBoost [38]. Then, the features with positive importance scores are chosen.

C. Model Training

As commonly occurring anomaly types are affected by various factors, such as specific solar farms and seasonality, the best combination of features and classification models can vary. Thus, a suitable classifier given a set of pre-defined models and features needs to be identified. This study trains three classification models, including support vector machine (SVM) [39], Bagging [40], and XGBoost based on original $D(k)$ features and extracted multimodal features, respectively. The goal of the training procedure is to seek a model and corresponding features with the highest classification performance.

VI. Experiments and Results

A. Evaluation Metrics and Experiment Setup

a) Anomaly Detection: As there is no prior knowledge about the total number of anomalies, the top-k detection accuracy defined in Eq. (7) is used to quantify the effectiveness of the proposed anomaly detection method.

$$Detection\ Accuracy=\frac{k_{correct}}{k}, \tag{7}$$

where $k_{correct}$ represents the number of true anomalies in the top-k detected anomalies. The top-k detected anomalies are the k identified anomalous PV strings with the highest LAI from a daily report. For the three baseline methods used in the following experiments, the total number of alarms for each PV string is first counted within a day. Then the k PV strings with the most frequent alarms are chosen as the top-k detected anomalies.

The proposed method is compared against three SCADA-based anomaly detection methods for PV systems [29]: Hampel identifier, 3-Sigma rule, and Boxplot outlier rule. These methods aim to find and report anomalous PV strings using instantaneous currents of all PV strings at every timestamp. To make an equal comparison setup, first, preprocessed SCADA data is used for all methods. Secondly, daily anomaly reports from the three baseline methods are generated by counting the total number of anomaly alarms for each PV string within a day and sorting their anomaly alarm numbers in descending order. Finally, the top-k detected anomalous PV strings are used to evaluate the performance of all methods.

b) Anomaly Classification: In this study, the multilabel-based macro-averaging metric defined in Eq. (8) [41] is used to quantify the overall performance of the proposed classification method.

$$B_{macro}(h)=\frac{1}{L}\sum_{j=1}^{L} B(TP_j,FP_j,TN_j,FN_j), \tag{8}$$

where $B(TP_j, FP_j, TN_j, FN_j)$ represents binary classification metrics ($B \in \{Precision, Recall, F_1\}$). L is the number of anomaly types, in this study, $L = 5$. TP_j, FP_j, TN_j, and FN_j denote the number of $true\ positive$, $false\ positive$, $true\ negative$, and $false\ negative$ test instances with respect to the j class label, respectively.

B. ADC Method Evaluation

1) Anomaly Detection Evaluation:

a) Overall Performance: Fig. 6 shows the mean detection accuracy from onsite monitoring for nearly a month. To show how the performance varies with different rankings of k for each method, k is set to vary from 10 to 100. As shown in Fig. 6, the proposed method consistently outperforms the three other methods, and the detection accuracy of the other methods decay more quickly as k increases. More specifically, the detection accuracy of the proposed method is 90.2% when k is up to 100, while it is 78.8% or lower for other methods.

In addition, the filtering algorithm can help improve the performance of the anomaly detection methods, and the proposed method (Proposed*) outperforms the unfiltered case (Proposed). That is because the filtering algorithm can partially

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSTE.2018.2867009, IEEE Transactions on Sustainable Energy

remove sensor noises and environmental variations, leading to larger current differences between normal and anomalous strings.

b) A Case Study of Anomaly Detection: The following study helps further clarify the two-stage anomaly detection method. Fig. 7 shows a combiner box containing 16 strings, and 6 of them are identified as anomalous candidates. Fig. 8 shows 20 clusters' centroid LAI values sorted in ascending order and the corresponding second order difference. The first significant peak is auto-detected as 0.21 at the 11th cluster. Therefore, only the 9th string shown in the combiner box is identified as a true anomalous string.

2) Anomaly Classification Evaluation: We collected 10-month operation data from the two PV sites, from which 1,034 anomalies were detected during this period. These anomalies are further classified into five types, summarized in Table II. The unrecoverable anomalies require repair or replacement of PV panels, while the recoverable ones can be recovered via routine maintenance, e.g., cleaning and mowing.

To evaluate the proposed classification method, the operation dataset is randomly divided into training and test sets by fixing the ratio between the training set and test set as 3:1. The results are averaged over 12 rounds of random training-test splits.

The proposed multimodal feature extraction process operates as follows. First, 541 features are extracted from each daily data sequence $D(k)$ (from 8AM to 5PM) with minute-level resolution. It then reduces the 541-dimension $D(k)$ sequence into 274 features, among which 4 of them are aggregation features, and the remaining 270 are spectrum features. Using the XGBoost method, the importance of each feature is further assessed, resulting in 254 features with positive importance score. The 254 features are then fed into classifiers.

Fig. 9 evaluates the classification performance of the proposed feature extraction method. It first evaluates the performance of the proposed $D(k)$ feature sequence. As shown in this figure, the SVM classifier achieves the best precision (92.0%) and recall (91.8%) using the proposed $D(k)$ feature sequence. Other classifiers, e.g., Bagging (BGG) and XGBoost (XGB), offer similar performance. In other words, the proposed $D(k)$ feature sequence consistently enables high-quality anomaly classification. Next, the proposed feature extraction method further reduces the 541-dimension $D(k)$ feature sequence down to the final 254 multimodal features, offering 53.1% feature dimension reduction. As shown in this figure, using the reduced 254-dimension multimodal features, among the three classifiers, the XBGoost offers the best classification precision (93.0%) and recall (92.8%). More importantly, it slightly outperforms against that of the 541-dimension $D(k)$ features. In other words, the proposed feature reduction method not only reduces classification computation complexity, but also maintains and slightly improves anomaly classification. Furthermore, using the 254-dimension multimodal features, other classifiers, i.e., SVM and Bagging (BGG) consistently offer similar classification performance.

TABLE II: Types of Anomalies Found in Two PV Systems

Property	Anomaly Source	Anomaly Type	Examples	Occurrence Frequency
unrecoverable	internal	type 1	sensor bias, aging	45.94%
	external	type 2	building shading	22.15%
	internal	type 3	hot spot, glass breakage	18.96%
recoverable	external	type 4	grass shading	12.57%
		type 5	surface soiling	0.39%

The following study aims to gain further insights of the proposed feature extraction method. Figs. 10 and 11 illustrate the top-2 components of the proposed D(k) features and the final 254 multimodal features using t-distributed stochastic neighbor embedding (t-SNE) algorithm [42], respectively. It can be seen that both feature sets provide clean separation for anomalies belonging to different types. Fig. 10 shows the proposed D(k) features contribute more in classifying type 1 and type 3 anomaly. Compared against D(k) features, type 2, type 4, and type 5 anomaly are more accurately classified using the multimodal features, as shown in Fig. 11. Figs. 12a and 12b provide further study using a confusion matrix. As can be seen, the classifier based on D(k) features misclassifies 9 testing samples of type 4 as type 3, while the classifier based on the multimodal features misclassifies 4 testing samples of type 4 as type 3 and type 5.

3) Sensitivity Analysis: The proposed method is more robust against irradiance and weather variations than previous methods. To better understand the sensitivity of our method in various scenarios, we compare our proposed method with the Hampel method, which has shown the best performance among the three methods currently available.

a) Impact of Irradiance: Fig. 13 shows the anomaly detection accuracy of our method and the Hampel method under different irradiance. We estimate the quantile regression model [43] for two quantiles ($\tau = .1$ and $\tau = .9$) to investigate the relationship between top-100 detection accuracy and irradiance. We can see that for both methods, the detection accuracy increases as irradiance increases, and the accuracy is more concentrated under higher irradiance. However, compared with the Hampel method, the proposed method achieves higher detection accuracy, and this accuracy exhibits less variation under the same irradiance. For example, when the irradiance is 4,000 Wh/m^2/day, the detection accuracy of the proposed method ranges approximately from 78% to 97%, while the detection accuracy of the Hampel method ranges from 40% to 94%. Also, the mean detection accuracy is 90.2% for the proposed method, while it is only 78.8% for the Hampel method.

b) Impact of Weather: The proposed method is also more robust to weather variations than the Hampel method. Fig. 14 shows a boxplot summarizing the distribution of top-100 anomaly detection accuracy for five different weather conditions: heavy rain, light rain, overcast, cloudy, and sunny. These conditions correspond to different amounts of cloud cover. As we can see in the figure, the proposed method achieves higher median detection accuracy (the horizontal bar within each box) and has less variation than the Hampel method under the same weather conditions, which demonstrates that the proposed method is less impacted by weather.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSTE.2018.2867009, IEEE Transactions on Sustainable Energy

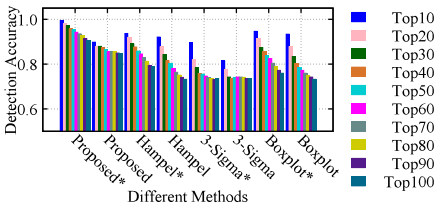


Fig. 6: Detection accuracy with top-k anomalies. $*$ indicates the use of filtered data.

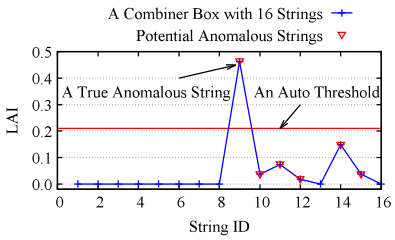


Fig. 7: A case study: LAIs for 16 strings in a combiner box.

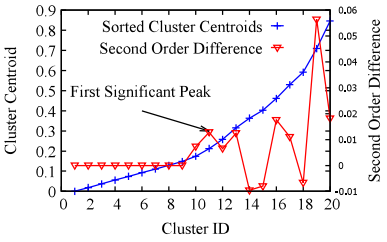


Fig. 8: An illustration of automatically identifying LAI threshold.

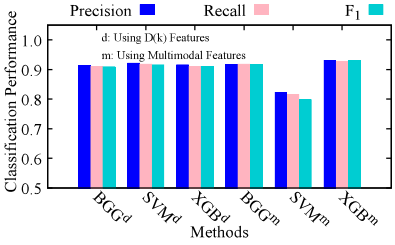


Fig. 9: Classification performance of different methods.

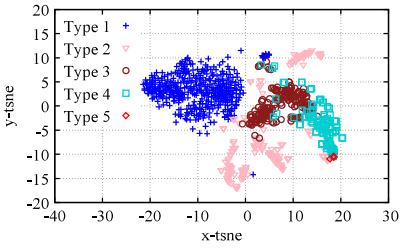


Fig. 10: Visualization of $D(k)$ features.

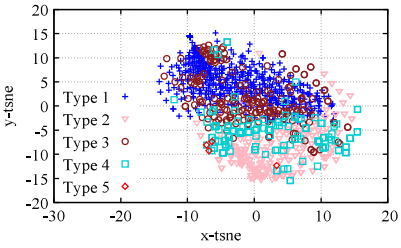
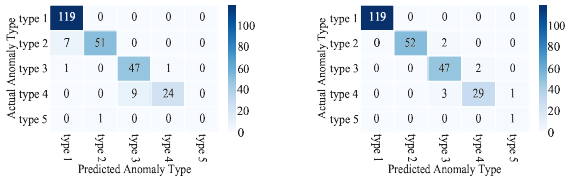


Fig. 11: Visualization of multimodal features.



(a) Using $D(k)$ Features (b) Using Multimodal Features

Fig. 12: Confusion matrix of individual anomaly types when using D(k) features vs. multimodal features.

C. Cost-benefit Analysis

Adopting the proposed solution can reduce the levelized cost of electricity (LCOE) of PV technology [44] by reducing O&M expenditures, increasing energy yield incomes, and improving soft benefits. To show the financial advantage of the proposed method, a cost-benefit analysis is performed to determine the net present value (NPV) [45], which is defined in Eq. (9).

$$NPV = \sum_{t=0}^{T} \frac{CI(t) - CO(t)}{(1+r)^t}, \tag{9}$$

where T is the expected lifetime of the PV system in years (20 years in this study), r is the discount rate (10% in this study), $CI(t)$ and $CO(t)$ are the benefits and costs in year t, respectively. Fig. 15 shows how the NPV varies with the DC nominal capacity when using the proposed solution versus conducting ADC manually. We can see that: (1) the proposed method are more financially viable than conducting ADC manually; and (2) the larger the DC nominal capacity, the greater the NPV value. Thus, it is clear that the proposed solution is financially advantageous when applied to more PV systems. The key benefits and costs are listed in Appendix A.

D. Efficiency Analysis

Computation efficiency is critical to support daily maintenance activities. The proposed ADC solution is implemented on a 2.66 GHz quad-core computer. The computation time of processing the daily collected data is measured as follows. The computation time of the LCAD stage for each sampling interval (1-minute, 5-minute, and 10-minute) is 179 min, 36 min, and 15 min, respectively. When using 10-minute downsampling interval, the computation time for site B is approximately 9 min in the LCAD stage. The computation time of GCAD is the same for all sampling intervals, 2.2 seconds. Under different sampling intervals, the proposed anomaly detection method achieves over 90% accuracy of the top-100 anomalous PV strings. To reduce computation and memory cost, the 10-minute downsampling interval is recommended. Also, the computation time in the test set for the proposed anomaly classification method with the best performance is less than 4.9 seconds (XGBoost method using multimodal features), which satisfies the real-time requirement of daily system O&M.

VII. Conclusions

This study proposes a data-driven solution for effective anomaly detection and classification, which utilizes PV string currents as indicators to detect and classify the 5 types of anomalies that occur most commonly in large-scale PV systems. Two PV systems located in China have adopted the proposed solution. Comprehensive theoretical and experimental analysis demonstrates the method is effective at detecting and classifying diverse types of faults as long as their maximum power point current I_{mpp} changes. Whether the method can detect and diagnose faults with negligible I_{mpp}, such as shunt defects, requires further investigation. We will continue to track the anomalies diagnosed from the two PV sites.

Future works include the investigation of combining I_{mpp} and maximum power point voltage V_{mpp} to detect and diagnose more diverse types of anomalies to facilitate O&M. Also, our proposed method could be complemented by visual & thermal methods given their high detection accuracy and exact fault localization. It is possible to combine aerial infrared

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSTE.2018.2867009, IEEE Transactions on Sustainable Energy

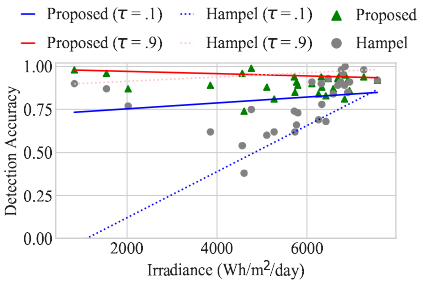


Fig. 13: Anomaly detection accuracy vs. irradiance.

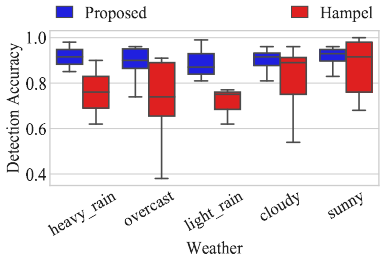


Fig. 14: Anomaly detection accuracy vs. weather.

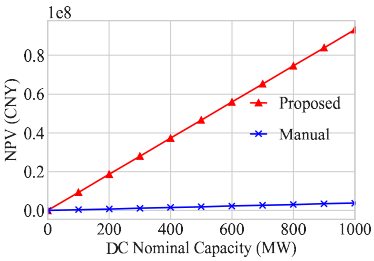


Fig. 15: NPV vs. DC nominal capacity.

thermography imaging with data-related methods for large-scale PV systems, so as to perform effective, efficient detection and diagnosis of diverse types of incipient faults.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under Grant No. 61233016, and the National Science Foundation (NSF) of United States under grant No. 1334351 and 1442971.

APPENDIX A
COMPARISON OF COSTS AND BENEFITS

Costs and Benefits	Proposed Solution (CNY/MW/year)	Manual ADC (CNY/MW/year)
O&M expenditures	3,600	6,000
software deployment costs	1,000	0
soft benefits	4,800	0
increased energy yield incomes	9,600	6,400

REFERENCES

[1] R. Platon *et al.*, "Online fault detection in PV systems," *IEEE Transactions on Sustainable Energy*, vol. 6, no. 4, pp. 1200–1207, 2015.

[2] Y. Zhao *et al.*, "Fault prediction and diagnosis of wind turbine generators using SCADA data," *Energies*, vol. 10, no. 8, pp. 1210–1210, 2017.

[3] S. Vergura *et al.*, "Descriptive and inferential statistics for supervising and monitoring the operation of PV plants," *IEEE Transactions on Industrial Electronics*, vol. 56, no. 11, pp. 4456–4464, 2009.

[4] Y. Zhao *et al.*, "Graph-based semi-supervised learning for fault detection and classification in solar photovoltaic arrays," *IEEE Transactions on Power Electronics*, vol. 30, no. 5, pp. 2848–2858, 2015.

[5] S. Firth *et al.*, "A simple model of PV system performance and its use in fault detection," *Solar Energy*, vol. 84, no. 4, pp. 624–635, 2010.

[6] J. Ahmed *et al.*, "An accurate method for MPPT to detect the partial shading occurrence in PV system," *IEEE Transactions on Industrial Informatics*, vol. PP, no. 99, pp. 1–1, 2017.

[7] B. Andò *et al.*, "Sentinella: Smart monitoring of photovoltaic systems at panel level," *IEEE Transactions on Instrumentation and Measurement*, vol. 64, no. 8, pp. 2188–2199, 2015.

[8] K. A. Kim *et al.*, "Photovoltaic hot-spot detection for solar panel substrings using AC parameter characterization," *IEEE Transactions on Power Electronics*, vol. 31, no. 2, pp. 1121–1130, 2016.

[9] J. A. Tsanakas, L. D. Ha, and F. A. Shakarchi, "Advanced inspection of photovoltaic installations by aerial triangulation and terrestrial georeferencing of thermal/visual imagery," *Renewable Energy*, vol. 102, pp. 224 – 233, 2017.

[10] Q. Liu *et al.*, "Hierarchical context-aware anomaly diagnosis in large-scale PV systems using SCADA data," *Proceedings of 15th International Conference on Industrial Informatics*, pp. 1025–1030, 2017.

[11] A. Chouder *et al.*, "Automatic supervision and fault detection of PV systems based on power losses analysis," *Energy Conversion and Management*, vol. 51, no. 10, pp. 1929–1937, 2010.

[12] I. Yahyaoui *et al.*, "A practical technique for on-line monitoring of a photovoltaic plant connected to a single-phase grid," *Energy Conversion and Management*, vol. 132, pp. 198–206, 2017.

[13] J. A. Tsanakas *et al.*, "Fault diagnosis and classification of large-scale photovoltaic plants through aerial orthophoto thermal mapping," *Proceedings of 31st European Photovoltaic Solar Energy Conference and Exhibition*, pp. 1783–1788, 2015.

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

[15] Y. Lei *et al.*, "An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data," *IEEE Transactions on Industrial Electronics*, vol. 63, no. 5, pp. 3137–3147, 2016.

[16] Z. Yi and A. H. Etemadi, "Fault detection for photovoltaic systems based on multi-resolution signal decomposition and fuzzy inference systems," *IEEE Transactions on Smart Grid*, vol. 8, no. 3, pp. 1274–1283, 2017.

[17] C. Alippi *et al.*, "Model-free fault detection and isolation in large-scale cyber-physical systems," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 1, no. 1, pp. 61–71, 2017.

[18] E. Garoudja *et al.*, "Statistical fault detection in photovoltaic systems," *Solar Energy*, vol. 150, pp. 485–499, 2017.

[19] L. Chen, S. Li, and X. Wang, "Quickest fault detection in photovoltaic systems," *IEEE Transactions on Smart Grid*, vol. 9, no. 3, pp. 1835–1847, 2018.

[20] M. Dhimish *et al.*, "Parallel fault detection algorithm for grid-connected photovoltaic plants," *Renewable Energy*, vol. 113, pp. 94–111, 2017.

[21] M. Bressan *et al.*, "A shadow fault detection method based on the standard error analysis of IV curves," *Renewable Energy*, vol. 99, pp. 1181–1190, 2016.

[22] L. Serrano-Luján *et al.*, "Case of study: Photovoltaic faults recognition method based on data mining techniques," *Journal of Renewable and Sustainable Energy*, vol. 8, no. 4, pp. 043 506–043 506, 2016.

[23] H. Mekki *et al.*, "Artificial neural network-based modelling and fault detection of partial shaded photovoltaic modules," *Simulation Modelling Practice and Theory*, vol. 67, pp. 1–13, 2016.

[24] W. Chine *et al.*, "A novel fault diagnosis technique for photovoltaic systems based on artificial neural networks," *Renewable Energy*, vol. 90, pp. 501–512, 2016.

[25] K. Jazayeri *et al.*, "Artificial neural network-based all-sky power estimation and fault detection in photovoltaic modules," *Journal of Photonics for Energy*, vol. 7, no. 2, pp. 025 501–025 501, 2017.

[26] Z. Yi and A. Etemadi, "Line-to-line fault detection for photovoltaic arrays based on multi-resolution signal decomposition and two-stage support vector machine," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 11, pp. 8546–8556, 2017.

[27] A. M. Pavan *et al.*, "A comparison between BNN and regression polynomial methods for the evaluation of the effect of soiling in large scale photovoltaic plants," *Applied Energy*, vol. 108, pp. 392–401, 2013.

[28] M. Dhimish and V. Holmes, "Fault detection algorithm for grid-connected photovoltaic plants," *Solar Energy*, vol. 137, pp. 236–245, 2016.

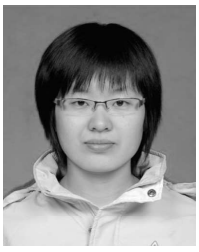
[29] Y. Zhao *et al.*, "Outlier detection rules for fault detection in solar photovoltaic arrays," *Proceeding of 28th Applied Power Electronics Conference and Exposition*, pp. 2913–2920, 2013.

[30] W. A. Omran *et al.*, "A clustering-based method for quantifying the effects of large on-grid PV systems," *IEEE Transactions on Power Delivery*, vol. 25, no. 4, pp. 2617–2625, 2010.

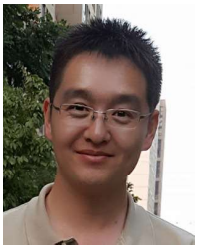
[31] Y. Zhao *et al.*, "Decision tree-based fault detection and classification in solar photovoltaic arrays," *Proceedings of 27th Applied Power Electronics Conference and Exposition*, pp. 93–99, 2012.

[32] E. Rahm and H. H. Do, "Data cleaning: Problems and current approaches," *IEEE Data Engineering Bulletin*, vol. 23, no. 4, pp. 3–13, 2000.
[33] G. R. Arce, *Nonlinear signal processing: a statistical approach*. John Wiley & Sons, 2005.
[34] S. Wang *et al.*, "A randomized response model for privacy preserving smart metering," *IEEE transactions on smart grid*, vol. 3, no. 3, pp. 1317–1324, 2012.
[35] K. P. Burnham and D. R. Anderson, "Multimodel inference: understanding AIC and BIC in model selection," *Sociological methods & research*, vol. 33, no. 2, pp. 261–304, 2004.
[36] D. Li, Q. Lv, L. Shang, and N. Gu, "Efficient privacy-preserving content recommendation for online social communities," *Neurocomputing*, vol. 219, pp. 440–454, 2017.
[37] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal processing magazine*, vol. 13, no. 6, pp. 47–60, 1996.
[38] G. Liu *et al.*, "Repeat buyer prediction for e-commerce," *Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining*, pp. 155–164, 2016.
[39] C. Cortes and V. Vapnik, "Support vector machine," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.
[40] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," *Machine learning*, vol. 40, no. 2, pp. 139–157, 2000.
[41] M. Zhang *et al.*, "A review on multi-label learning algorithms," *IEEE transactions on knowledge and data engineering*, vol. 26, no. 8, pp. 1819–1837, 2014.
[42] Y. Bengio *et al.*, "Learning deep architectures for AI," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.
[43] R. Koenker and K. F. Hallock, "Quantile regression," *Journal of economic perspectives*, vol. 15, no. 4, pp. 143–156, 2001.
[44] M. Villarini *et al.*, "Optimization of photovoltaic maintenance plan by means of a FMEA approach based on real data," *Energy Conversion and Management*, vol. 152, pp. 1 – 12, 2017.
[45] A. Nottrott, J. Kleissl, and B. Washom, "Energy dispatch schedule optimization and cost benefit analysis for grid-connected, photovoltaic-battery storage systems," *Renewable Energy*, vol. 55, pp. 230–240, 2013.

Yingying Zhao received the B.E. degree from Guilin University of Electronic Technology, Guilin, China, in 2003. She is currently pursuing the Ph.D. degree with the Department of Computer Science and Technology, Tongji University, Shanghai, China. Her current research interests include renewable and sustainable energy, recommender systems, machine learning applications, and data mining. She is a student member of IEEE.



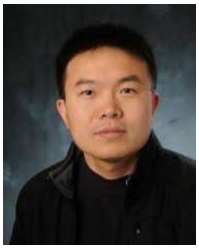
Qi Liu received the B.E. degree from Harbin Institute of Technology, China, in 2010. She is currently pursuing the Ph.D. degree with the Department of Electrical and Computer and Energy Engineering, University of Colorado, Boulder. Her current research interests include wearable computing, signal processing, and data mining. She is a student member of IEEE.



Dongsheng Li is now an adjunct associate professor with the School of Computer Science, Fudan University, Shanghai, China. Meanwhile, he is a Research Staff Member with IBM Research - China since April 2015. He obtained the Ph.D. from the School of Computer Science, Fudan University, China, in 2012. His research interests include recommender systems, machine learning applications, and data analysis in energy systems. He has won three consecutive IBM outstanding technical achievement awards from 2016 to 2018. In April 2018, he won one of the highest technical awards in IBM - the IBM Corporate Award. He is a member of IEEE.

Dahai Kang received the B.E. degree from Tsinghua University, Beijing, China, in 2000. Since 2016, he has been the chief energy Internet, Concord New Energy Group Limited - China, where he is currently working on information management of wind power and photovoltaic power plants, as well as the development of energy Internet cloud platform POWER+.

Qin Lv received the B.E. degree (Hons.) from Tsinghua University, Beijing, China, in 2000, and the Ph.D. degree in Computer Science from Princeton University, Princeton, NJ, in 2006. She is currently an Associate Professor with the Department of Computer Science, University of Colorado Boulder. Her research integrates systems, algorithms, and applications for effective and efficient data analytics in ubiquitous computing and scientific discovery. Topics of interest include mobile/wearable computing, social networks, spatial-temporal data, anomaly/misbehavior detection, recommender systems, and multi-modal data fusion. Her research is interdisciplinary in nature and interacts closely with a variety of research domains including environmental research, Earth sciences, renewable and sustainable energy, materials science, as well as the information needs in people's daily lives, such as mobile environmental sensing, indoor localization, driving behavior analysis, user profiling, and cybersafety. Lv is an associate editor of PACM IMWUT and has served on the technical program committee and organizing committee of many international conferences. Lv has received 2017 Google Faculty Research Award, VLDB 2017 Ten Year Best Paper Award, ICTAI 2017 Best Student Paper Award, two Best Paper Award nominations, and Pervasive 2012 Computational Sustainability Award. Lv has published more than 60 papers with over 5000 citations. She is a member of IEEE.



Li Shang (S'99—M'04) received the B.E. degree (Hons.) from Tsinghua University, Beijing, China, and the Ph.D. degree from Princeton University, Princeton, NJ. He is currently an Associate Professor with the Department of Electrical, Computer, and Energy Engineering, University of Colorado Boulder. He has authored or co-authored over 100 publications in computer systems, mobile computing and design for high-performance information systems. Dr. Shang served as an Associate Editor of the IEEE Transactions on Very Large Scale Integration Systems and the ACM Journal on Emerging Technologies in Computing Systems. He was a recipient of the Best Paper Award in IEEE/ACM DATE 2010 and IASTED PDCS 2002. His work on FPGA power modeling and analysis was selected as one of the 25 Best Papers from FPGA. His work on temperature-aware on-chip networks was selected for publication in the MICRO Top Picks 2006. His work was a recipient of the Best Paper Award nominations at ISLPED 2010, ICCAD 2008, DAC 2007, and ASP-DAC 2006. He was a recipient of the Provost's Faculty Achievement Award in 2010 and his department's Best Teaching Award in 2006. He was a recipient of the NSF CAREER Award. He is a member of IEEE.