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.2020.2974404, IEEE Transactions on Sustainable Energy

Collaborative Fault Detection for Large-scale Photovoltaic Systems

Yingying Zhao, *Member, IEEE,* Dongsheng Li, *Member, IEEE,* Tun Lu†, *Member, IEEE,* Qin Lv, *Member, IEEE,* Ning Gu, *Member, IEEE,* and Li Shang, *Member, IEEE*

***Abstract*—Data-driven approaches have gained increasing interests in fault detection of photovoltaic systems due to the availability of sensor data. However, the noise introduced by environmental variations and measurement variabilities pose significant challenges on effective fault detection. Furthermore, the change in electrical signal magnitude of a faulty photovoltaic component is usually small, making it difficult to distinguish an anomaly from normal ones. As such, incipient faults are nearly undetectable when they cause less loss of electricity generation. This paper proposes a collaborative fault detection solution based on collaborative filtering techniques. Specifically, the proposed solution first predicts photovoltaic strings' current values according to similar strings using historical data. Faults are then detected by long-term differences between the predicted and actual values. A key advantage of the proposed solution is its ability to capture similarities among different photovoltaic strings under noisy and spatial-temporally variant conditions, which significantly enhances fault detection performance. The proposed solution has been deployed in two large-scale solar farms (39.36 MWp and 51.04 MWp). The results show that the proposed solution is superior to existing data-driven solutions in terms of efficiency, effectiveness, and robustness.**

***Index Terms*—Data-driven, collaborative filtering, fault detection, photovoltaic, noise**

I. Introduction

RECENT years have witnessed growing deployment of photovoltaic (PV) systems in terms of number and scale [1], [2]. The scale of a PV system can be quantified by its installed capacity and determined by the number of PV panels, sensors required to monitor those panels, and the complexity of infrastructure for operating and maintaining those components. For instance, a MW-level PV system may contain thousands of 300 W PV panels, hierarchically connected through PV strings, combiner boxes, and inverters [1]. These PV components are exposed to varying weather conditions, and therefore, are prone to diverse faults. If the faults are not detected in a timely fashion, they will not only reduce system electricity production and cause secondary damage, but also accelerate system aging and even cause fire hazards [3]. Therefore, it is highly desirable to develop fault detection solutions in order to detect and locate incipient faults, thus helping operators schedule operation and maintenance (O&M) activities as well as improving the reliability of PV systems.

Fault detection techniques have been widely studied and recent works have demonstrated their effectiveness in improving the reliability of PV systems (see [4] for a comprehensive review). Researchers have found that PV systems reliability can be effectively improved and O&M cost can be greatly reduced by adopting proper fault detection solutions [1]. However, fault detection is a challenging problem in PV systems because (1) commonly-occurring faults are complex and diverse and (2) initially-installed supervisory control and data acquisition (SCADA) systems collect limited amount of information regarding the system health management [1].

Recently, both data-driven methods [1] and model-based methods [5] have been proposed to tackle fault detection in PV systems. These methods work well when detecting the most serious faults that cause significant loss of power generation. However, if an incipient fault or a specific fault (e.g., sensor bias faults) only causes small loss of power generation, it is difficult for existing methods to distinguish them from normal ones, especially under low irradiance or high cloud cover weather condition. That is because environmental variations (e.g., drifting clouds) and measure variabilities introduce noise, which can likely cause the power generation of normal PV components to deviate from normal values [1]. Meanwhile, the change in electrical signal magnitude of a faulty PV component is much lower, and the immediate impact of the fault might be minimal [6]. That minimal current magnitude, coupled with noise, renders an anomaly difficult to distinguish from normal ones. The focus of this work is to develop a noise-robust solution to detect faults in PV systems when the fault-induced power generation loss is difficult to distinguish from noise-induced power generation loss.

This paper proposes a collaborative fault detection solution to address the above challenge using the information solely collected by the SCADA system. Recent studies have shown that collaborative filtering (CF) techniques [7] can accurately predict the behaviors of a targeted user based on the behaviors of similar users of the targeted user in e-business applications. Motivated by the above idea, this paper proposes a CF-based method to predict the power generation of a targeted PV component based on the power generations of similar PV components. The key idea is that, PV components' power generation should be similar in the future if they are similar in the past no matter how temporal conditions change. That is, a faulty PV component can be detected if it does not gen-

† Corresponding author.

Y. Zhao, T. Lu, and N. Gu are with School of Computer Science, Fudan University, Shanghai, China, Shanghai Key Laboratory of Data Science, Fudan University, Shanghai, China, and Shanghai Institute of Intelligent Electronics & Systems, Shanghai, China.

D. Li is a senior researcher with Microsoft Research Asia, Shanghai, China and an adjunct professor with School of Computer Science, Fudan University, Shanghai, China.

Q. Lv and L. Shang are with University of Colorado Boulder, Boulder, CO, USA.

erate similar power as its historically similar neighbors. More specifically, the proposed solution first quantifies historical similarities among neighboring PV components' power in the case of environmental variations and measure variabilities. The proposed method pays special attention to design a problem-specific similarity measure since measuring similarity is the first step towards efficient and accurate prediction. Then, the proposed solution predicts individual PV components' power values using neighbor values and their historical similarities. After that, the accumulated residuals between the predicted values and the actual values can be used to suggest faults, since instantaneous anomaly may not mean the occurrence of a fault, while long-term underperformance indicates a fault.

The proposed solution has been adopted by two large-scale PV systems with DC nominal capacity of 39.36 MW and 51.04 MW, respectively. Multi-month operation demonstrates the effectiveness, efficiency, and robustness of the proposed solution. The main contributions of this work are summarized as follows:

1) This work introduces CF-based techniques to detect commonly-occurred faults in a robust way in PV systems. To our best knowledge, it is the first work to leverage CF techniques for fault detection applications in PV systems.
2) Field SCADA data are used to evaluate the proposed solution. Results demonstrate that the proposed solution outperforms existing methods. Furthermore, the proposed solution is more robust against irradiance and cloud cover. Specifically, at a flat terrain PV site, the proposed method achieves 96.3% detection accuracy for the top-100 faults, while existing methods achieve 90.3% accuracy or lower. At a mountainous PV site, the proposed method achieves 97.1% detection accuracy, significantly outperforming existing methods with 72.1% detection accuracy.

The rest of this paper is organized as follows: Section II formulates the targeted problem. Section III details the proposed collaborative fault detection method. Section IV presents experimental results. Section V surveys related work and differentiates our work from prior. Finally, Section VI concludes this work.

II. Problem Formulation

In this section, we first introduce the fault detection problem in large-scale PV systems. Then, we analyze the challenge caused by noisy data. After that, we motivate the targeted problem and our solution.

A. Problem Description

As mentioned in [1], a grid-connected large-scale PV system is connected hierarchically — multiple PV strings are connected into a combiner box in parallel, and multiple combiner boxes are connected to an inverter. Fault detection should be performed at the most fine-grained component level to help O&M. At that device level, unique information is required such that one device can be distinguished from others. For a large-scale PV system, the PV string level is the finest-grained level of information collected by existing SCADA systems, and a PV string's current value is the unique factor that can distinguish it from others. As such, this study leverages PV strings' current value to implement fault detection.

Let us consider a PV system composed of s sensors collecting PV strings' current values and monitoring a varying process. Each sensor i provides the measurement vector $\mathbf{X}_i = [\mathbf{x}_{0,i}^{(j)}, \mathbf{x}_{1,i}^{(j)}, \ldots, \mathbf{x}_{N-1,i}^{(j)}]_{1\times N}^T$, denoting the current values generated from PV string i in combiner box j during period N (the number of timestamps). If PV string i is affected by a fault at time t^*, the current value of the PV string is changed. Different types of faults may have different effects on current values, but usually produce lower current values compared with normal ones. Additionally, an instantaneously lower current value may not indicate a fault, while long-term underperformance does.

B. Noisy Data in PV Systems

PV strings' currents are typically noisy due to environmental variations (e.g., drifting clouds) and measurement variabilities [8]. Existing fault detection methods usually solve the noise problem by smoothing data to enlarge current differences between normal and faulty PV strings. A recent work [8] pointed out that averaged data would reduce the measurement variabilities and improve the effectiveness of fault detection methods. Inspired by this work, Zhao et al. adopted the filtering algorithm to smooth environmental variations and measurement variabilities as well, leading to improved effectiveness of their proposed hierarchal detection (HD) method [1]. They pointed out that although filtering algorithm increases the performance of detection method, noise is the foremost reason for lower detection accuracy [1], [8].

Figure 1 (left) shows the currents of 7 normal PV strings in the same combiner box on a specific day in a 39.36 MWp PV system. The 7 PV strings' currents are slightly different due to noise introduced by sensors at the same time instant, among which PV string No. 3 (dashed green line) is detected as a faulty PV string by the HD method [1]. Figure 1 (right) shows the ranking variations of PV string No. 3 under different irradiance. Here, the ranking can be viewed as fault seriousness that causes the most of power loss. We estimate the quantile regression model [9] to observe the relationship between the irradiance and rankings. The quantile τ is set as .5. We can see that the filtering algorithm decreases its rankings (with mean from 37 to 136). Also, the ranking is lower under lower irradiance. This shows that noise makes a faulty PV string difficult to distinguish from normal ones, especially under low irradiance conditions. The higher the rankings of PV strings, the higher the possibility that it will be maintained by operators. Therefore, if we can properly address the noise issue, it is very promising to improve the performance of fault detection accuracy and facility O&M.

C. Collaborative Filtering Techniques for Noise Data

As one of the most important techniques in recommender systems, collaborative filtering [7] is an algorithm that predicts a user's preferences on items that have not been seen by 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.2020.2974404, IEEE Transactions on Sustainable Energy

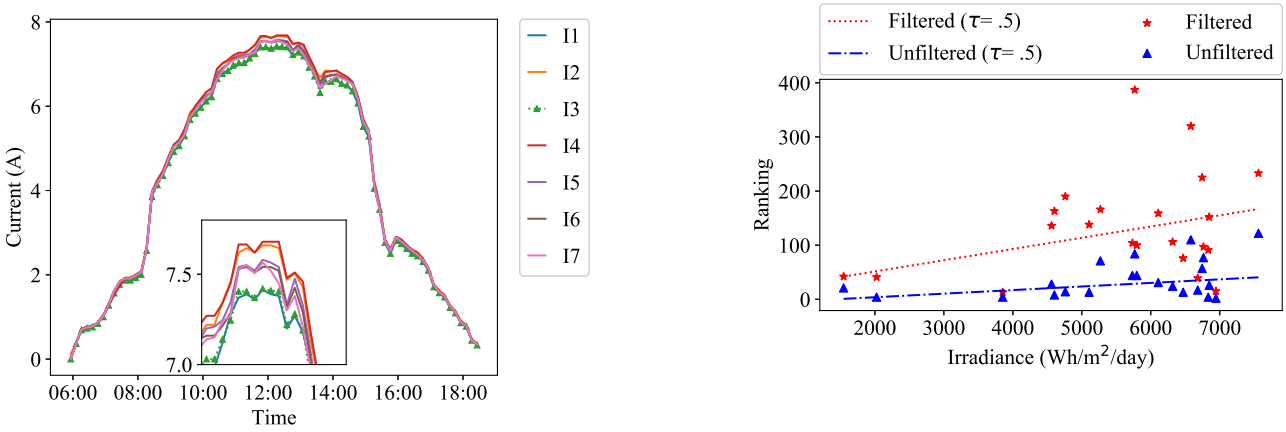


Fig. 1: Normal current values in the same combiner box on a specific day (irradiance = 6469.44 Wh/m^2/day) (Left) and anomaly ranking (Right) of PV string No. 3 during a monitoring period.

user. Among existing CF techniques, user-based collaborative filtering [10] is one of the most widely adopted methods, which generally takes two steps: (1) compute user similarities among all users and (2) compute the recommendation scores for the targeted user based on the ratings from the k neighbors with highest similarities. The basic idea behind this design is the assumption that: users who are similar in the past will also be similar in the future. Similarly, we can place the following assumption on the strings in PV systems, i.e., PV strings that have similar current values in the past will also have similar current values in the future no matter how the other conditions change. In other words, if a PV string suddenly does not generate similar current values as its neighbors, we can suggest that a fault might occur on this PV string. If such abnormal pattern continues for a period of time, e.g., several hours, we should have higher confidence to suggest a fault for this PV string.

Following the above idea, we can predict PV strings' current values according to the similarities calculated using data in the past, and the gaps between predicted and actual values can be used to suggest whether a fault occurs. However, two challenges arise here due to the differences between PV systems and traditional recommender systems. (1) Similarity measure. The PV string's current values are noisy and spatial-temporal, as such it is not clear how to accurately and effectively measure the similarity between different PV strings. (2) Fault suggestion. Based on the predicted current values and actual current values, it is also not clear how to define a gap to suggest a fault from real-time fault detection at different PV sites.

III. Proposed Method

This section first gives an overview of the proposed collaborative detection method. Then, it details the steps of the method, coupling with how it addresses the noisy data issue.

A. Method Overview

Figure 2 illustrates the flow of the proposed method, which consists of the following steps: (1) Current prediction. This step first quantifies pairwise neighbor PV strings' similarities. Then, it predicts each PV string's current values using the current values generated by its neighbors and similarities between itself and its neighbors. (2) Fault seriousness evaluation. This step evaluates each PV string's fault seriousness, which is defined as the daily accumulated residuals between the predicted and actual current values. If the residuals are induced by noises, the accumulated residuals will be around 0 by assuming a Gaussian distribution over the noise-induced residuals. If the residuals are induced by faults, the accumulated residuals will be monotonically increasing, which is clearly different from the noise-induced residuals. Theoretically, the higher the fault seriousness, the higher possibility of a PV string is faulty. (3) Fault detection. The PV strings with the highest fault seriousness are detected as faults. Here, we use an auto-thresholding method to identify faulty strings.

B. Current Prediction

This study uses the CF technique to predict each PV string's daily current values.

First, it measures the similarities among neighboring PV strings' current values. Here, neighboring PV strings are defined as other PV strings that are connected to the same combiner box. To address the noisy data issue in PV systems, we should choose an appropriate similarity measure that is robust under noises. A variety of methods can be adopted to describe similarity, in which correlation-based similarity measures, e.g., Pearson correlation and Cosine similarity, are most widely used [10]. Given two PV strings i and k connected to the same combiner box j, their Pearson correlation and Cosine similarity are defined in Equation 1 and Equation 2, respectively.

$$Corr_{i,k} = \frac{\sum_{l=0}^{n-1}(\mathbf{x}_{l,i}^{(j)} - \bar{\mathbf{x}}_i^{(j)})(\mathbf{x}_{l,k}^{(j)} - \bar{\mathbf{x}}_k^{(j)})}{\sqrt{\sum_{l=0}^{n-1}(\mathbf{x}_{l,i}^{(j)} - \bar{\mathbf{x}}_i^{(j)})^2}\sqrt{\sum_{l=0}^{n-1}(\mathbf{x}_{l,k}^{(j)} - \bar{\mathbf{x}}_k^{(j)})^2}}, \quad (1)$$

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.2020.2974404, IEEE Transactions on Sustainable Energy

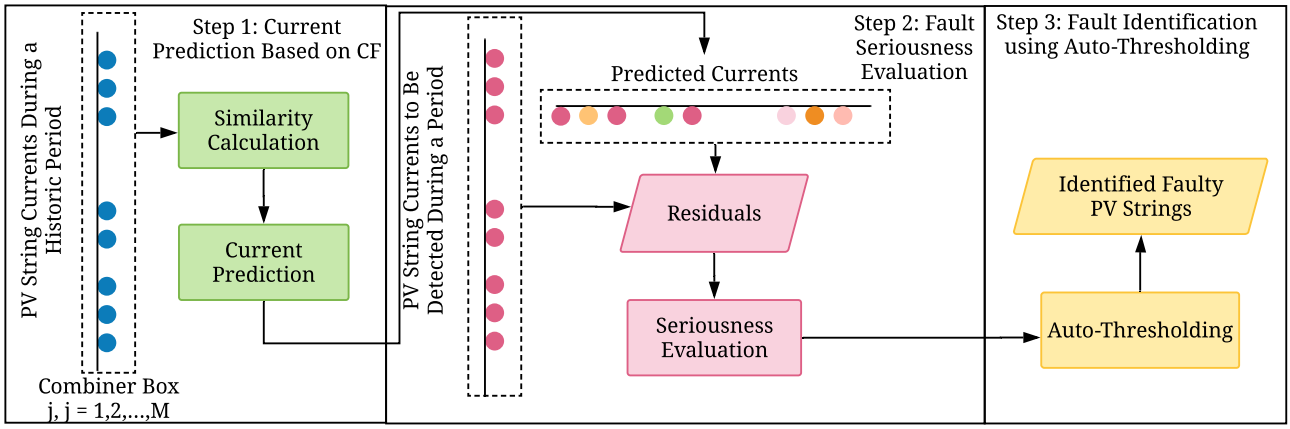


Fig. 2: Overview of the collaborative fault detection process.

$$Cosine_{i,k} = \frac{\sum_{l=0}^{n-1} \mathbf{x}_{l,i}^{(j)} \mathbf{x}_{l,k}^{(j)}}{\sqrt{\sum_{l=0}^{n-1} (\mathbf{x}_{l,i}^{(j)})^2} \sqrt{\sum_{l=0}^{n-1} (\mathbf{x}_{l,k}^{(j)})^2}}, \quad (2)$$

where n is the number of timestamps, $\bar{\mathbf{x}}_i^{(j)}$ and $\bar{\mathbf{x}}_k^{(j)}$ is the mean current value of PV string i and k in combiner box j during period n, respectively.

However, based on our investigation, Pearson correlation and Cosine similarity are not robust to noises. Given two noisy observations $\hat{\mathbf{x}}_i = \mathbf{x}_i + \epsilon_i$ and $\hat{\mathbf{x}}_k = \mathbf{x}_k + \epsilon_k$, where ϵ_i and ϵ_k are two noise vectors with zero-mean Gaussian distributions, the numerator of Equation 1 on the noisy data can be expressed as follows:

$$\sum_{l=0}^{n-1} (\hat{\mathbf{x}}_{l,i}^{(j)} - \bar{\mathbf{x}}_i^{(j)})(\hat{\mathbf{x}}_{l,k}^{(j)} - \bar{\mathbf{x}}_k^{(j)}) =$$
$$\sum_{l=0}^{n-1} (\mathbf{x}_{l,i}^{(j)} - \bar{\mathbf{x}}_i^{(j)})(\mathbf{x}_{l,k}^{(j)} - \bar{\mathbf{x}}_k^{(j)}) +$$
$$\underbrace{\sum_{l=0}^{n-1} \epsilon_{l,i}^{(j)} (\mathbf{x}_{l,k}^{(j)} - \bar{\mathbf{x}}_k^{(j)}) + \sum_{l=0}^{n-1} \epsilon_{l,k}^{(j)} (\mathbf{x}_{l,i}^{(j)} - \bar{\mathbf{x}}_i^{(j)}) + \sum_{l=0}^{n-1} \epsilon_{l,i}^{(j)} \epsilon_{l,k}^{(j)}}_{noise\ terms}. \quad (3)$$

The numerator of Equation 2 on the noisy data can be expressed as follows:

$$\sum_{l=0}^{n-1} \mathbf{x}_{l,i}^{(j)} \mathbf{x}_{l,k}^{(j)} =$$
$$\sum_{l=0}^{n-1} \mathbf{x}_{l,i}^{(j)} \mathbf{x}_{l,k}^{(j)} + \underbrace{\sum_{l=0}^{n-1} \epsilon_{l,i}^{(j)} \mathbf{x}_{l,k}^{(j)} + \sum_{l=0}^{n-1} \epsilon_{l,k}^{(j)} \mathbf{x}_{l,i}^{(j)} + \sum_{l=0}^{n-1} \epsilon_{l,i}^{(j)} \epsilon_{l,k}^{(j)}}_{noise\ terms}. \quad (4)$$

From Equation 3 and Equation 4, we can see that the noise terms are large since $\mathbf{x}_{l,i}^{(j)} - \bar{\mathbf{x}}_i^{(j)}$, $\mathbf{x}_{l,k}^{(j)} - \bar{\mathbf{x}}_k^{(j)}$, $\mathbf{x}_{l,i}^{(j)}$ and $\mathbf{x}_{l,k}^{(j)}$ are large. Therefore, we can know that the above two similarity measures are not suitable for PV systems where data noises are prevalent.

Due to the above reason, we choose to use the Euclidean distance-based similarity measure, which has been shown to be effective when the noise follows Gaussian distribution [11], [12]. The Euclidean distance can be formally described as follows:

$$Dist_{i,k} = \sqrt{\sum_{l=0}^{n-1} (\mathbf{x}_{l,i}^{(j)} - \mathbf{x}_{l,k}^{(j)})^2}. \quad (5)$$

Again, we can analyze its robustness on noisy data as follows:

$$\sum_{l=0}^{n-1} (\hat{\mathbf{x}}_{l,i}^{(j)} - \hat{\mathbf{x}}_{l,k}^{(j)})^2 = \sum_{l=0}^{n-1} (\mathbf{x}_{l,i}^{(j)} - \mathbf{x}_{l,k}^{(j)})^2 +$$
$$\underbrace{\sum_{l=0}^{n-1} 2(\epsilon_{l,i}^{(j)} - \epsilon_{l,k}^{(j)})(\mathbf{x}_{l,i}^{(j)} - \mathbf{x}_{l,k}^{(j)}) + \sum_{l=0}^{n-1} (\epsilon_{l,i}^{(j)} - \epsilon_{l,k}^{(j)})^2}_{noise\ terms}. \quad (6)$$

We can see from Equation 6 that the noise terms can be small because $\mathbf{x}_{l,i}^{(j)} - \mathbf{x}_{l,k}^{(j)}$ is usually small for similar PV strings, i.e., the Euclidean distance is more robust to noises and thus more applicable in this work. The Euclidean distance can be converted to similarity measure as follows:

$$sim_{i,k} = 1 - norm(\sqrt{\sum_{l=0}^{n-1} (\mathbf{x}_{l,i}^{(j)} - \mathbf{x}_{l,k}^{(j)})^2}), \quad (7)$$

where $norm(\cdot)$ denotes the min-max normalization function [13]. More empirical comparisons among the three similarity measures will be presented in the experiment section (Section IV).

Secondly, we predict a PV string's current value $\hat{\mathbf{x}}_i^{(j)}$ at time instant t using the CF technique, as defined in Equation (8) [14]:

$$\hat{\mathbf{x}}_{t,i}^{(j)} = \frac{\sum_{k \in C_j, k \neq i} \mathbf{x}_{t,k}^{(j)} * sim_{i,k}}{\sum_{k \in C_j, k \neq i} sim_{i,k}}, \quad (8)$$

where C_j is the set of neighbors of PV string i.

C. Fault Seriousness Evaluation

The fault seriousness of a PV string is defined as an accumulation function that takes two factors into account: (1) the residuals between the PV string's predicted and the actual current values; and (2) the duration that the residuals last. The larger the residual, the higher the possibility that a PV string is faulty. Also, a transiently high residual may not suggest a fault, while a long-lasting high residual may mean a fault. Based on such thinking, Equation (9) shows the definition of fault seriousness.

$$Fault\ Seriousness = \frac{1}{n}|\int_{t=0}^{n-1}(\hat{\mathbf{x}}_{t,i}^{(j)} - \mathbf{x}_{t,i}^{(j)})|dt, \quad (9)$$

D. Fault Detection

Theoretically, the higher the fault seriousness, the higher the possibility of the PV string being faulty. Here, we need a threshold to help identify the higher fault seriousness from daily fault detection at different PV sites.

The auto-thresholding method, proposed in [1], is capable of adjusting itself from day-to-day fault detection at different PV sites. The intuition behind the auto-thresholding method is based on the observation that most PV strings are fault-free and the total number of faulty strings is far fewer than that of normal ones. The authors in [1] used their proposed metric to measure the anomaly degree of PV strings. The metric has such characteristics: faulty PV strings' anomaly degree are significantly greater than the normal ones'. Thus, the auto-thresholding method designed for capturing a significant "divergence" from the second-order difference for daily ascending fault degrees of PV strings.

This work follows the idea. Fault seriousness measures the daily-accumulated difference between the measured current value and the predicted current value (under normal operation) of each PV string. Since the majority of PV strings are fault free with similar fault seriousness value (close to zero), we can then detect the faulty ones as their fault seriousness values deviate significantly from the majority of fault-free ones.

Since the auto-thresholding method is not the key contribution of this work, we refer readers to [1] for details.

IV. Experiments and Results

A. Experimental Setup

Data Description. This study makes use of SCADA data collected from two real-word solar farms (site A and site B) located in China. Site A is located in a flat terrain. It 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. Site B is located in a mountainous area. The DC nominal capacity, 51.04 MW, of site B is generated by 170,118 300 W PV panels connected to 9,451 PV strings, and 790 combiner boxes. Measurements are recorded every minute.

Compared Method. The proposed method is compared against the HD method [1], which has shown the best performance among the data-driven methods currently available in the literature. This work also compares the proposed Euclidean-based method with the Cosine-based and the Pearson-based method.

Data Preprocessing. Before fault detection, data preprocessing is used in all methods. The details of data preprocessing can be referred to [1]. Since the fault detection method with a 10-minute interval has been widely adopted and recommended [1], [8], we implement the fault detection under a 10-minute downsampling interval.

Evaluation Metric. Both methods are scored in terms of top-K detection accuracy [1], which is defined as follows:

$$Detection\ Accuracy = \frac{K_{correct}}{K}, \quad (10)$$

where $K_{correct}$ represents the number of correctly-detected faults in the top-K detected faults. For the proposed method, the top-K detected faults are the K PV strings with the highest fault seriousness from a daily report.

B. Effectiveness Evaluation

1) Overall Performance: Figures 3a and 3b exhibit the top-K detection accuracy of different methods at site A and site B, respectively. Here, K ranges from 10 to 100. Compared with the HD method, the proposed Euclidean-based method consistently outperforms the HD method and the proposed method using Cosine-based or Pearson-based similarity. Furthermore, the detection accuracy of the proposed method decays more slowly as K increases. More specifically, the top-100 detection accuracy of the proposed method is 96.3% at site A and 97.1% at site B, compared with 90.3% and 72.1% for the HD method at site A and B, respectively.

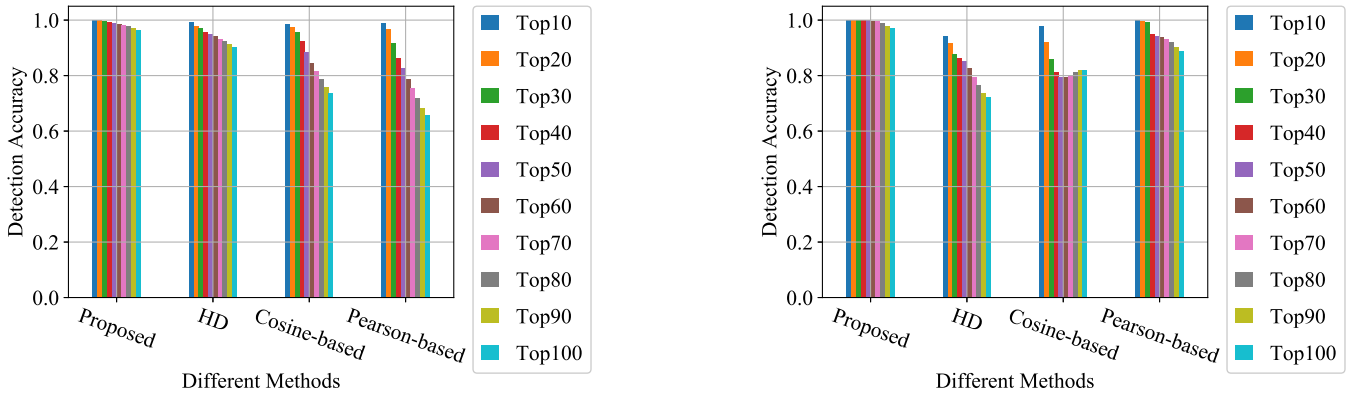
2) Case Studies:

a) Comparison against the HD method: The lower performance of the HD method is mainly caused by noise, which we explain in details below. The HD method identified faulty PV strings from a set of fault candidates whose current values are lower than that of their normal neighbors at most of time during a detected period. However, noise and faults are both likely to cause a PV string's current values to be lower than its normal neighbors most of the time. The proposed method identifies faults if their current values are dissimilar currently while similar in the past. Also, the specific similarity measure is designed to calculate the similarities among neighboring PV strings, which is robust to noise.

Figures 4, 5, and 6 further help to illustrate why the proposed method outperforms the HD method. Figure 4 shows 16 normal PV strings' current values on Jun 2^{nd}. The 16 PV strings are connected to the same combiner box. For the HD method, 3 PV strings (named I10, I11, and I9) are misidentified as faults. As shown in Figure 5, the 3 PV strings' fault seriousness are higher than the threshold, and the rankings of the 3 incorrectly detected PV strings are 64 (I10), 61 (I11), and 62 (I9), respectively. While the proposed method correctly detects the 16 PV strings as normal PV strings, and their fault seriousness are lower than the threshold, as shown in Figure 6.

b) Comparison against other similarity-based methods: The following study helps to better understand why the proposed Euclidean-based similarity measure outperforms 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.2020.2974404, IEEE Transactions on Sustainable Energy



(a) Site A (b) Site B

Fig. 3: Detection accuracy with top-K faults at two PV sites.

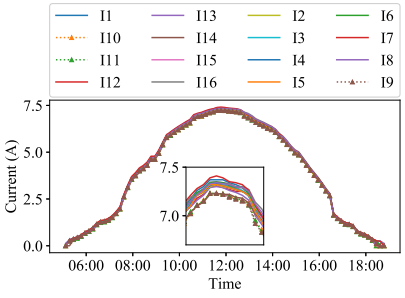


Fig. 4: Normal current values in the same combiner box on a day.

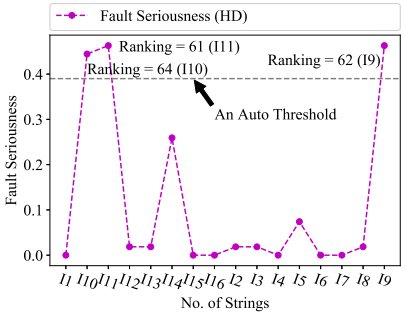


Fig. 5: Fault seriousness of 16 normal PV strings for the HD method.

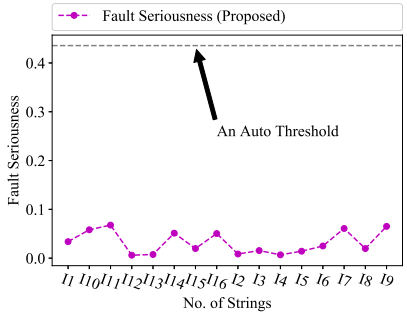


Fig. 6: Fault seriousness of 16 normal PV strings for the proposed method.

Cosine-based method and the Pearson-based method. Figure 7 shows the comparison of three similarity measures among 16 PV strings connected to the same combiner box during continuous 4 days. A PV string (I13) had a fault from day 2 to day 4. As shown in Figure 7, the incipient fault (I13) can be observed to be different from other PV strings using the Euclidean-based method from the second day, and this difference is becoming more apparent on the third and fourth days. In contrast, the faulty PV string cannot be clearly distinguished from other PV strings from day 2 to day 4 using the Cosine-based method and the Pearson-based method.

C. Sensitivity Analysis

Here, we further compare the proposed method with the HD method to better understand the sensitivity of the two methods under various weather conditions. The following experimental results show that the proposed method is more robust against irradiance and cloud cover than the HD method. It is necessary to note that there are no irradiance sensors at site B, this paper only compares the detection accuracy vs. cloud cover at site B.

1) Impact of Irradiance: Figure 8 shows the detection accuracy of the proposed method and the HD method at site A under different irradiance. To ensure fair comparison, we adopt the same experimental setup as that of [1], which estimates the quantile regression model [9] for two quantiles ($\tau = .1$ and $\tau = .9$) to investigate the relationship between top-100 detection accuracy and irradiance. We can see from Figure 8 that, compared with the HD method, the proposed method achieves higher detection accuracy, and this accuracy exhibits smaller variations under the same irradiance. For instance, when the irradiance is $4000\,Wh/m^2/day$, the detection accuracy for the proposed method ranges approximately from 90% to 98%, while the detection accuracy of the HD method ranges from 78% to 97%.

2) Impact of Weather: The proposed method is also more robust to weather variations than the HD method. Figures 9 and 10 show a boxplot summarizing the distribution of top-100 detection accuracy for different weather conditions at site A and B, respectively. These conditions correspond to different amounts of cloud cover. As we can see in the two figures, the proposed method achieves higher median detection accuracy (the horizontal bar within each box) and has less variation than the HD method under the same weather conditions at both sites.

Also, we can see from Figure 10 that site B has more variations of cloud cover during the evaluation period, which introduces more noise to the collected data. The noises significantly decrease the performance of the HD method, but shows less impact on the proposed method.

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.2020.2974404, IEEE Transactions on Sustainable Energy

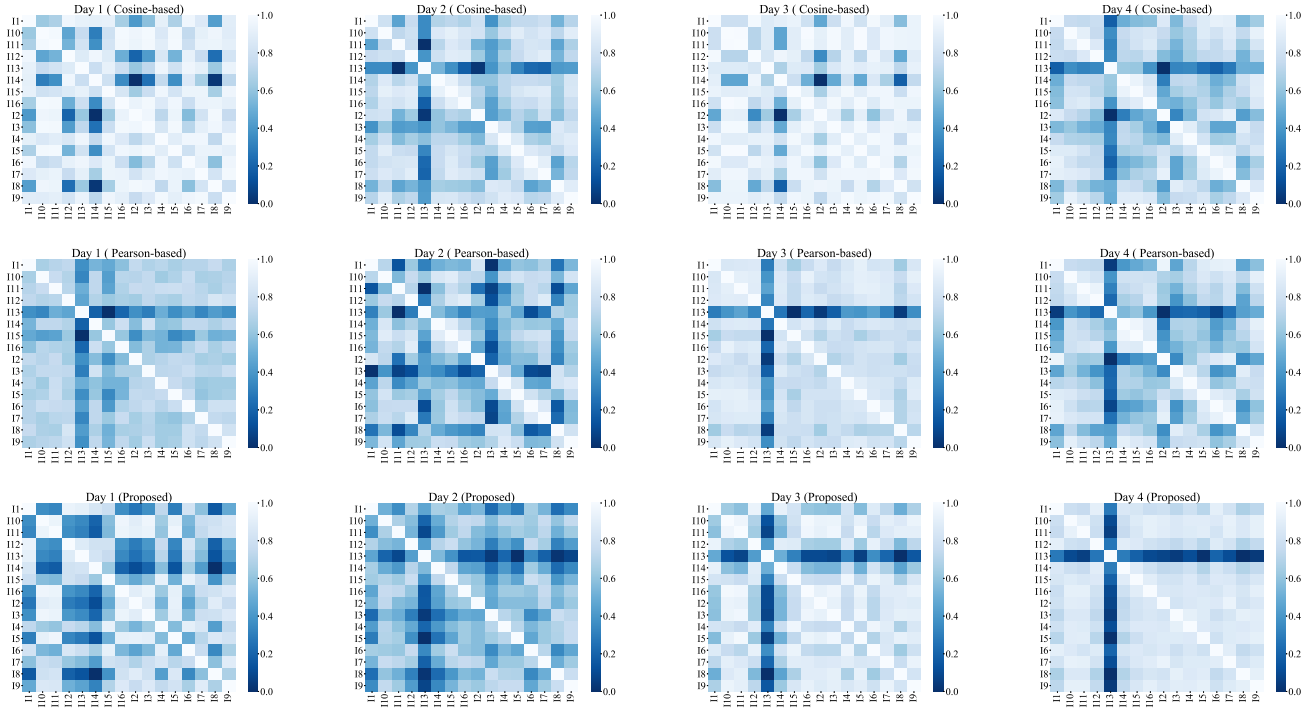


Fig. 7: Comparison of different similarity measures on the PV strings.

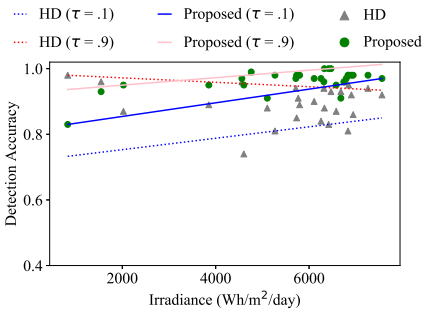


Fig. 8: Detection accuracy vs. irradiance at site A.

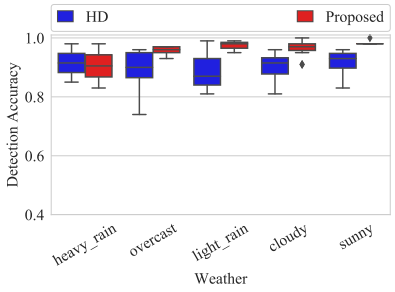


Fig. 9: Detection accuracy vs. weather at site A.

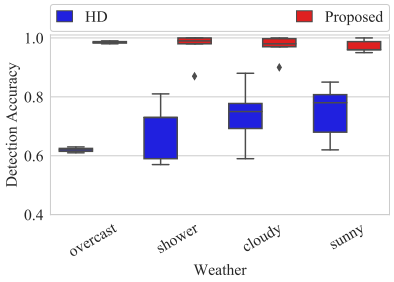


Fig. 10: Detection accuracy vs. weather at site B.

D. Efficiency Analysis

The proposed collaborative fault detection method is implemented on a server with 3.60 GHz CPU. The computation time of processing 30-day collected data is shown in Figure 11. At site A, the computation time is 7.2 min (proposed Euclidean-based method), 6.5 min (proposed Cosine-based method), 6.4 min (proposed Pearson-based method), and 395.5 min (the HD method), respectively, and it is 8.3 min, 7.3 min, 7.4 min, and 491.6 min, respectively, at site B. We can see that: (1) the proposed method achieves approximately 55X and 59X efficiency improvement compared with the HD method at site A and site B, respectively; and (2) the proposed Euclidean-based method achieves almost the same computation efficiency compared with the Cosine-based and the Pearson-based method, while achieving superior accuracy.

V. Related Work

Recent fault detection approaches in PV systems can be categorized into two classes: model-based approaches [5] and data-driven approaches [1].

A. Data-driven Approaches

Data-driven fault detection approaches in PV systems can be further classed into two categories: electrical methods and visual & thermal methods. These methods mainly rely on the understanding of data with a limited requirement of prior domain knowledge [15].

Power loss analysis is one of the widely used electrical methods to detect faults occurred in PV systems. The rationale is that faulty PV devices generate less electricity compared with normal ones. Chouder et al. analyzed power loss to detect faults and further identified faults' severities by measuring

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.2020.2974404, IEEE Transactions on Sustainable Energy

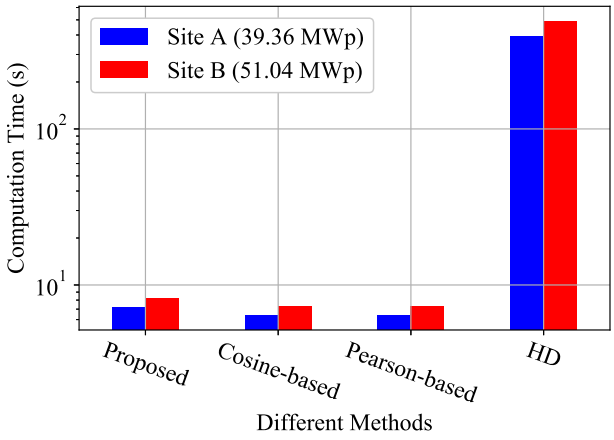


Fig. 11: Comparison of computation time.

their loss durations [16]. There are also fault detection methods based on statistical methods [17], and machine learning techniques, such as artificial neural network [18], [19], [20], Bayesian Neural Network [21], and decision tree [15]. Yi et al. developed a fault detection method based on multi-resolution signal decomposition [22]. Zhao et al. proposed a hierarchical anomaly detection method based on unsupervised learning methods. One common drawback of these machine learning-based methods is that they either require a large amount of pre-labeling to establish models or have high false alarms. However, the proposed method solves the fault detection problem in an unsupervised way, so that no labeled data are required. Another drawback of the existing methods is the high false alarms caused by noises, which can be alleviated by the proposed method as demonstrated in the experiments.

There are also visual & thermal methods (e.g., infrared, thermal imagining), which are capable of accurately detecting and locating the occurrence of faults at the PV module level. As such, these methods have become increasingly popular in recent years [23], [24], [25]. Specifically, these methods detect visual-related and thermal-related faults by implementing the orthophoto infrared thermography by light unmanned aerial vehicle and a thermal imaging system. However, these methods are difficult to deploy in large-scale PV systems due to some practical limitations, such as cost and time-consuming experimental set-up [24].

B. Model-based Approaches

Model-based methods detect faults based on residuals between the actual values and the predicted values generated from a-prior (physical or mathematical) model leveraging domain knowledge [5].

Platon et al. used irradiance and PV module temperature to predict the AC power production, and the residuals between the predicted values and the actual ones are used for online fault detection [8]. Similarly, a model-based method was proposed leveraging temperature and irradiance to predict the healthy PV panel's maximum power [26]. Differently, Dhimish et al. detected faulty PV modules and strings using power ratio and voltage ratio [27]. The main advantage of these model-based methods is high efficiency, and the main drawbacks are: (1) these methods are designed for specific faults, hence have limited scope [1]; and (2) these methods require non-SCADA data collection, such as module temperatures, which increases the overall system's O&M cost.

VI. Conclusions

This paper presents a collaborative fault detection method for PV systems to detect faulty PV strings. The solution has been deployed at two ground-mounted PV systems located in China. Comprehensive theoretical analysis and experimental results demonstrate that the proposed solution outperforms previous methods in terms of effectiveness, efficiency, and robustness. Preliminary on-site measure shows that the proposed method is also applicable to roof/surface mount building integrated photovoltaic systems (BIPV). Also, the similarity metric is the key feature of the proposed method, which is typically more effective than Cosine- and Pearson-based methods for fault detection of PV strings. Future work includes the investigation of applying CF-based techniques to more fault detection scenarios.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant No. 61932007, No. 61233016, and the National Science Foundation (NSF) of United States under grant No. 1442971.

References

[1] Y. Zhao *et al.*, "Hierarchical anomaly detection and multimodal classification in large-scale photovoltaic systems," *IEEE Transactions on Sustainable Energy*, vol. 10, no. 3, pp. 1351–1361, Jul. 2019.

[2] H. Hua *et al.*, "Optimal energy management strategies for energy internet via deep reinforcement learning approach," *Applied Energy*, vol. 239, pp. 598–609, Apr. 2019.

[3] 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.

[4] A. Mellit, G. M. Tina, and S. A. Kalogirou, "Fault detection and diagnosis methods for photovoltaic systems: A review," *Renewable and Sustainable Energy Reviews*, vol. 91, pp. 1–17, 2018.

[5] 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.

[6] 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.

[7] D. Goldberg, "Using collaborative filtering to weave an information tapestry," *Communications of the ACM*, vol. 35, no. 12, pp. 61–70, 1992.

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

[9] R. Koenker and K. F. Hallock, "Quantile regression," *Journal of economic perspectives*, vol. 15, no. 4, pp. 143–156, 2001.

[10] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl, "An algorithmic framework for performing collaborative filtering," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999, pp. 230–237.

[11] N. Sebe, M. S. Lew, and D. P. Huijsmans, "Toward improved ranking metrics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1132–1143, 2000.

[12] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization," in *Proceedings of the 20th International Conference on Neural Information Processing Systems*, 2007, pp. 1257–1264.

[13] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognition*, vol. 38, no. 12, pp. 2270–2285, 2005.

[14] D. Li *et al.*, "Interest-based real-time content recommendation in online social communities," *Knowledge-Based Systems*, vol. 28, no. 2, pp. 1–12, 2012.

[15] 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, p. 043506, 2016.

[16] 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.

[17] Y. Zhao *et al.*, "Outlier detection rules for fault detection in solar photovoltaic arrays," in *Twenty-Eighth Annual Applied Power Electronics Conference and Exposition*. IEEE, 2013, pp. 2913–2920.

[18] 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.

[19] 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.

[20] 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, p. 025501, 2017.

[21] 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.

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

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

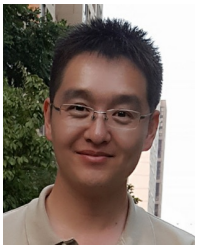
[24] J. A. Tsanakas, L. Ha, and C. Buerhop, "Faults and infrared thermographic diagnosis in operating c-Si photovoltaic modules: A review of research and future challenges," *Renewable and Sustainable Energy Reviews*, vol. 62, pp. 695–709, 2016.

[25] 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.

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

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

Yingying Zhao received the Ph.D. degree in computer science and technology from Tongji University, Shanghai, China, in 2019. She is currently a Post Doctor with Fudan University, Shanghai, China. Her current research interests include renewable and sustainable energy, recommender systems, and machine learning applications.

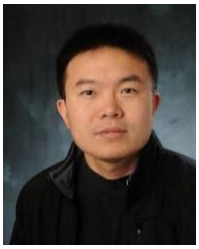


Dongsheng Li is a senior researcher with Microsoft Research Asia (MSRA) since Febuary 2020. Before joining MSRA, he was a research staff member with IBM Research – China since April 2015. He is also an adjunct professor with School of Computer Science, Fudan University, Shanghai, China. He obtained Ph.D. from School of Computer Science of Fudan University, China, in 2012. His research interests include recommender systems and general machine learning applications. In April 2018, he won one of the highest technical awards in IBM – the IBM Corporate Award.

Tun Lu received the Ph.D. degree in computer science from Sichuan University, Sichuan, China, in 2006. He was a Visiting Scholar with HCI Institute, Carnegie Mellon University, USA, from 2014. 9 to 2015. 8. He is currently an Associate Professor with School of Computer Science, Fudan University, Shanghai, China. His research interests include computer supported cooperative works (CSCW), social computing, and humancomputer interaction (HCI). He shared a Best Paper Award at CSCW'15 and a Honorable Mention Award at CSCW'18. He is a Senior Member of China Computer Federation (CCF) and a Member of ACM. He is the Secretary General of CCF Technical Committee of Cooperative Computing. He has been active in professional services by serving as PC Co-Chairs (e.g. ChineseCSCW'17 & 18 & 19, CSCWD'10), Associate Chairs (e.g. CHI'19 & 20, CSCW'19 & 20), PC members (e.g., GROUP'18, CRIWG'17 & 2018, CSCWD'16), Guest Editors (e.g. International Journal of Cooperative Information Systems, Chinese Journal of Computers) and reviewers for many well-known journals and conferences.

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, USA, in 2006. She is currently an Associate Professor with the Department of Computer Science, University of Colorado Boulder, Boulder, CO, USA. She has authored or coauthored more than 60 papers with over 5000 citations. Her research integrates systems, algorithms, and applications for effective and efficient data analytics in ubiquitous computing and scientific discovery. Her research interests include mobile/wearable computing, social networks, spatial-temporal data, anomaly/misbehavior detection, recommender systems, and multimodal 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. She is an Associate Editor of the *PACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* and has served on the technical program committees and organizing committees of many international conferences. She has received the 2017 Google Faculty Research Award, the VLDB 2017 Ten Year Best Paper Award, the ICTAI 2017 Best Student Paper Award, two Best Paper Award nominations, and the Pervasive 2012 Computational Sustainability Award.

Ning Gu received the Ph.D. degree from Institute of Computing Technology, Chinese Academy of Sciences, China, in 1995. He is currently a professor with School of Computer Science, Fudan University, Shanghai, China. His research interests include human-centered cooperative computing, CSCW and social computing, and human computer interaction.



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, Boulder, CO, USA. He has authored or co-authored more than 100 publications in computer systems, mobile computing, and design for high-performance information systems. He was an Associate Editor for the IEEE Transactions on Very Large Scale Integration (VLSI) 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 field-programmable gate array (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. 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 National Science Foundation CAREER Award.