# Hierarchical Context-Aware Anomaly Diagnosis in Large-Scale PV System Using SCADA data (OUTLINE)

***Abstract*— Accurate anomaly diagnosis is essential for reducing operation and maintenance (O&M) cost, while improving safety and reliability of a large-scale photovoltaic (PV) system. This paper presents a hierarchical context-aware anomaly diagnosis approach in PV system, which requires no additional hardware support beyond widely adopted supervisory control and data acquisition (SCADA) system. The proposed approach aims to implement accuracy anomaly diagnosis based on unsupervised machine learning techniques for strings of a PV system. It is motivated from strings patterns that are related with their electrical characteristics in space and time. The spatial and temporal distribution of different strings provides rich indications for understanding theirs operation patterns so as to implement anomaly diagnosis. Evaluations using a 40MW PV system located in Northeast China reveal that the proposed diagnosis approach can implement string-level anomaly diagnosis with accuracies higher than 97% on a daily basis, which satisfy PV plants requirements.**

***Index Terms*— Context-aware, anomaly diagnosis, PV system.**

## I. Introduction

Photovoltaic (PV) systems, as one of mature technologies for power production from renewable energy sources, has shown a rapid growth over the past years. However, as shown in recent studies [?], the large-scale PV system incurs ever-increasing risks of operation and maintenance (O&M) concerns. How to minimize O&M cost, while improve safety and reliability of a large-scale photovoltaic (PV) system has become major concerns for PV plants. Strings, as one of the most important parts in PV system, and the failures of which occurs at a high frequency and is hard to implement diagnosis for the large area of a large-scale PV system. More seriously, these failures is extremely damaging to a PV system, such as fire hazards [?]. To this end, the goal of this study is to develop effective diagnosis approach for large-scale PV system on string-level basis to reduce O&M cost, while improve safety and reliability of PV systems.

Recent research work has tackled the anomaly diagnosis for PV systems. High-cost is the primary limiting factor to one of the existing solutions [?], as expensive, extra sensing devices are required [?]. For instance, Radu et al. [online fault detection in pv systems] developed an online fault detection method by developing a model to analyze solar irradiance and PV panel temperature provided by extra a sunshine pyranometer (SPN) and temperature transducers on each PV panel. Due to the high installation and maintenance cost, to date, the adoption rate of extra sensing devices has been low in existing PV plants. Compared with model-based methods, data-driven methods do not rely on expensive extra devices, but utilize the SCADA system, which is considered as a standard installation equipment. These methods, unfortunately, either have a high false alarm rate [hampel identifier], or are not computation-efficient [LOF] for large-scale PV systems. More importantly, these methods are usually based on statistical techniques[], supervised[??], or semi-supervised [ zhaoye] machine learning techniques. Numerous true labelled data is needed. In a real-world PV plants, it is expensive and difficult to collect the labelled data. In addition, there are still some hybrid methods combining model-based and data-driven methods that implement performance evaluation of PV systems. However, on the one hand, these methods are based on simulation environment and lack practice application [?]. On the other hand, these methods usually localize on a system-level anomaly detection, which can not conduct anomaly isolation, identification and classification on a fine-grained anomaly diagnosis for offering better scheduling repair, such as string-level.

This work aims to tackle string-level anomaly diagnosis problem for a large-scale PV system using SCADA data. The proposed work is based on a hierarchical context-aware anomaly diagnosis model. First, our model is based on a data-driven method such that its implementation does not require any redundant purpose-built sensing device. Second, the hierarchical model ensures the combination between local context and global context, which can minimize the ambient effect so as to reduce false alarm effectively while achieve more accuracy diagnosis rate. Third, our proposed model is based on unsupervised machine learning techniques so that the labelled data is needless ??????. Most importantly, the proposed approach is developed and validated using real world 40MW PV plants data (8199 strings over 1 year). Experimental results demonstrate that the proposed approach can perform accurate diagnosis for string-level anomaly on a daily basis, which is sufficient to schedule maintenances activities.

In this paper, our key contributions are as follows:

1) We are the first to identify and analyze which contexts can be exploited for string operation patterns recognition and anomaly diagnosis using SCADA system theoretically. Our studies show that statistical information in temporal and spatial, based on string-level, which can be adopted to infer strings operation patterns on a daily basis;
2) A hierachical context-aware anomaly diagnosis method is proposed to automatic identify strings operating state, in which spatio-temporal information are considered in a Gaussian Mixture Model (GMM) while combining unsupervised machine learning algorithm to achieve accurate anomaly diagnosis;

3) The proposed approach is evaluated using a real-world PV plant data collected by SCADA system, and experimental results demonstrate that the proposed diagnosis method can accurately infer strings state with accuracies higher than 97%;

The rest of this paper is organized as follows. Section II outlines the current related works. Section III discusses the challenges in anomaly diagnosis of PV systems. Section IV provides the model formulation. Section I presents the hierarchical context-aware anomaly diagnosis model in detail. Experimental results are presented in Section VI. Finally, we conclude this work in Section VII.

## II. Related Work

Existing anomaly diagnosis methods for PV systems can be categorized into three classes: model-based approaches [?], data-driven approaches [?] and hybrid approaches [?].

Model-based approaches typically use an explicit mathematical model to reason the causality of the components inherent operation principle. A significant difference between the model output and measured signal was considered an anomaly. ..... In summary, these model-based approaches typically require operation features that can only be collected by redundancy sensing device, which can not be supported by SCADA system.

The data-driven methods utilizes statistical approaches and machine learning techniques, which learn models directly from the data. ..... On the one hand, it requires large amount of data and numerous examples to develop an accurate model. In practice, it is very expensive and difficult to obtain labelled data to establish an accurate model.

The hybrid approaches that combines model-based approaches with data-driven approaches has been widely studied. ..... Overall, these hybrid approaches usually localize on a system-level anomaly detection, which can not conduct anomaly diagnosis on a fine-grained anomaly diagnosis for offering better scheduling repair, such as string-level.

## III. Challenges

The primary challenges to tackle anomaly diagnosis has always been the exploration of a set of meaningful and actionable data analysis methods to indicate the hidden relationship among limited few signals collected by SCADA system so as to recognize strings operation patterns. Compared with techniques using redundancy sensing device to collect signals (e.g, voltage of each string[?], temperature of each module[?].), SCADA system is not initially designed for anomaly diagnosis purpose so that it collects limited signals, e.g. current of each string and voltage of each combiner box. Therefore, it is challenging to capture anomaly patterns of stings from the raw SCADA data without effective data analysis method.

Nevertheless, SCADA system provides information which can potentially reveal strings operations states. In PV systems, power output has been widely used to diagnose system-level anomaly. Thus, anomaly detection can be conducted, anomaly identification, isolation and classification can not be implemented using the method. Current is an another widely used feature to diagnose string-level or module-level anomaly. However, on the one hand, the high false alarm rate is main limitation of these methods. On the other hand, It is hard to set a threshold to differ the anomaly string and normal string in the same combiner box because the current varies randomly. Figure ?? shows 15 strings that operates properly and one that experience an anomaly in the same combiner box [need a figure here]. Figure ?? shows 16 strings operates properly in the same combiner box. Figure ?? shows 1 normal string and 1 anomaly string in two different combiner boxes.

The diversity and complexity of anomaly in string-level poses another challenge in diagnosis methods. In general, anomalies in string-level fall into three classes: visual anomaly (browning, discoloration, surface soiling, and delamination), thermal (hot spot)[ a novel fault diagnosis technique ], and electrical (ground fault, line-line fault, open-circuit, and mismatch fault)[zhao ye ]. [give some pictures in real world pv plants]. Most current approaches are built on case-by-case basis that can not be possible to conduct a general anomaly detection. On the other hand, the aforementioned anomalies are interrelated and not fixed. For instance, the long-term partial shading cause to hot spot effects.

The lack of true labelled data is is another challenges we need to consider. Many elaborate works use statistical methods or supervised or semi-supervised machine learning techniques, which are based on many true labelled data. However, it is very expensive and difficult to obtain true labelled data in a real world PV plants. Therefore, it is challenging to establish an unsupervised-based method to conduct .....

## IV. MODEL FORMULATION

### A. Preprocessing

Current is the major concern. However, currents show high fluctuation. An instantaneous current may not decide whether there is an anomaly is a string.

TABLE I: PARAMETERS IN OUR REAL WORLD PV SYSTEM

| Parameters | Symbol | Value |
|---|---|---|
| Area of the PV module | A | 1.941 $m^2$ |
| Maximum Power | $P_{mppt}$ | 300 W |
| Maximum Power Voltage | $V_{mppt}$ | 36.50 V |
| Maximum Power Current | $I_{mppt}$ | 8.22 A |
| Open-circuit voltage | $V_{OC}$ | 45.3 V |
| Short-circuit Current | $I_{SC}$ | 8.79 A |
| The number of series solar cells per module | $N_S$ | |
| temperature factor | $\delta$ | -0.43% |
| Standard Test Conditions Tempture | $T_{STC}$ | 25°C |
| Standard Test Conditions irradiance | $G_{STC}$ | 1000 W/$m^2$ |
| conversion efficiency for module | $\eta_r$ | 15.45% |

### B. Local Context

All strings in the same combiner box are supposed to exhibit the similar characteristics. Thus, the local context in the same combine box

### C. Global Context

In a PV system, most of strings function properly most of the time, and anomaly strings number is relative few to the total number of strings. combiner box -level.

# V. PROPOSED HIERARCHICAL CONTEXT-AWARE ANOMALY DIAGNOSIS MODEL

## A. Local Context-Aware Algorithm

....an algorithm goes here

We can calculate the normal module temperature.

$$T_{cell} = \frac{f(v) \cdot T_a + G(t) \cdot (\alpha \cdot \tau - \eta_r - \delta \cdot \eta_r \cdot T_{STC})}{f(v) - \delta \cdot \eta_r \cdot G(t)} \quad (1)$$

$$f(v) = 17.1 + 5.7 \cdot v \quad (2)$$

Where $v$ is wind speed; $T_a$ is the ambient temperature; $\alpha \cdot \tau = 0.9$.

## B. Global Context-Aware Algorithm

....an algorithms goes here

$$CPR = \frac{PR}{1 + \delta \cdot (T_{cell} - T_{STC})} \quad (3)$$

$$PR = \frac{E(t)_{AC}}{E_{input}} \quad (4)$$

$$E_{input} = \int G(t) \cdot A \cdot \eta_r \cdot dt \quad (5)$$

Where $\delta$ is the corrective factor of temperature; $T_cell$ is the module temperature. $T_{STC}$ is the standard temperature; $E(t)_{AC}$ (Wh) is the electrical energy output from combiner box recorded over a time minute period $t$; $G(t)$ is the in-plane solar irradiation received on an unshaded surface of the same location; $A$ ($m^2$) is the area of the PV module; $\eta_r$ is the conversion efficiency of an array.

## C. Hierarchical Context-Aware Algorithm

...an algorithms goes here

# VI. EXPERIMENTS AND RESULTS

## A. Dataset Description and Evaluation Metrics

There are 74 inverters, 553 combiner boxes, 8199 strings , 131184 modules in a real-world PV system. There are 16 modules in each string. There are 72 cells in each module.

## B. Overall Performance

## C. A Case Study

# VII. CONCLUSION