ABSTRACT
Solar arrays often experience faults that go undetected for long periods of time, resulting in generation and revenue losses. In this paper, we present SunDown, a sensorless approach for detecting per-panel faults in solar arrays. SunDown’s model-driven approach leverages correlations between the power produced by adjacent panels to detect deviations from expected behavior, can handle concurrent faults in multiple panels, and performs anomaly classification to determine probable causes. Using two years of solar data from a real home and a manually generated dataset of solar faults, we show that our approach is able to detect and classify faults, including from snow, leaves and debris, and electrical failures with 99.13% accuracy, and can detect concurrent faults with 97.2% accuracy.

CCS CONCEPTS
• Hardware → Renewable energy; • Computing methodologies → Anomaly detection.

KEYWORDS
Solar anomaly detection; data-driven modeling; machine learning

1 INTRODUCTION
Recent technological advances and falling prices has led to a significant increase in deployments of both large utility-scale and smaller residential solar arrays. Large utility-scale solar farms tend to be instrumented with sensors for monitoring real-time generation to identify production issues. Due to cost reasons, smaller residential-scale systems lack such sensing and instrumentation and may only have coarse-grain monitoring capabilities, at best, to detect system-level faults. Thus, it is not uncommon for residential solar arrays to encounter power anomalies or other local faults that go undetected for long periods, resulting in generation and revenue losses.

To address these challenges, we present SunDown, a sensorless approach for detecting per-panel faults in small-scale solar arrays. Prior work on per-panel solar anomaly detection are based on time series [15] or statistical [3, 30] analysis of a panel’s output or use of sensors such as a pyranometer [12] to detect faults. In contrast, our approach uses the actual output from other nearby panels to estimate each panel’s expected output and find anomalous deviations from this estimate. Our model-driven approach is based on machine learning and, similar to [15], can detect physical anomalies, such as snow, leaves, and electric faults at panels. In designing, implementing, and evaluating our SunDown system we make the following contributions.

1. We present a model-driven approach that leverages correlations in the generated output between adjacent panels to predict the expected output of a particular panel and flags anomalies when the model predictions deviate from the expected values. Further, our approach can handle and detect multiple concurrent faults in the system, a key challenge that has not been addressed by prior work. We present a random forest-based classification technique to classify the probable cause of the observed fault.

2. We construct a real-world labelled dataset of solar anomalies that we release to the community. Using this dataset, we show that SunDown has a MAPE of 2.98% when predicting per-panel output, demonstrating the efficacy of using nearby panels to perform model-driven predictions. Furthermore, SunDown is able to detect and classify faults such as snow cover, leaves, and electrical failures with 99.13% accuracy for single faults and is able to handle concurrent faults in multiple panels with 97.2% accuracy.

2 BACKGROUND
This section presents background on solar anomaly detection.

Residential Solar Arrays. Our work primarily focuses on residential solar arrays that are typically small-scale installations with capacities of 10kW or less and comprise a few to a few dozen solar panels (see Figure 1). We assume that the power generation of the array can be monitored at a per panel level. This is a reasonable assumption since many residential arrays are equipped with micro-inverters (e.g. Enphase micro-inverters [1]) or DC power optimizers [2]. As shown in Figure 1, such systems provide real-time per-panel
We assume that the panels are mounted on a residential roof across well as site specific factors, such as shading caused by nearby trees highly correlated weather conditions, and produce similar output. If the deviation is “large” and persists over an extended period of time, it is indicative of a fault, rather than an error in the model prediction. The cause of the fault can be separately determined by analyzing the amount of loss or the power pattern exhibited by the panel. Such a model-driven approach only uses the panels’ observed output to detect anomalies—no other instruments or sensors are needed for anomaly detection unlike some approaches [5].

3.2 Model-Based Predictions

We now present two model-driven techniques for predicting the power output of an individual panel using neighboring panels.

3.2.1 Linear Regression-Based Model. Since the power generated by solar panels in close proximity of one another are highly correlated, we can use regression to predict the output of a panel given the observed output of neighboring panels. Let \( P_i \) denote the observed power output of panel \( i \) at time instant \( t \). Let us assume we wish to predict the output of panel \( i \) using \( n \) other panels. A linear regression model allows us to estimate the output of desired panel as a linear function of the others:

\[
\hat{P}_i = w_1 P_{i1} + w_2 P_{i2} + w_3 P_{i3} + \ldots + w_n P_{in} + \epsilon_i
\]

where \( X = \{i_1, i_2, \ldots, i_n\} \) is the set of \( n \) panels used to model the output of the \( i^{th} \) panel. We can use linear regression to estimate the weight \( w_i \) that minimizes the error term \( \epsilon_i \).

Such an approach yields \( N \) distinct regression models, one for each panel in the system, where each model makes a prediction using the observed output of \( n \) other panels. To determine if a panel has a fault, we compare the model prediction \( \hat{P}_i(t) \) at time \( t \) with the observed value \( \hat{P} \). If the difference between the model’s prediction and observed value is large and persists over a period of time (e.g., a day or multiple days), the approach flags that panel as faulty.

3.2.2 Graphical Model and Half-Sibling Regression. Our second model is based on a recently proposed machine learning technique called half-sibling regression that uses a Bayesian approach to remove the effects of confounding variables [28]. This approach is based on our prior work on SolarClique [17] that predicted the output of an entire array using nearby solar arrays. We draw inspiration from the half-sibling regression method [28] and SolarClique [17] for SunDown’s per-panel anomaly detection. Additional details of our approach, which is summarized below, can be found in [11]. Using the Bayesian approach, our algorithm to estimate the amount of production loss due to anomalies is as follows.

We first use regression to estimate the power output of a particular panel, denoted by a random variable \( P \), using the power output of \( n \) other panels in the system, denoted by a random variable \( X \) (a vector of size \( n \)). The regression yields \( E[P|X] \) - an estimate of \( P \) given the observed output of \( n \) neighboring panels that constitute
A model based on panels with normal output, and a model based on panels that is faulty. The classifier needs to distinguish between three types of faults: snow, partial occlusion and open circuit. Note that errors leading to higher false positive as compared to the Bayesian model. Given anomalies detected by our Bayesian model we use a random forest classifier to label the possible cause of the fault for each panel. Our approach can also label system-wide faults, caused either by a system-wide electrical failure or full snow cover, both of which cause near total loss of power output.

Figure 3: Machine Learning Model

\[ \text{MAPE} = \frac{1}{m} \sum_{t=1}^{m} \left| \frac{P_O(t) - P_I(t)}{P_O} \right| \]  

Here, \( m \) is the number of samples, \( P_O(t) \) is the observed solar power at time \( t \), \( P_I(t) \) is the inferred power at time \( t \), and \( P_O \) is the mean of observed power generation. We use three different metrics to quantify different aspects of the classification task: accuracy, sensitivity, and specificity.

Solar Anomaly Open Dataset. Since there are no datasets of solar faults available for research use, we constructed a labelled dataset using two arrays: a 31-panel production residential site, and a 20-panel ground-mounted site where we introduced anomalies, such as dust, leaves, and electrical faults, to mimic real-world faults. Our dataset is available at http://traces.cs.umass.edu and details of our dataset construction can be found in [11].

5 EXPERIMENTAL EVALUATION

We evaluate SunDown by quantifying (1) the accuracy of model-based power inference where we infer the output of a single panel using nearby panels and (2) the accuracy of our anomaly classification. We quantify the accuracy of predicting a panel’s output using Mean Absolute Percentage Error (MAPE) between the inferred output and the actual solar generation, as below.

5.1 Prediction Model Accuracy

We begin by evaluating the accuracy of predicting the power output of an individual panel using neighboring panels.

5.1.1 Machine Learning Model. To evaluate the accuracy of model inference, we choose test data only from the days where the site experiences no anomaly. We then use the normal days of the home dataset to train our linear regression and graphical model. We also compare their performance with a naive approach that infers the power output of a panel as the mean output of \( n \) other panels. As...
We next evaluate the accuracy of our model-driven approach and cover and use them as inputs to our random forest classifier. Table 1 shows our model can classify single faults with an accuracy of 98.78%, specificity of 97%, and sensitivity of 100%. For concurrent faults, the model obtains accuracy of 97.2%, specificity of 97.06%, and sensitivity of 97.26%.

### 5.2 Anomaly Classification Accuracy

We next evaluate the accuracy of our model-driven approach and classifier in detecting and classifying anomalies, respectively. The common anomalies we consider include snow fault, open circuit, and partial occlusions due to leaves.

Our home dataset already includes real snow faults that are labelled and we evaluate the accuracy of our classifier on identifying these snow faults. We then use the synthetic faults from our solar anomaly dataset and inject them into the home data set by introducing synthetic single panel faults as well as concurrent fault and evaluate the accuracy of our classifier. Figure 5 presents per-panel data for a typical day when an electric fault or object covering anomaly has been injected into one or many panels.

5.2.1 Snow Fault Detection. We first evaluate the ability of our classifier in detecting snow faults in the home dataset. We extract the features from daily power output, which include Pearson’s correlation coefficient, ratio of maximum observed power and the nominal panel capacity, and weather data such as snow and cloud cover and use them as inputs to our random forest classifier. Figure 5 presents per-panel data for a typical day when an electric fault or object covering anomaly has been injected into one or many panels.

5.2.2 Single and Concurrent Fault Classification. We next show that our approach is capable of fine-grain anomaly detection and classification of a single fault and it is also capable of detecting concurrent faults in a subset of the panels. To do so, we use our solar anomaly dataset and choose the partial occlusion and open circuit anomaly from the dataset and inject these faults into a single, randomly chosen, panel of the array; different panels have faults injected into them on different days. We use our model to detect the presence of the fault and our random forest classifier to identify the type of fault. We next inject multiple concurrent faults of all types (snow, occlusion, open circuit) into the array using a similar methodology and attempt to detect and classify each fault using our model and classifier. Note that, in this case, we need to use our concurrent fault detection approach. Table 1 shows our model can classify single faults with an accuracy of 98.78%, specificity of 97%, and sensitivity of 100%. For concurrent faults, the model obtains accuracy of 97.2%, specificity of 97.06%, and sensitivity of 97.26%.

### 6 RELATED WORK

There has been significant work on predicting power output for solar sites [4, 6, 10, 23, 24, 27, 29]. All of these studies predict only system level output and generally report 20-30% error. These high errors and inability to predict panel level output would cause their prediction for all panels to be the same, and limit their ability to detect and classify anomalies. There is also significant prior work on anomaly detection and classification in solar photovoltaic systems, which can be broadly classified into model-based approaches [9, 13, 16, 19, 20] and machine learning based [7, 8, 12, 14, 21, 22, 25, 26, 31, 32] approaches. Some of these studies use power data from nearby solar sites [17, 30] to detect and classify anomalies. In [30], authors compare the performance of different solar arrays at the same site, but do not do anomaly classification. Our work uses the output of other nearby panels to predict a panel’s output for detecting faults and can classify various types of faults, i.e. snow, object covering, and electrical faults, on a single or multiple panels.

### 7 CONCLUSIONS

In this paper, we proposed SunDown, a sensorless approach to detecting per-panel anomalies in residential solar arrays. We take a model-driven approach that leverages correlations between the power produced by adjacent panels to detect deviations from expected behavior. We constructed and released an open dataset of solar anomaly faults for experimental use. Finally, we showed that our approach can predict panel level output with a MAPE of 2.98% and can correctly classify anomalies with >97% accuracy.

**Acknowledgements**

This research was supported by NSF grant 1645952, DOE grants DE-EE0007708, DE-EE0008277, and MA DOER.
REFERENCES