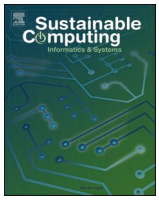
Sustainable Computing: Informatics and Systems 32 (2021) 100595

Contents lists available at ScienceDirect

Sustainable Computing: Informatics and Systems

journal homepage: www.elsevier.com/locate/suscom



SolarDiagnostics: Automatic damage detection on rooftop solar photovoltaic arrays

Qi Li *, Keyang Yu, Dong Chen

Knight Foundation School of Computing and Information Sciences, Florida International University, United States

ARTICLE INFO

Keywords:
Solar energy
Deep learning
Machine learning
Anomaly detection

ABSTRACT

Homeowners are increasingly deploying rooftop solar photovoltaic (PV) arrays due to the rapid decline in solar module prices. However, homeowners may have to spend up to ~$375 to diagnose their damaged rooftop solar PV system. Thus, recently, there is a rising interest to inspect potential damage on solar PV arrays automatically and passively. Unfortunately, recent approaches that leverage machine learning techniques have the limitation of distinguishing solar PV array damages from other solar degradation (e.g., shading, dust, snow). To address this problem, we design a new system—SolarDiagnostics that can automatically detect and profile damages on rooftop solar PV arrays using their rooftop images with a lower cost. In essence, SolarDiagnostics first leverages an K-Means algorithm to isolate rooftop objects to extract solar panel residing contours. Then, SolarDiagnostics employs a convolutional neural networks to accurately identify and characterize the damage on each solar panel residing contour. We evaluate SolarDiagnostics by building a lower cost prototype and using 60,000 damaged solar PV array images generated by deep convolutional generative adversarial networks. We find that SolarDiagnostics is able to detect damaged solar PV arrays with a Matthews correlation coefficient (MCC) of 1.0. In addition, pre-trained SolarDiagnostics yields an MCC of 0.95, which is significantly better than other re-trained machine learning-based approaches and yields as the similar MCC as of re-trained SolarDiagnostics. We make the source code and datasets that we use to build and evaluate SolarDiagnostics publicly-available.

1. Introduction

Homeowners are increasingly deploying rooftop solar photovoltaic (PV) arrays due to the rapid decline in solar module prices. To illustrate, the cost of solar energy in $/W dropped an estimated ~80% from 2010 to 2018, resulting in a ~700% increase in solar energy capacity in U.S. over the same period [1]. Solar power prices have now fallen below retail electricity rates in many areas, further increasing the incentive for homeowners to install solar modules [1]. In the first quarter of 2019, over ~70% of solar deployments in U.S. are continuously small-scale solar PV arrays from residential rooftops. Recent news [2] showed solar owners may spend up to $375 per year on the services to maintain their "degraded" rooftop solar PV systems, including damaged solar PV panel inspection, wiring damage, annual inspection, damage localization, and solar PV array cleaning, which typically are not covered in their purchase warranty. Thus, recently, there is a rising interest to inspect potential damage on rooftop solar PV arrays automatically and passively with a lower cost.

Traditional approaches [3–5] relied on *I–V* curve and *P–V* characteristic monitoring of the target rooftop solar PV system in nominal and faulty conditions and had the accuracy as ~60%. These approaches require user expertise in measuring model parameters and hardware installation such as cameras and solar radiation sensors to collect training data. Thus, significant recent work focuses on data-driven approaches [6–12] that leverage machine learning techniques to train accurate classifiers to identify damages. These approaches require significant amount of historical pure solar generation data, which may not be available due to the new solar sites become online, to calibrate their models, and also can not accurately distinguish solar PV array damage from other degradation, such as shading, dust, snow, cloudy, and so on. Thus, new techniques are necessary.

To address this issue, we design a new system—SolarDiagnostics that can automatically detect and localize damage on solar PV arrays with a lower cost. Our hypothesis is that solar PV arrays are visually identifiable in their rooftop images such that we can leverage image processing and deep learning techniques to automatically profile information. In evaluating our hypothesis, this paper makes the following contributions.

Detection Challenge. We highlight numerous challenges that we

* Corresponding author.
E-mail addresses: qli027@fiu.edu (Q. Li), kyu009@fiu.edu (K. Yu), dochen@cis.fiu.edu (D. Chen).

https://doi.org/10.1016/j.suscom.2021.100595
Received 11 June 2020; Received in revised form 22 April 2021; Accepted 3 August 2021
Available online 9 August 2021

met to detect damages on solar PV arrays automatically using only their rooftop images. Rooftop solar PV arrays segmentation and damage identification are affected by numerous unknown variables, e.g., imbalanced training dataset, inaccurate shape and color features, shade, size, orientation, topography, and other outliers on rooftop.

Damaged Solar PV Array Image Dataset Buildup. We leverage deep convolutional generative adversarial networks (DCGANs) to build an unsupervised approach to automatically generate 60,000 damaged solar PV array rooftop images. By carefully designing this approach, we insure that the discriminative model of the DCGANs has the worst accuracy to distinguish the "real" and "fake" images. In addition, we also collected groundtruth for each image, such as damage level and brand. In doing so, we are preparing a balanced dataset to train and evaluate our new approaches.

SolarDiagnostics Design. We design a new system—SolarDiagnostics that can automatically detect and localize damages on rooftop solar PV arrays with a lower cost (~$35). In essence, SolarDiagnostics first leverages an unsupervised segmentation algorithm to isolate rooftop objects each image and extract solar panel residing contours. Then, SolarDiagnostics employs a convolutional neural networks (CNNs) to accurately identify and characterize the damages in each contour.

Implementation and Evaluation. We implement SolarDiagnostics in `python` using widely available open-source frameworks, including Pandas [13], OpenCV [14], Scikit-learn [15] and PyCUDA [16,17]. We evaluate SolarDiagnostics using our large damaged solar PV array image dataset and also by building a SolarDiagnostics prototype. We find that SolarDiagnostics can accurately detect damaged rooftop solar PV arrays and also learn the damage profiling information for each solar site. We evaluate SolarDiagnostics using multiple ways: (1) We compare SolarDiagnostics's results with the groundtruth labeled 60,000 rooftop images that are generated using our DCGANs model. (2) We validate SolarDiagnostics's detection results using the groundtruth image data for 350 sites that we fetched using Google Images API [18] and our prior work [19]. (3) We validate SolarDiagnostics's profiling accuracy for 10 "mock" solar PV arrays using the SolarDiagnostics prototype.

Releasing Datasets and Source Code. We release all the datasets that are comprised of over 60,360 solar PV array rooftop images and the source code of SolarDiagnostics on our website [20].

2. Background and related work

2.1. Problem statement

Given a solar-powered home, we first want to build a new approach that can automatically fetch its rooftop image. We then present a new approach that can accurately segment rooftop objects and focus on solar panel residing contours in each image. We further seek to build a deep learning classifier to accurately identify the damage on each solar PV array. For each reported solar PV array, we also want to learn its profiling information, such as damage location, damage level and manufacture brand, which can be used to assist solar owners to repair or replace their solar PV arrays promptly. Formally, given a solar PV array-powered home S_i, we first need to segment its rooftop objects $O_i(1 \leq i \leq N)$ on rooftop image I_i into small "contours"—C_i, where N is the number of objects. Then, we will identify the contours that have damaged solar PV panels and report their damage level, damage location, and brand information.

2.2. Related work

Traditional approaches [3–5] focusing on engineering methods to detect faults of solar PV arrays can be classified into the following categories: (1) statistical signal processing based approaches [21,22]; (2) *I–V* characteristics analysis [23–25]; (3) power loss function-based analysis [26–28]; and (4) voltage and current measurement based approaches [29–31]. The prior approaches [4,5] leverage these electrical methods to monitor the rooftop solar PV system in nominal and faulty conditions, respectively. However, these approaches all require hardware installation such as cameras, and solar radiation sensors to collect training data, and also user expertise in measuring model parameters when building classifiers.

In contrast, significant recent work focuses on data-driven approaches [6–12] that leverage machine learning and deep learning techniques to train accurate machine learning or deep learning classifiers. The major issue is that these approaches typically require significant amount of historical pure solar generation data to calibrate their models, which may not be available due to the new solar sites become online, and also have the difficulty to distinguish solar PV array damage from other degradation, e.g., shading, dust, snow, and cloudy. Thus, new techniques are necessary.

2.3. Summary

As we had discussed in this section, the prior approaches that leverage electrical engineering methods typically require hardware installation to monitor and model the faulty and normal conditions. While, recent machine learning and deep learning based data analytics methods require significant amount of training data which is not always available. These valuable insights guide the development of our proposed technique—SolarDiagnostics.

3. Detection challenges

Highly imbalanced data. The prior work mainly builds machine learning or deep learning classifier using a high imbalanced datasets with ~80% rooftop images having no damage on their solar PV arrays. This is mainly due to the fact that damaged solar PV array image dataset are not immediately publicly-available. Thus, these approaches may not be able to accurately and reliably identify damaged solar PV arrays. These data-driven approaches may need more "negations" to achieve a reasonable accuracy. To address this challenge, we design a deep convolutional generative adversarial networks (DCGANs)-based approach that can automatically generate a large amount of damaged solar PV array images. In doing so, we are able to build our SolarDiagnostics using a balanced training dataset that has the ratio of positive and negative samples as 1:1. More details are in Section 4.1.

Automatic rooftop image segmentation. In addition to solar PV arrays, many other objects, such as ridge, structure, chimney, shade and window, may also present on each rooftop. In particular, the shape features of the ridges, structures and shades have significant overlapping with solar PV arrays. This makes these data analytical approaches having more difficulty to train a reasonably accurate classifier to distinguish different solar degradation. To address this issue, we design an unsupervised machine learning-based approach that integrates with our recent SolarFinder work [19] to automatically segment objects in each rooftop image and only focus on solar panel residing contours. More details can be found in Section 5.1.

Inaccurate shape and color features. The shape and color features of rooftop solar PV arrays from different manufactures might look different in their rooftop images. In addition, the gap lines including gaps, bus bars and fingers on solar PV arrays, which are the spaces between the solar cells and necessary to allow for thermal expansion of the cells when the panels heat in the sun, always present white rectangle noise bars in their rooftop images. These white noises in rooftop solar PV array images significantly reduce the damage detection accuracy. To address this challenge, we design an algorithm that leverages unsupervised learning approach to automatically remove these white "outliers" when performing damage detection using SolarDiagnostics. More details can be found in Section 5.2.

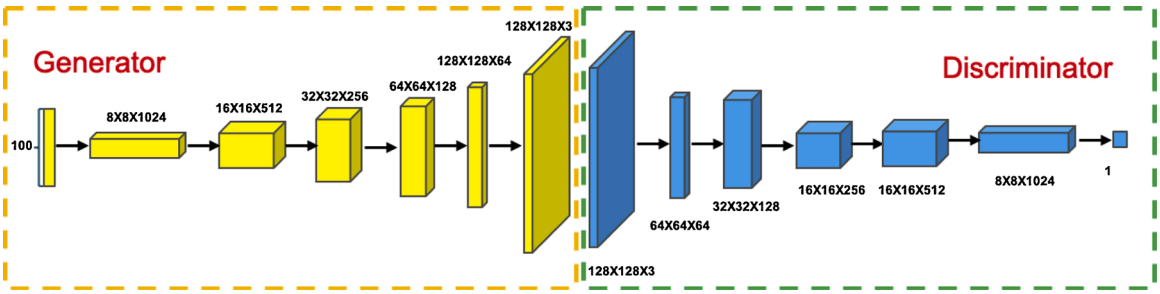


Fig. 1. The pipeline of our DCGANs system structure.

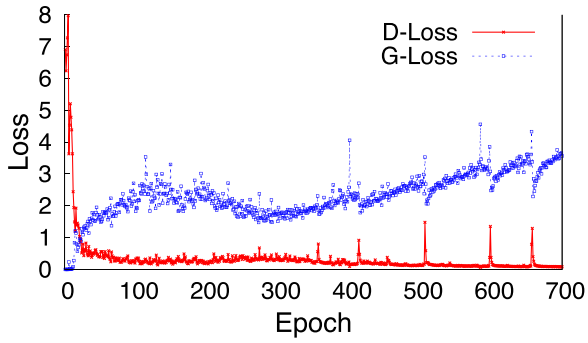


Fig. 2. The loss comparison results of discriminator and generator in our DCGANs.

4. Building large damaged image dataset

In this section, we describe how we address the imbalanced data challenge. We design a new approach that leverages deep convolutional generative adversarial networks (DCGANs) to generate a large and balanced solar PV array damage image dataset.

4.1. Image generator

To represent image data more effectively, we use generative adversarial networks (GANs) [32] architecture to build our damaged solar PV array image generator. However, recent work [33] has shown that GANs model has some performance limitations. For instance, GANs might be unstable to train and thus resulting in generators that produce nonsensical, noisy and incomprehensible new artificial images. The recent work [34] presented a new GANs—deep convolutional GANs (DCGANs) that has mitigated these issues by replacing the deterministic pooling function to strided convolution and using Rectified Linear Units (ReLU) activation in all generator layers, and leaky rectified linear unit (Leaky ReLU) activation in all discriminator layers. DCGANs has become the standard architecture to solve image generation problems. Fig. 1 shows the pipeline of our DCGANs network. Our DCGANs architecture is comprised of convolutional layers without max pooling or fully connected layers. We leverage convolutional stride and transposed convolution for downsampling and upsampling, respectively. The generator network uses a 100×1 noise vector. Our first layer is to project and

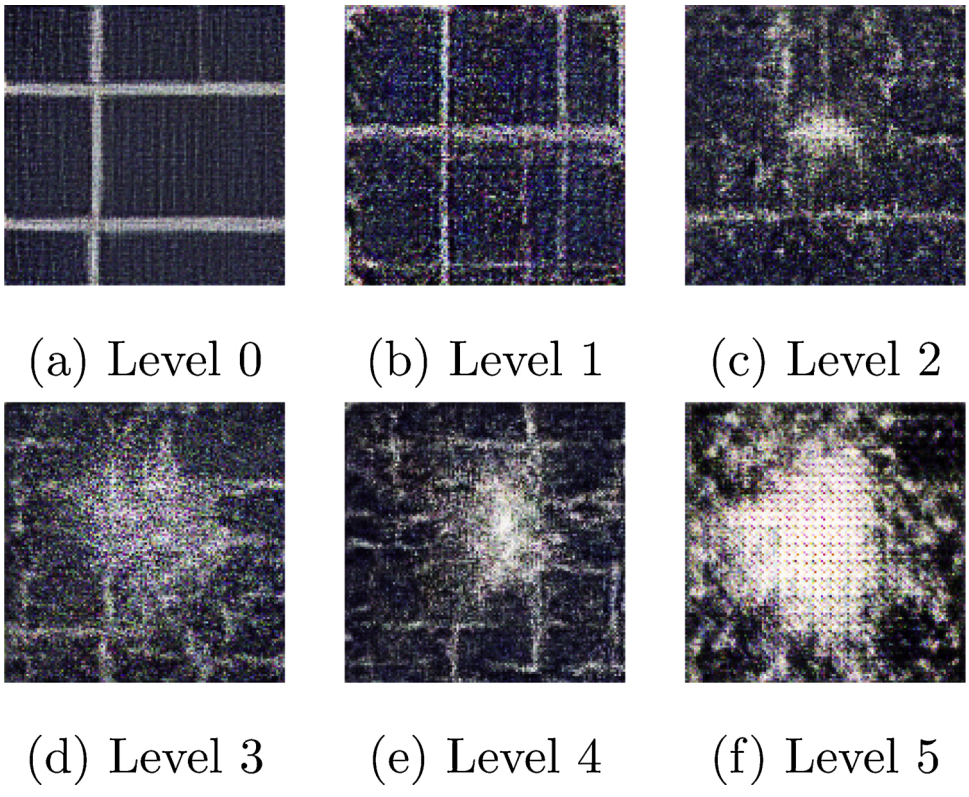


Fig. 3. The DCGANs generated rooftop solar PV array images that have six different damage levels.

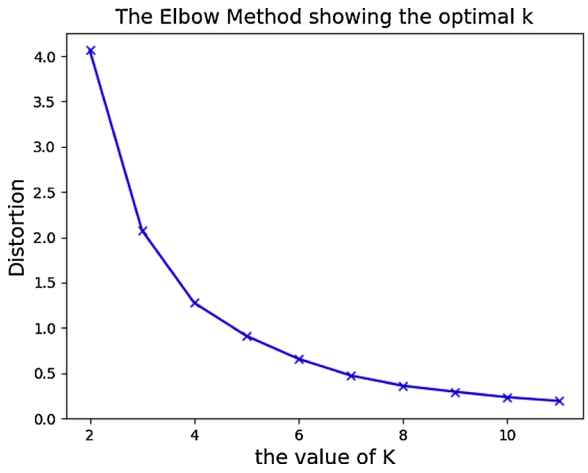


Fig. 4. The relationship between K and WCSS errors for K-Means clustering.

reshape inputs, following this layer, we have five convolutional layers. For generator model, we use the ReLU activation function for all the layers except the final one, where we employ the Tanh activation function. Generator and Discriminator have almost the same architectures, but reflected. For discriminator model, we use the Leaky ReLU activation function for all the layers except the last layer where we use the Sigmoid activation function.

4.2. Loss function

Given a DCGANs model that is comprised of one image generator and one image discriminator. Our goal is to train a generator that can generate image samples that others (e.g., discriminator) cannot distinguish them from the genuine source images. The generator—G is optimized to generate image samples that would be identified by the discriminator—D as the real image sample. Thus, the objective to design DCGANs model can be described as follows:

$$\max_{D}\{E_s(logD(s)) + E_n(log(1 - D(G(n))))\} \quad (1)$$

$$\min_{G}\{E_n(log(1 - D(G(n))))\} \quad (2)$$

where $D(s)$ denotes the discriminator output that s is a real image and $G(n)$ indicates the generated image sample using a noise—n. Note that, $E_s(logD(s)) = \sum_m^1 [logD(x^i)]$, where m denotes the number of the source image samples. As shown in Fig. 2, we find that our DCGANs model has achieved stable discriminator loss and generator loss after reaching at ~300 epochs.

4.3. Data augmentation

We first collect 350 damaged solar PV array images using Google Images API [18] and our prior work—SolarFinder [19]. We use the Google Image API to search the healthy and damaged rooftops, then we use thresholding approach to classify the images. However, it is rare to find damaged rooftop images and we can only find 350 images. In particular, this is also the motivation that we design a DCGANs-based model to generate the large solar PV array rooftop images. In addition, the actual images may not always look like these rooftop solar PV array images due to their different orientation, tilt, shading and other physical characteristics. To mimic these effects, we leverage multiple data argument techniques from TensorFlow [35], Open CV [14] and scikit-image [36], such as flip, rotation, crop, and translation operations.

We first apply K-Means clustering algorithm to characterize the collected 350 damage image dataset. The key problem is to determine the optimal cluster size—K. In particular, we leverage elbow method [37] to find the optimal—K. Fig. 4 shows the relationship between K and within-cluster sum of square (WCSS) errors for K-Means clustering. We find that when choosing $K = 6$, the K-Means algorithm yields at the minimum WCSS. Fig. 3shows six different solar PV array images that are generated by our DCGANs model and have different damage levels.

4.4. Summary

In this section, we present a deep convolutional generative adversarial networks (DCGANs)-based approach that leverages multiple image processing techniques to automatically generate a large amount (~60,000) of damaged solar PV array images. The generator is optimized to generate image samples which would be classified by the discriminator as belonging to the real data distribution. By doing this, we are able to build our deep learning-based damage detection system—SolarDiagnostics using a well-balanced training dataset.

5. Solar SolarDiagnostics design

In addressing the detection challenges, we design a new system—SolarDiagnostics that can automatically detect damage on a solar PV

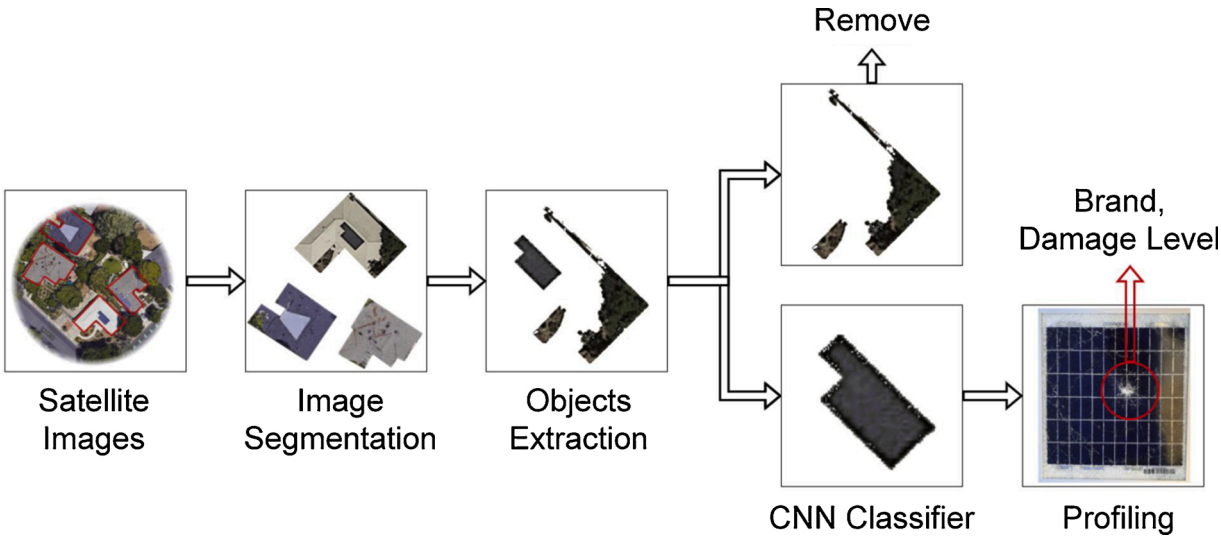


Fig. 5. The pipeline overview of SolarDiagnostics.

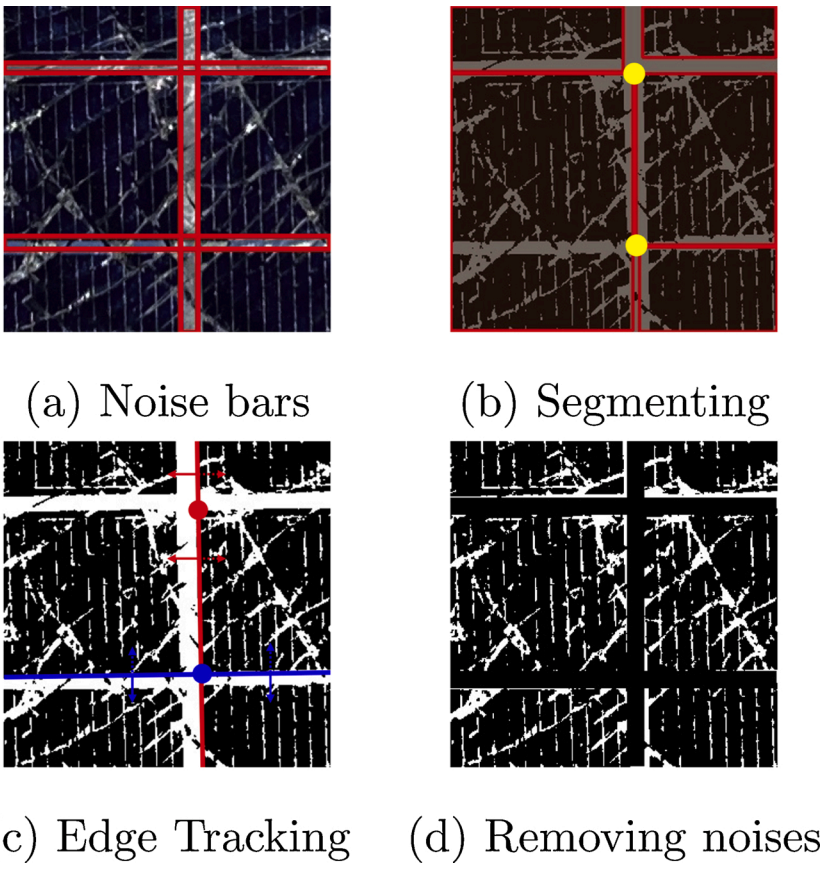


Fig. 6. Noise detection and removing.

array using only its rooftop image. In essence, SolarDiagnostics first integrates with our SolarFinder [19] to automatically segments rooftop images to solar PV array image contours and other rooftop object contours. Then, SolarDiagnostics leverages a CNNs-based deep learning classifier to detect any potential solar damage in each solar residing image contour. Finally, SolarDiagnostics applies solar damage level estimation, damage locator, and other characteristics estimators to further profile each damaged solar PV array. Fig. 5 shows the SolarDiagnostics's pipeline of the above operations.

5.1. Segmenting rooftop images

In addition to solar PV arrays, many other "outliers" objects such as ridge, structure, chimney, shade, and window may also present on solar PV array rooftops. Thus, after fetching the rooftop solar PV array images, SolarDiagnostics leverages an unsupervised multi-dimensional k-Means algorithm [38] to automatically segment each rooftop solar PV array image into a set of image contours C_i such that objects on the rooftop R_i are isolated. The goal of this segmentation is to ensure: given a rooftop R_i, we can assign each pixel based on its grayscale value into its best "fitting" cluster. We had found the optimal cluster size—$k = 5$ for a typical residential solar powered homes in our most recent work [19]. Eventually, SolarDiagnostics focuses on solar residing image contours only.

5.2. Preprocessing solar PV array images

Although SolarDiagnostics is only focusing on solar residing image

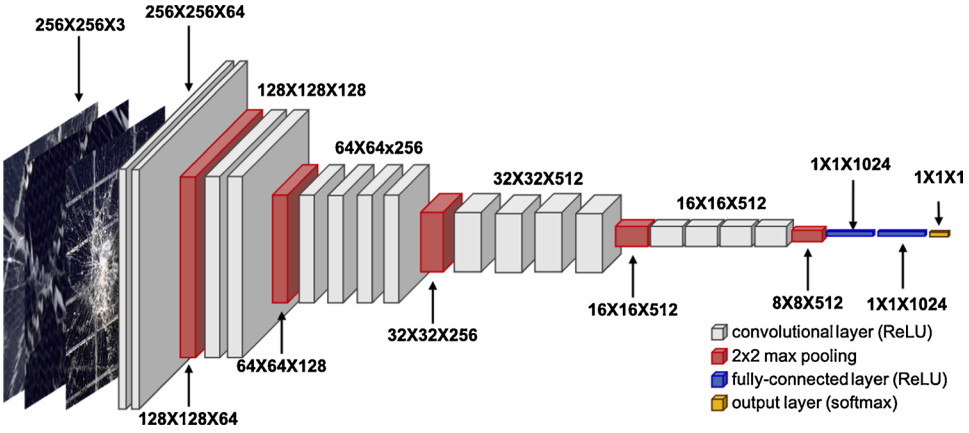
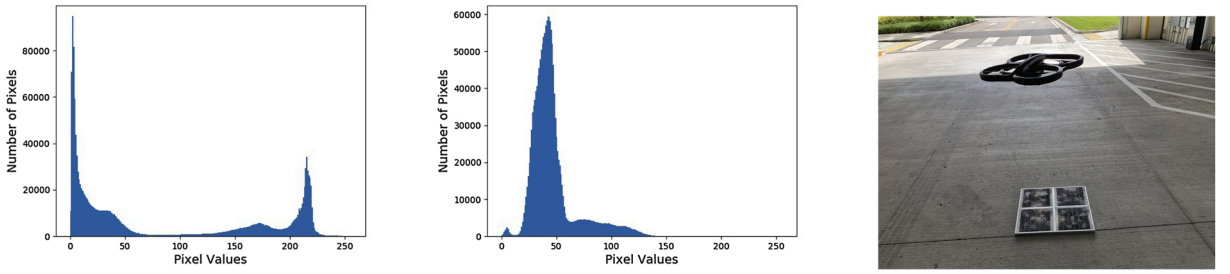


Fig. 7. The overview of our CNNs architecture design.



(a) ECO-WORTHY (b) Newpowa (c) SolarDiagnostics prototype

Fig. 8. The comparison of statistical learning results for 2 different brand manufactures' solar PV arrays using their grayscales.

contours, SolarDiagnostics may still see "outliers" in their image contours. Fig. 6(a) shows examples of these "outliers" that are mainly white lines such as gaps, bus bars, and fingers on solar PV arrays. These white lines always present while rectangle noise bars in their rooftop solar PV array images and their contours. Thus, as shown in Fig. 6(b), SolarDiagnostics leverages K-Means clustering approach to filter out those white rectangle contours as shown in Fig. 6(a). SolarDiagnostics identifies the vertex set for each rectangle contour, and then leverage noise reduction technique to track edge caused by hysteresis. As shown in Fig. 6(c), SolarDiagnostics tries to enlarge the rectangle "outlier" white bars at four different directions until no significant hysteresis is detected. After removing those True Negative "outliers", SolarDiagnostics is able to detect damage in solar PV array images shown in Fig. 6(d) using their actual greyscale value distributions.

5.3. Detecting damaged solar arrays

Next, we build a new convolutional neural networks (CNNs)-based deep learning classifier that can accurately identify damaged solar cells/regions in each solar residing image contour. Below, we describe the design of our CNNs architecture. As shown in Fig. 7, our CNNs architecture is comprised of input, convolutional layers (ReLU), max pooling, fully-connected layers (ReLU) and output. The inputs are 256×256 solar PV array images, and the first two layers are convolutional layers which have 256×256 neurons with a rectified linear unit (ReLU). Then, we have another two convolutional layers that have 128×128 neurons with ReLUs. After these layers, we employ another 4 convolutional layers with ReLUs, and all these layers have 32×32 neurons. Finally, we leverage 4 convolutional layers with ReLUs, and these layers all have 16×16 neurons. Among the different groups of convolutional layers, we have 2×2 max pooling which is used to down sample input solar PV array images and reduce its dimensionality. Two fully-connected layers with ReLUs have 1024 channels per each. The final layer in our design of CNNs architecture is the softmax layer that performs damage identification and thus contains 1 channel. Note that, this structure is our minimum recommendation to implement our SolarDiagnostics approach. More layers structure design may have better detection accuracy, however, the training time may be significantly increased.

5.4. Profiling damaged solar arrays

In addition to detecting damaged solar PV arrays, SolarDiagnostics can also profile each reported solar PV array. The profiling information for a damaged solar PV array may include damage level, damage location, brand detection and other information that is useful to assistant solar owners to better identify and make decisions to repair their rooftop solar PV systems. The proposed CNNs-based model in Section 5.3 is used to identity whether there is any damage on the given solar PV array. While, the following SVMs- and Random Forest-based reporting approaches are used to further profile the damage situation.

Reporting damage level. To classify the damage level for each damaged solar PV array, we leverage the supervised machine learning approach such as SVMs with linear kernel. The difference for SolarDiagnostics to report binary damage detection results and different level damage detection is the input features when training the deep learning and machine learning classifier. In the evaluation section, we evaluate the performance accuracy for these two different scenarios, respectively.

Reporting damage location. To localize damaged "portion" on solar PV arrays, we track the longitude and latitude for each image contour's vertex in C-language library—SQLite3 database that implements a small, fast, self-contained, high-reliability, full-featured, SQL database engine. Therefore, when SolarDiagnostics is reporting a damaged pixel in an image contour, we are able to track and cluster all the damage pixels/cells. We have included evaluation for this damage location reporting using Jaccard Similarity Index (JSI).

Reporting manufacture brand. To assist solar owners to repair their damaged solar PV arrays, the solar panel manufacture brand information might be necessary for the replacement and cost analytics purpose. In addition to report damaged solar PV arrays, SolarDiagnostics can also report their associated brand information simultaneously. As shown in Fig. 8(a) and (b), our insight is that different brand solar PV arrays have significant different patterns in their grayscale statistical learning results. SolarDiagnostics uses the pixel grayscale distribution features and leverages the Random Forest modeling to identify manufacture brand information for each damaged solar PV array. Note that, rather than other analytics, SolarDiagnostics leverages regular non-damaged regions on a damaged solar PV array to predict the manufacture brand information.

6. Implementation

We implement SolarDiagnostics in `python` using open-source frameworks, including Pandas [13], OpenCV [14], Scikit-learn [15] and PyCUDA [16,17]. SolarDiagnostics leverages Google Image API [18]. Our current implementation fetches damaged solar PV array images (800x800 pixels) from Internet. We use OpenCV, NumPy and Pandas for grayscale and RGB channel image data processing. We use the Scikit-learn [15] machine learning library in python to build our machine learning and deep learning approaches. The library supports multiple techniques including support vector machines (SVMs) with different kernel functions, multiple linear regression models, Random Forest, Decision Tree, KNN, and principal component analysis (PCA). For the CNNs-based SolarDiagnostics approach, we implement based on the framework from VGGnet [39], Scikit-learn [15], and OpenCV [14]. Finally, we schedule the batch jobs on our GPU servers to compare the MCC accuracy of 5 different approaches using CUDA. The server that we use to get all the benchmarking and evaluation results has resources as follows: (1) CPU: 2× Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10 GHz, (2) GPU: nVidia TITAN X (Pascal) (×8), (3) RAM: 128 GB, (4) OS: Linux CentOS 7.

In addition, we also build a SolarDiagnostics prototype. As shown in Fig. 8 (c), our prototype uses down facing camera on the drone to

capture images when flying over rooftops that have solar PV arrays installed. Our prototype has both of image output and video output. In particular, the video output can have HD 30 fps resolution which is sufficient for capturing ≥10 snapshots per rooftop. Our drone is flashed with Linux firmware which is open and easy for end users to control via mobile phone or Python scripts. The images and videos are synchronized via Wi-Fi to either Pi-3 based local SolarDiagnostics detecting system implementation or our SolarDiagnostics public APIs to be processed for identifying any potential damages.

7. Experimental evaluation

Below we describe our datasets, experimental setup, metrics used to evaluate our approaches, and evaluation results.

7.1. Datasets

Dataset 1. We use the large damaged solar PV array image dataset which is built in Section 4 and comprised of ∼60,000 rooftop solar PV array images with the resolution as 1024 × 1024. In this dataset, overall, we have 10,000 images per each damage level. In addition, we also include damage level, damage location, brand information, and other installation details for each rooftop image.

Dataset 2. We collect ∼500 publicly-available solar PV array rooftop images using Google Images API. The ratio of the damaged to non-damaged solar PV array images is 1:1. These images are indicating the actual damaged solar PV arrays from U.S. Given a solar-powered home listed in this dataset, we also prepare its groundtruth data, including the damage levels, damage locations, brand information, and other installation details for each solar PV array rooftop image.

Dataset 3. We also use our drone-based SolarDiagnostics prototype to test the performance of SolarDiagnostics at 10 "mock" rooftops. The dataset has 10 "mock" residential rooftop images which are taken by HD camera of our prototype. Note that, our approach may achieve better accuracy if the camera and memory storage support video recording, which significantly increases the amount of the damage images.

7.2. Experimental setup

To better understand the benefits of different damage detecting approaches, we implement and compare a group of "re-trained" and "pre-trained" solar PV array damage detection approaches, including the CNNs, SVMs (RBF), Random Forest, Decision Tree and KNN based approaches.

Re-trained approaches. In this case, all of solar PV array damage approaches can access to damaged solar PV array images from their testing sites. For CNNs approaches, we also fine-tune the VGGnet using the information from the testing sites. In doing so, we are bench-marking the best performance of different approaches.

Pre-trained approaches. In this case, all of solar PV array damage approaches cannot access to damaged solar PV array images from their testing sites. For CNNs approaches, we do not fine-tune the VGGnet using the information from the testing sites. In doing so, we are bench-marking the practical performance of different approaches.

7.3. Evaluating metrics

Below we describe the metrics that we use to evaluate SolarDiagnostics and other approaches.

Matthews correlation coefficient (MCC). To quantify the accuracy of different solar PV array damage(s) detection approaches, we use the Matthews correlation coefficient (MCC) [40], a standard measure of a binary classifier's performance, where values are in the range −1.0 to 1.0, with 1.0 being perfect solar PV array damage detection, 0.0 being random solar PV array damage prediction, and −1.0 indicating solar PV array damage detection is always wrong. The expression for computing MCC is below, where TP is the fraction of true positives, FP is the fraction of false positives, TN is the fraction of true negatives, and FN is the fraction of false negatives, such that TP+FP+TN+FN = 1.

$$\frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \tag{3}$$

Cohen's kappa. The Cohen's kappa [41] is a measure of the agreement between two raters who each classify N items into C mutually exclusive categories. The definition is as follows,

$$\kappa = 1 - \frac{1 - p_o}{1 - p_e} \tag{4}$$

where p_o is the relative observed agreement among raters, and p_e is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly seeing each category. If the classifiers are in complete agreement then κ should be close to 1. If there is no agreement among the classifiers other than what would be expected by chance, $\kappa = 0$.

Jaccard Similarity Index (JSI). To quantify the accuracy of SolarDiagnostics to predict the damage size of solar PV arrays, we use Jaccard Similarity Index (JSI) which is widely used in prior work to measure the similarly between detected damaged regions and groundtruth damaged regions. As a measure of similarity for the two sets of pixel data, with a range from 0% to 100%. The higher the percentage, the more precise predictions that SolarDiagnostics can do. It can be defined as follows,

$$\mathrm{JSI} = \frac{r_d \cap r_g}{r_d \cup r_g} \tag{5}$$

where r_d denotes the detected damage region for a solar PV array, and r_g indicates the groundtruth damage region for a solar PV array.

Table 1
The detection accuracy comparison of SolarDiagnostics when employing different classifiers.

	Model	TP	FN	TN	FP	MCC
Re-trained	CNNs	100%	0%	100%	0%	1
	SVMs-RBF	99.1%	0.9%	78.6%	21.4%	0.803
	Random Forest	98.0%	2.0%	80.7%	19.3%	0.807
	Decision Tree	90.7%	9.3%	86.2%	13.8%	0.772
	KNN	100%	0%	85.2%	14.8%	0.870
Pre-trained	CNNs	94.2%	5.8%	100%	0	0.947
	SVMs-RBF	0.2%	99.8%	29.8%	70.2%	−0.744
	Random Forest	1.5%	98.5%	27.1%	72.9%	−0.749
	Decision Tree	23.2%	76.8%	19.3%	80.7%	−0.574
	KNN	0.2%	99.8%	29.8%	70.2%	−0.695

7.4. Experimental results

7.4.1. Comparing re-trained approaches

We first compare SolarDiagnostics's performance with fully re-trained machine learning approaches that have complete access to the rooftop images from testing solar sites. In this case, the 5 approaches split the dataset into training and testing using a ratio as 7:3 after cross-validation. Unsurprisingly, as shown in Table 1, SolarDiagnostics (with CNNs classifier) yields the best MCC—1.0, and is the best performing and the most sophisticated solar PV arrays damage detection approach. In addition, SolarDiagnostics (with CNNs classifier) has False Negative (FN) as 0%. We can also see that the re-trained SVMs-RBF, Random Forest, Decision Tree and KNN approach yields a MCC of 0.80, 0.80, 0.77, and 0.87, respectively.

Results: *Comparing with the re-trained SVMs, Random Forest, Decision Tree and KNN approaches, re-trained SolarDiagnostics (with CNNs) yields the best MCC as 1.0 with False Negative as 0%, and thus is the best performing approach.*

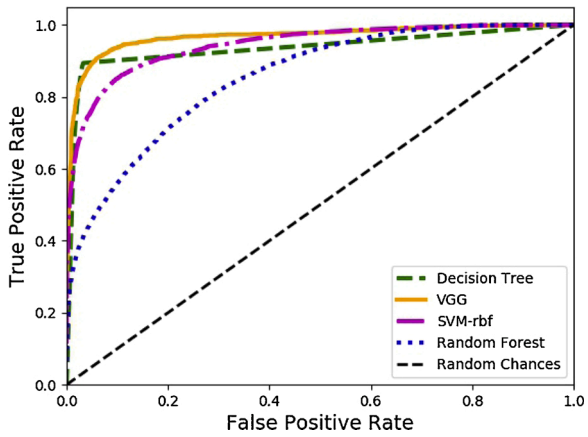


Fig. 9. The comparison of receiver operating characteristic (ROC) curves when applying different classifiers.

7.4.2. Comparing pre-trained approaches

We then compare the accuracy of the pre-trained ML approaches and SolarDiagnostics that do not have any access to the rooftop images from testing solar sites. In this case, the 5 approaches split the dataset into training dataset and testing dataset using a ratio of 7:3 without cross-validation between the two datasets. Note that, SolarDiagnostics (with CNNs) does not use any information from testing sites to fine-tune its CNNs model at this time. As shown in Table 1, SolarDiagnostics (with CNNs) yields the best MCC as 0.947 and all other 5 machine learning classifiers have reported negative MCC values. This simply indicates the machine learning approaches that leverage SVMs-RBF, Random Forest, Decision Tree, or KNN classifier are not able to reliably distinguish the damaged from the non-damaged solar PV arrays.

Results: *Comparing with the pre-trained machine learning approaches, SolarDiagnostics is the best performing approach and it yields the best MCC as 0.95, which is ~2 times better than decision tree based machine learning approach.*

7.4.3. Pre-trained vs re-trained approaches

The goal of this evaluation is to evaluate the ability of SolarDiagnostics to work in real practice. Note that the major difference between pre-trained and re-trained approaches is whether new solar PV array testing images are included in the training dataset of SolarDiagnostics or not. Re-trained approaches have all the access to the testing solar PV array images to calibrate their models. While, pre-trained approaches do not have any access to any testing solar PV array images, and thus is to identity the damage on solar PV array using their ready-to-use models. In real practice, SolarDiagnostics works in the pre-trained manner.

Table 1 shows that the MCC reported by the pre-trained SolarDiagnostics approach is significantly better than that of the re-trained machine learning (ML) approaches, including SVMs-RBF, Random Forest, Decision Tree, and KNN. In addition, the pre-trained SolarDiagnostics (CNNs) approach yields the MCC (~0.95) which is slight worse than that (~1.0) of the re-trained SolarDiagnostics (CNNs) approach. This is mainly due to the fact that the pre-trained CNNs approach cannot leverage any information from testing images to fine-tune its neural network. Among all the pre-trained approaches, pre-trained CNNs approach has minimum FN as only 5.8%. Note that "re-trained" results are provided as the upper bound reference of our SolarDiagnostics approach, in real practice, SolarDiagnostics works in the "pre-trained" mode.

Results: *Comparing with both of the re-trained and pre-trained approaches, SolarDiagnostics is the best and stable pre-trained performing approach and it yields the best MCC as 0.95, which is almost the same as re-trained SolarDiagnostics approach.*

Table 2
The comparison of SolarDiagnostics detection accuracy when employing pre-trained SVMs, Random Forest, Decision Tree, KNN, and CNNs to detect 6 different damage levels for solar PV arrays.

Model	Accuracy	F1	Cohen_Kappa	Precision	Recall	MCC
CNNs (VGG)	85.9%	0.859	0.830	0.872	0.859	0.832
SVMs (RBF)	74.0%	0.738	0.685	0.761	0.840	0.690
Random Forest	52.5%	0.456	0.424	0.559	0.525	0.451
Decision Tree	64.0%	0.641	0.568	0.643	0.640	0.568
KNN	85.2%	0.854	0.822	0.876	0.852	0.826

7.4.4. Quantifying SolarDiagnostics's accuracy

We then plot the receiver operating characteristic (ROC) curves for 5 different approaches. The goal of this examination is to evaluate the output quality for these 5 different approaches. ROC curves typically feature TP rate on the *Y*-axis, and FP rate on the *X*-axis. Thus, that says, the top left corner of the plot is the "ideal"—a false positive rate of zero, and a true positive charge of one. In addition, a larger area under the curve (AUC) is typically better. As shown in Fig. 9, for the re-trained comparison, our new approach—SolarDiagnostics stays at the top left corner and overlaps with the SVMs-RBF approach. In addition, the AUC under the SolarDiagnostics curve has the largest area. Therefore, among all the approaches examined in Fig. 9, SolarDiagnostics is the best binary classier when detecting damages on rooftop solar arrays.

Results: *SolarDiagnostics's ROC curve stays on the top of the left corner and has the largest AUC. Thus, comparing with other ML-based approaches, SolarDiagnostics is the best binary classifier for solar PV array damage detection.*

7.4.5. Profiling the damage of solar PV arrays

We examine the accuracy of SolarDiagnostics when predicting solar PV array damage using Dataset 2 as discussed previously in Section 7.1. SolarDiagnostics first fetches the 500 homes rooftop images and then segments them into contours. SolarDiagnostics then apply the unsupervised hybrid approach over those contours to identify solar panels and learning the physical characteristics, e.g., size, orientation, and shade.

Identifying damage level. Rather than comparing binary detection results for the machine learning approaches and SolarDiagnostics approach, we also compare the performance accuracy of all approaches when they reporting damage levels. We employ the metric—MCC to report the accuracy when SolarDiagnostics identifying the 6 different damage levels. As discussed in Section 7, to report the damage level of solar PV array, SolarDiagnostics first examines the pixels that are identified as damage in each solar PV array contours Then, SolarDiagnostics performs a union operation to add up all the contours for the same rooftop to report the damage level. As shown in Table 2, SolarDiagnostics also yields the best MCC as 0.83, the best Cohen_Kappa as 0.83, and the best F1 as 0.86. Note that, to report the results in Table 2, we re-sampled the damaged solar PV array dataset such that each damage level has 200 images.

Results: *Comparing with the pre-trained approaches, SolarDiagnostics yields the best MCC as 0.83, the best Cohen-Kappa as 0.83, and the best F1 as 0.86. Thus, SolarDiagnostics is the best performing pre-trained approach when reporting 6 different damage levels.*

Localizing detected solar PV arrays. We employ the metric—Jaccard Similarity Index (JSI) to report the accuracy. To report the location of solar PV array damage, SolarDiagnostics first examines the pixels that are identified as solar PV array contours. Then, SolarDiagnostics performs a union operation to add up all the contours for the same rooftop to report the damage size. We find that SolarDiagnostics is able to report a JSI as 78.52% when averaging on the results of Dataset

Table 3
The brand detection accuracy comparison of SolarDiagnostics on solar PV array arrays with 6 different damage levels.

Damage level	F1	MCC
Level 0	0.859	0.832
Level 1	0.840	0.690
Level 2	0.525	0.451
Level 3	0.640	0.568
Level 4	0.852	0.826
Level 5	0.852	0.826

#3.

Detecting brands of solar PV arrays. We leverage statistical learning methods (e.g., KNN, SVMs) to identify the manufacture brand for each reported damaged solar PV array. As shown in Fig. 8, the two different brand solar PV arrays has significant different pattern in their grayscale distribution. Using Dataset 3, we find that SolarDiagnostics can report the solar PV array's manufacture brand within an accuracy—MCC of 1.0. Note that, the prior data analytics-based approaches do not provide or discuss the damage localization. The goal of this brand profiling is to assist solar owners to better estimate the cost to repair damages on their rooftop solar PV arrays. In addition, we also exam SolarDiagnostics' brand detection accuracy on different damage level images. As shown in Table 3, SolarDiagnostics yields the MCC of brand detection in the range of 0.69-0.83 over the solar PV array images from different damage levels.

Results: *In addition to accurately detect damage on solar PV arrays, SolarDiagnostics is able to report accurate profiling information, e.g., damage location, and manufacture brand, simultaneously*.

Note that, SolarDiagnostics's approach to profiling damaged solar PV arrays is orthogonal to the other aspects of the techniques and is "pluggable," such that we could use other new machine learning approaches to better estimate the profiling information.

8. Conclusion and future work

We design a new defense system—SolarDiagnostics that can automatically detect and localize damage on rooftop solar PV arrays using only their rooftop images. In essence, SolarDiagnostics first leverages an unsupervised segmentation algorithm to isolate the objects on rooftops to extract solar panel residing contours. Then, SolarDiagnostics employs a deep convolutional neural networks (CNNs) to accurately identify and characterize any damage that may exist in each image contour. We evaluate SolarDiagnostics using 60,360 damaged solar array images generated and by building a prototype. We found that SolarDiagnostics is able to yield an MCC as of 1.0 when detecting damage on solar PV arrays. In addition, pre-trained SolarDiagnostics yields a MCC of 0.95, which is significantly better than the re-trained ML approaches and is the same as the re-trained SolarDiagnostics. Thus, Solar- Diagnostics achieves similar accuracy without access to any training data from testing solar sites as a fully re-trained approach with complete access to such training image data. We plan to implement the optimization of SolarDiagnostics profiling module to report more accurate solar PV array damage estimations. We also plan to learn the performance accuracy of SolarDiagnostics using different type of images (e.g., Tesla roof shingles). In addition, we are also planning to host a SolarDiagnostics API server such that users can directly use our SolarDiagnostics damage detecting service directly via their remote API calls.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Partially supported by Cyber Florida Collaborative Seed Program and National Security Agency's National Centers of Academic Excellence in CybersecurityProgram.

References

[1] U.S. Energy Information Administration, Electricity Monthly Update, 2016. https://www.seia.org/solar-industry-research-data.
[2] How Much Does It Cost to Clean and Maintain Solar Panels?, 2020. https://www.seia.org/solar-industry-research-data.
[3] A. Mellit, G.M. Tina, S.A. Kalogirou, Fault detection and diagnosis methods for photovoltaic systems: a review, Renew. Sustain. Energy Rev. 91 (2018) 1–17.
[4] S. Sarikh, M. Raoufi, A. Bennouna, A. Benlarabi, B. Ikken, Photovoltaic system fault identification methodology based on iv characteristics analysis, AIP Conference Proceedings, vol. 2123 (2019) 020037.
[5] M. Dhimish, V. Holmes, B. Mehrdadi, M. Dales, The impact of cracks on photovoltaic power performance, J. Sci.: Adv. Mater. Dev. 2 (2) (2017) 199–209.
[6] S. Rao, S. Katoch, P. Turaga, A. Spanias, C. Tepedelenlioglu, R. Ayyanar, H. Braun, J. Lee, U. Shanthamallu, M. Banavar, et al., A cyber-physical system approach for photovoltaic array monitoring and control, 2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA) (2017) 1–6.
[7] S. Iyengar, S. Lee, D. Sheldon, P. Shenoy, Solarclique: detecting anomalies in residential solar arrays, Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies (2018) 1–10.
[8] M. Gan, C. Wang, et al., Fault feature enhancement for rotating machinery based on quality factor analysis and manifold learning, J. Intell. Manuf. 29 (2) (2018) 463–480.
[9] A. Livera, M. Theristis, G. Makrides, G.E. Georghiou, Recent advances in failure diagnosis techniques based on performance data analysis for grid-connected photovoltaic systems, Renew. Energy 133 (2019) 126–143.
[10] X. Zhao, J. Liang, F. Cao, A simple and effective outlier detection algorithm for categorical data, Int. J. Mach. Learn. Cybern. 5 (3) (2014) 469–477.
[11] R. Hariharan, M. Chakkarapani, G. Saravana Ilango, C. Nagamani, A method to detect photovoltaic array faults and partial shading in pv systems, IEEE J. Photovolt. 6 (5) (2016) 1278–1285.
[12] R. Platon, J. Martel, N. Woodruff, T.Y. Chau, Online fault detection in pv systems, IEEE Trans. Sustain. Energy 6 (4) (2015) 1200–1207.
[13] Pandas. https://pandas.pydata.org/.
[14] OpenCV. https://opencv.org/.
[15] Scikit-Learn Machine Learning in Python. https://scikit-learn.org/stable/.
[16] PyCUDA. https://mathema.tician.de/software/pycuda/.
[17] An Even Easier Introduction to Cuda. https://devblogs.nvidia.com/even-easier-introduction-cuda/.
[18] Google Images API: Serpapi. https://serpapi.com/images-results.
[19] Q. Li, Y. Feng, Y. Leng, D. Chen, Solarfinder: automatic detection of solar photovoltaic arrays, Proceedings of the 19th ACM/IEEE International Conference on Information Processing in Sensor Networks (2020) 100–111.
[20] Solardiagnostics. https://github.com/cyber-physical-systems/SolarDiagnostics.
[21] M. Davarifar, A. Rabhi, A. El-Hajjaji, M. Dahmane, Real-time model base fault diagnosis of pv panels using statistical signal processing, 2013 International Conference on Renewable Energy Research and Applications (ICRERA) (2013) 599–604.
[22] M. Banavar, H. Braun, S.T. Buddha, V. Krishnan, A. Spanias, S. Takada, T. Takehara, C. Tepedelenlioglu, T. Yeider, Signal processing for solar array monitoring, fault detection, and optimization, Synthesis Lectures on Power Electronics 7 (1) (2012) 1–95.
[23] D. Stellbogen, Use of pv circuit simulation for fault detection in pv array fields, Conference Record of the Twenty Third IEEE Photovoltaic Specialists Conference-1993 (Cat. No. 93CH3283-9) (1993) 1302–1307.
[24] E.A. Kawam, E.A. Kawam, Photvoltaic solar array health monitor, US Patent A 12/156,935 (2008).
[25] S. Fadhel, C. Delpha, D. Diallo, I. Bahri, A. Migan, M. Trabelsi, M. Mimouni, Pv shading fault detection and classification based on iv curve using principal component analysis: application to isolated pv system, Solar Energy 179 (2019) 1–10.
[26] M.R. Maghami, H. Hizam, C. Gomes, M.A. Radzi, M.I. Rezadad, S. Hajighorbani, Power loss due to soiling on solar panel: a review, Renew. Sustain. Energy Rev. 59 (2016) 1307–1316.
[27] S. Rao, A. Spanias, C. Tepedelenlioglu, Solar array fault detection using neural networks, in: 2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS), IEEE, 2019, pp. 196–200.
[28] Y. Zhao, L. Yang, B. Lehman, J.-F. de Palma, J. Mosesian, R. Lyons, Decision tree-based fault detection and classification in solar photovoltaic arrays, in: 2012 Twenty-Seventh Annual IEEE Applied Power Electronics Conference and Exposition (APEC), IEEE, 2012, pp. 93–99.
[29] X. Xu, H. Wang, Y. Zuo, Method for diagnosing photovoltaic array fault in solar photovoltaic system, in: 2011 Asia-Pacific Power and Energy Engineering Conference, IEEE, 2011, pp. 1–5.
[30] K.A. Kim, G.-S. Seo, B.-H. Cho, P.T. Krein, Photovoltaic hot-spot detection for solar panel substrings using ac parameter characterization, IEEE Trans. Power Electron. 31 (2) (2015) 1121–1130.

[31] Y. Zhao, B. Lehman, R. Ball, J. Mosesian, J.-F. de Palma, Outlier detection rules for fault detection in solar photovoltaic arrays, in: 2013 Twenty-Eighth Annual IEEE Applied Power Electronics Conference and Exposition (APEC), IEEE, 2013, pp. 2913–2920.
[32] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Advances in Neural Information Processing Systems (2014) 2672–2680.
[33] Y. Wang, C. Wu, L. Herranz, J. van de Weijer, A. Gonzalez-Garcia, B. Raducanu, Transferring gans: generating images from limited data, Proceedings of the European Conference on Computer Vision (ECCV) (2018) 218–234.
[34] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, 2015. arXiv:1511.06434.
[35] TensorFlow. https://www.tensorflow.org/.
[36] scikit-image, 2020. https://scikit-image.org/.
[37] Elbow Method. https://www.scikit-yb.org/en/latest/api/cluster/elbow.html.
[38] K-means. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html/.
[39] Very Deep Convolutional Networks for Large-scale Visual Recognition. https://www.robots.ox.ac.uk/vgg/research/very_deep/.
[40] S. Boughorbel, F. Jarray, M. El-Anbari, Optimal classifier for imbalanced data using Matthews correlation coefficient metric, PLOS ONE 12 (6) (2017) e0177678–e0177678.
[41] M.L. McHugh, Interrater reliability: the kappa statistic, Biochem. Med.: Biochem. Med. 22 (3) (2012) 276–282.