Open Data Sets for Assessing Photovoltaic Reliability

Xin Chen *†, Baojie Li *, Jennifer L. Braid §, Brandon Byford §, Dylan J. Colvin ¶, Andrew Glaws ‡, Norman Jost §, Benjamin Pierce §‖, Salil Rabade ‡, Martin Springer ‡, Anubhav Jain *

*Lawrence Berkeley National Laboratory, Berkeley, California, U.S.
†University of California, Berkeley, Berkeley, California, U.S.
‡National Renewable Energy Laboratory, Golden, Colorado, U.S.
§Sandia National Laboratories, Albuquerque, New Mexico, U.S.
¶University of Central Florida, Florida Solar Energy Center, Cocoa, Florida, U.S.
‖Case Western Reserve University, Cleveland, Ohio, U.S.

Abstract

Photovoltaic (PV) systems have become a cornerstone of renewable energy strategies, particularly due to the significant reduction in solar power costs over the past decade. However, the long-term reliability of PV installations presents a persistent challenge, requiring the development of advanced monitoring and predictive maintenance strategies. A wide range of data types are used to evaluate the health of PV systems, including environmental conditions, electrical performance, and inspection imagery. These data enable methodologies such as machine learning (ML) models for lifetime prediction and computer vision techniques for defect detection. However, the acquisition of high-quality and comprehensive data is difficult, particularly in terms of long-term consistency and data variety. Publicly available data sets serve as valuable resources for addressing these challenges, but they often suffer from fragmentation and are difficult to access. This paper presents a comprehensive review of existing open-source data sets related to PV degradation, analyzing their features, functionalities, and potential applications. We categorize these data sets based on the specific aspects of PV system information they cover, such as environmental conditions, operational monitoring, and material inspection, and propose relevant tools and ML models for processing them. In addition, we propose practices for future data collection and usage, while also discussing potential directions in data-driven research. Our aim is to enhance data utilization and publication among researchers and industry professionals, promoting a deeper understanding of the role of data in enhancing the performance and durability of PV systems.

Keywords

Photovoltaic degradation, Solar module durability, Open-source data set, Machine learning

The project was primarily funded and intellectually led as part of the Durable Modules Consortium (DuraMAT), an Energy Materials Network Consortium funded under Agreement 32509 and 38259 by the U.S. Department of Energy (DOE), Office of Energy Efficiency & Renewable Energy, Solar Energy Technologies Office (EERE, SETO). Lawrence Berkeley National Laboratory is funded by the DOE under award DE-AC02-05CH11231. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This work is authored in part by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308.
The authors declare no conflicts of interest. Xin and Baojie co-organize the review work. All authors contribute to the writing of this paper. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. government. Instruments and materials are identified in this paper to describe the experiments. In no case does such identification imply recommendation or endorsement by LBL.

Nomenclature

ANN	Artificial neural network
API	Application programming interface
BOKU	University of Natural Resources and Life Sciences, Vienna
BOM	Bill of material
c-Si	Crystalline silicon
CCD	Charge-coupled device
CdTe	Cadmium telluride
CEC	California Energy Commission
CMOS	Complementary metal-oxide-semiconductor
CNN	Convolutional neural network
CWRU	Case Western Reserve University
DHI	Diffuse horizontal irradiance
DIAS	Data Integration and Analysis System
DMX	Dynamic mechanical acceleration
DNI	Direct normal irradiance
ECMWF	European Centre for Medium-Range Weather Forecasts
EDA	Ensemble of Data Assimilations
EL	Electroluminescence
EPRI	Electric Power Research Institute
ETL	Electron transport layer
EUMETSAT	European Organisation for the Exploitation of Meteorological Satellites
FARMS	Fast all-sky radiation model for solar applications
FAU	University of Erlangen–Nuremberg
FSEC	Florida Solar Energy Center
GAN	Generative adversarial network
GES DISC	Goddard Earth Sciences Data and Information Services Center
GHCN	Global Historical Climate Network
GHI	Global horizontal irradiance
I-V	Current-voltage
IFS	Integrated Forecast System
INESC TEC	The Institute for Systems and Computer Engineering, Technology and Science
IR	Infrared
JRA-55	Japanese 55-year Reanalysis
KIT	Karlsruhe Institute of Technology
LBNL	Lawrence Berkeley National Laboratory
LCOE	Levelized cost of electricity
LLM	Large language model
LSPV	Large-scale solar photovoltaic
LSTM	Long short-term memory
MERRA-2	Modern Era-Retrospective Analysis
ML	Machine learning
MPP	Maximum power point
NASA	National Aeronautics and Space Administration
NDA	Nondisclosure agreement
NIST	National Institute of Standards and Technology
NOAA	National Oceanic and Atmospheric Administration
NMOT	Nominal module operating temperature
NREL	National Renewable Energy Laboratory
NSRDB	National Solar Radiation Database
OEDI	Open Energy Data Initiative
O&M	Operations and maintenance
PCE	Power conversion efficiency
PL	Photoluminescence
POA	Plane-of-array
PREDICTS2	Physics of Reliability: Evaluating Design Insights for Component Technologies in Solar 2
PRISM	Parameter-elevation Regressions on Independent Slopes Model
PSL	Physical Sciences Laboratory
PV	Photovoltaic
PVDAQ	Photovoltaic Data Acquisition
PVDeg	PV Degradation Tools
PVEL	PV Evolution Labs
PVRDB	Photovoltaic Rooftop Database
PVRDB-PR	PV Rooftop Database for Puerto Rico
PVROM	PV Reliability Operations and Maintenance
QMUL	Queen Mary University of London
Qual Plus	Qualification Plus
REMPD	Renewable Energy Materials Properties Database
RAG	Retrieval-augmented generation
RNN	Recurrent neural network
RTC	Regional Test Center
SAM	System Advisor Model
SARAH-3	Surface Solar Radiation Data Set – Heliosat
SNL	Sandia National Laboratories
STC	Standard test conditions
TMY	Typical meteorological year
UAV	Unmanned aerial vehicles
UCF	University of Central Florida
UCSD	UC San Diego
UPV	Utility PV
USGS	United States Geological Survey
USPVDB	United States Large-Scale Solar Photovoltaic Database
UVF	Ultraviolet Fluorescence
VAE	Variational autoencoder
Wits	The University of the Witwatersrand, Johannesburg

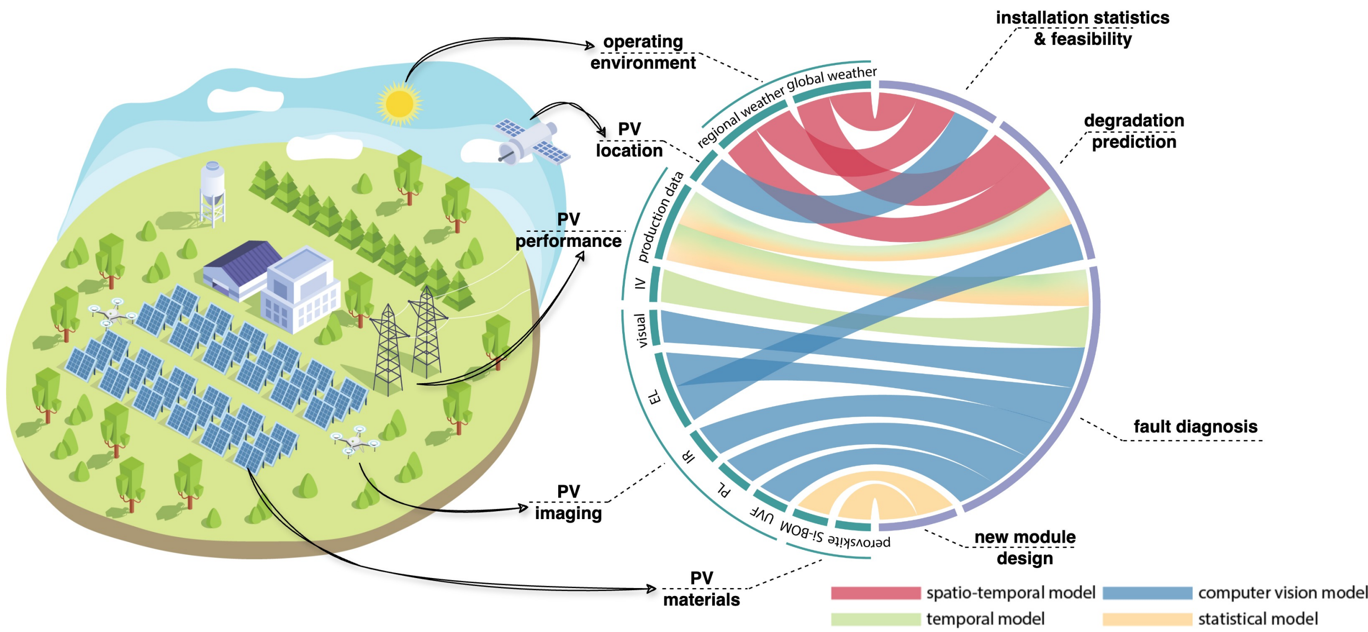

Fig. 1. An overview of data types, project objectives, and data-driven methodologies. The left-hand diagram provides a schematic representation of data sources pertinent to solar module degradation. The right-hand chord diagram elucidates the interconnections between data types and their corresponding project objectives. The colors of connection differentiate the various prevalent methods employed in data processing.

I. Introduction

Solar energy has emerged as a critical renewable energy source, making substantial contributions to global initiatives aimed at reducing carbon emissions and promoting sustainable development [1]. In recent years, the global installation of photovoltaic (PV) systems has experienced rapid growth, with the global weighted average levelized cost of electricity (LCOE) for utility-scale solar power decreasing from 0.445 USD / kWh in 2010 to 0.049 USD / kWh in 2022, which represents a reduction of approximately 29% compared to the most economical fossil fuel-fired alternatives [2]. As PV deployment continues to expand, ensuring long-term durability and performance becomes critical to optimizing its economic and environmental advantages. However, demanding operational environments and gradual material degradation present significant challenges in extending the operational lifespan of solar modules, thereby necessitating rigorous monitoring and predictive maintenance strategies. These tasks include root cause analysis using production data, fast inspection of solar farms, and prediction of PV system lifetime. The effective implementation of these tasks relies heavily on the availability of high-quality data, which allows detailed analysis and the application of data-driven methods such as machine learning (ML) modeling to enhance PV system reliability.

Various categories of data are utilized in PV health monitoring and analysis, each serving distinct purposes and requiring specific data-driven methods. For instance, performance data from PV systems can be used in statistical models such as physics-based regression models to predict future performance trends [3]. In addition, high-resolution images can facilitate defect detection using computer vision techniques [4], [5], [6], [7], [8]. Figure 1 illustrates the relationship between different data sets, the corresponding data-driven methods, and the tasks they support.

Despite the potential of data-driven methods, collecting comprehensive and high-quality PV degradation data poses several challenges. Long-term data collection is particularly difficult due to the need for consistent monitoring over extended periods to accurately capture degradation patterns. In addition, obtaining advanced imaging data and ensuring a high data variance that covers different module types further complicate the process.

The sheer volume of data required to develop reliable models also presents challenges in storage and processing. In this context, existing public data sets become invaluable resources. They not only provide benchmarks for validating new models but also ease the burden of data collection for researchers. However, these data sets often suffer from issues such as being unorganized and difficult to locate, which can hinder their effective utilization.

Previous papers [9], [10], [11], [12], [13], [14], [15], [16] have provided valuable reviews of the application of data-driven methods in solar module degradation. However, these publications, such as [16], [10], [14], focus mainly on the review of methodologies rather than the aspect of data. In addition, most of the existing work focuses on specific aspects of data processing rather than encompassing all types. For instance, papers [12], [13] primarily review state-of-the-art ML techniques in module power prediction, while papers [11], [14] focus on ML techniques used for fault detection. The paper [9] mentions some open-source data sets, but covers only a small part of the data types in solar module degradation. To the best of our knowledge, a thorough review of open-source data sets in this field has not yet been conducted.

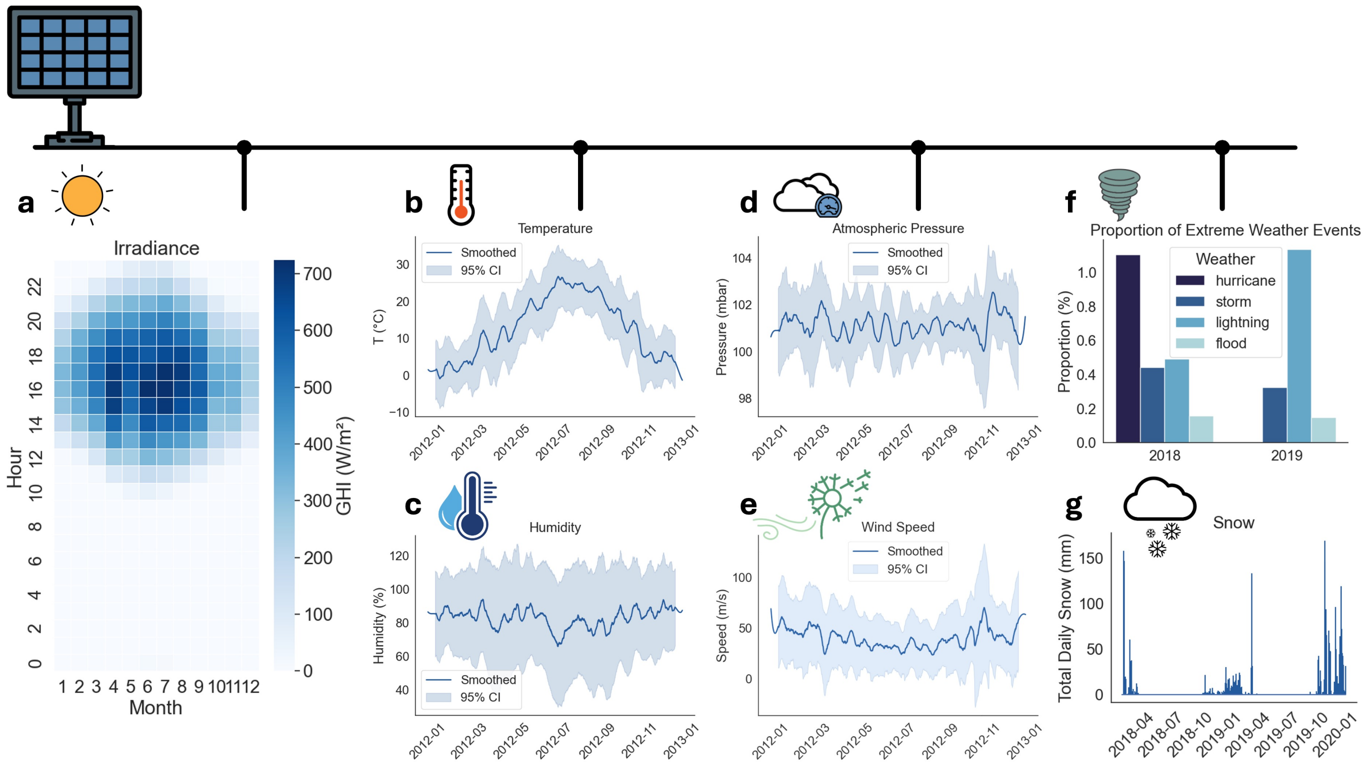

Fig. 2. Visualization of common types of environmental data, including (a) irradiance, (b) ambient temperature, (c) humidity, (d) atmospheric pressure, (e) wind speed, (f) records of extreme weathers and (g) depth of snow. The extreme weather and snow data are from [22] and other data are from [23].

Our paper aims to bridge this gap by providing a comprehensive review of available data sets related to PV performance and degradation, along with insights into future data-driven projects using these data sets. We summarize the data sets by highlighting their key features and functionalities, covering a wide scope of topics in PV degradation such as environmental conditions, PV installations, current-voltage monitoring, image inspection, and materials. We also suggest corresponding tools or ML models that can be used to process these data sets. Since this paper focuses primarily on reviewing the data rather than methodologies, we present only some typical tools or models. In doing so, we offer a solution to current findability issues in existing data sets, making it easier for researchers to access and use these resources. In addition to reviewing availability and quality of existing data sets, we discuss potential research directions, suggest best practices for data collection, publication, and utilization, and highlight innovative ML techniques that can be applied to PV degradation data in the future. We hope that this paper can inspire and guide researchers in developing more effective and efficient data-driven methods for monitoring and improving the durability of the PV system.

II. Operation environment

This section reviews data sets and processing tools related to the operational environment of a PV system. The data sets examined here include spatial and temporal information on various environmental factors that influence PV performance and degradation.

The operational environment where PV systems are deployed is a key driver of material degradation and damage, in addition to determining the power output of the array. Figure 2 provides a visual representation of some environmental data. Solar irradiance, measured by direct normal irradiance (DNI), diffuse horizontal irradiance (DHI), or global horizontal irradiance (GHI), is a primary factor in PV system degradation [17]. While the intensity and variability of solar irradiance are major contributors to degradation, other environmental factors such as temperature, humidity, and wind speed and direction also play significant roles [18]. For example, studies have shown statistical correlations between degradation pathways and environmental factors such as air temperature and humidity in different geographical locations [19]. Similarly, high or turbulent winds can impose significant mechanical stress on PV panels, potentially leading to aeroelastic flutter, which accelerates damage and power loss [20].

Traditional physics-based models used to assess the impact of environmental conditions on PV degradation have often relied on limited or sparse data sets [21]. In contrast, ML models typically require large and comprehensive data sets for more accurate model fitting and prediction. Therefore, there is a pressing need to expand the availability and accessibility of comprehensive environmental data.

Platform	data set	REF|URL	Spatial Coverage	Temporal Coverage	Spatial Resolution	Temporal Resolution	Remark
NREL	NSRDB	[23]|[25]	Global	1998-Present	2 km-10 km	5 min-1 hr	API and data viewer
ECMWF	ERA5	[26]|[27]	Global	1940-Present	$0.25° \times 0.25°$	1 hr	API access
NASA	MERRA-2	[28]|[29]	Global	1980-Present	$0.5° \times 0.625°$	1 hr	Access from webpage
NOAA	NCEP/DOE Reanalysis II	[30]|[31]	Global	1979-Present	$2.5° \times 2.5°$	6 hr-1 month	Website access and FTP
JMA	JRA-55	[32]|[33]	Global	1958-Present	$1.25° \times 1.25°$	3 hr	Website access
EUMETSAT	SARAH-3	[34]|[35]	Global	1983-Present	$0.05° \times 0.05°$	30 min	API access

TABLE I
OVERVIEW OF NOTABLE GLOBAL METEOROLOGICAL DATA SETS.

A. Global weather data

The summary of open-source data sets is presented in Table I. We first examine open data sets related to environmental and meteorological information. Meteorological reanalysis, one of the primary approaches for constructing these data sets, integrates sparse observational data with numerical weather prediction models to generate gridded data that typically spans multiple years and covers nearly the entire globe [24]. This type of comprehensive data set is valuable for long-range or geographically diverse studies. They are often used to generate typical meteorological year (TMY) data sets that provide a statistically representative (*e.g.*, median) time series of meteorological data for specific locations or regions. In this section, we explore several global data sets containing extensive spatio-temporal solar irradiance and other meteorological data.

1) The National Solar Radiation Database: The National Solar Radiation Database (NSRDB) is a global spatio-temporal database that includes solar radiation values along with various meteorological parameters, such as temperature, relative humidity, atmospheric pressure, and wind speed and direction [23]. The NSRDB contains several data sets covering different regions and time periods, with spatial resolutions ranging from 2 km to 10 km and temporal resolutions from 5 minutes to 1 hour. While NSRDB provides global data, its highest-resolution and longest-running data focus primarily on the continental United States and nearby regions. Users can access the data interactively through the NSRDB Viewer or programmatically via an application programming interface (API), both of which are available on the NSRDB website [25], as well as through the US Department of Energy's Open Energy Data Initiative (OEDI) [36].

The current version of NSRDB integrates multiple data sources with the Physical Solar Model developed by the National Renewable Energy Laboratory (NREL). This model first generates cloud and aerosol data [37], which are then used to produce solar irradiance outputs through the fast all-sky radiation model for solar applications (FARMS) [38]. Additional variables needed for running models such as REST2 and FARMS (*e.g.*, aerosol optical depth, precipitable water vapor, and albedo) are derived from NASA's Modern-Era Retrospective Analysis (MERRA-2) data set [28].

2) ERA5: The European Centre for Medium-Range Weather Forecasts (ECMWF) has released a series of reanalysis data sets, with its most recent iteration being ERA5 [26]. ERA5 is generated using ECMWF's Integrated Forecast System (IFS) model Cy41r2 [39], offering high-resolution, ensemble meteorological forecasts. The IFS model is enhanced through a data assimilation approach using a hybrid incremental 4D-Var system [40]. ERA5 provides global data on solar irradiance, temperature, humidity, wind speed, precipitation, and albedo, covering the period from 1940 to the present. The data are available at a spatial resolution of 31 km and an hourly temporal resolution. Additionally, ERA5 includes the Ensemble of Data Assimilations (EDA) [41], a ten-member ensemble data set at a spatial resolution of 63 km and a time resolution of three hours, which can be used for uncertainty quantification and statistical analyses. ERA5 data can be downloaded from the Climate Data Store [27], managed by the Copernicus program [42], and can also be accessed programmatically through an API or through the ERA5 Explorer tool.

3) Other global data sets: Beyond NSRDB and ERA5, several other global data sets cover different regions, time periods, or meteorological quantities. The National Aeronautics and Space Administration (NASA) maintains the Modern Era-Retrospective Analysis (MERRA-2) data set [28], which emphasizes advancements in aerosol observation and stratospheric and cryospheric processes. MERRA-2 data are utilized by NSRDB for cloud formation modeling and can be accessed directly from NASA's Goddard Earth Sciences Data and Information Services Center (GES DISC).

The National Oceanic and Atmospheric Administration (NOAA) provides the NCEP/DOE Reanalysis II data set, covering over 40 meteorological and environmental variables at various atmospheric pressure levels from 1979 to the present [30]. This data is available through the NOAA Physical Sciences Laboratory (PSL) website [31]. Additionally, the Japanese 55-year Reanalysis (JRA-55), another NOAA project, covers the period from 1958 to the present and is accessible through the Data Integration & Analysis System (DIAS) [32].

The Surface Solar Radiation Data Set–Heliosat (SARAH-3), now in its third version, focuses on solar surface irradiance, normalized direct horizontal irradiance, and other solar-related metrics based on satellite observational data [34]. This data set is managed by the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT) and is available

Platform	data set	REF|URL	Quantities	Location	Temporal Coverage
Sandia	Global Normal Spectral Irradiance Data	[46]|[47]	Spectral Irradiance (direct, diffuse, global horizontal, global normal), meteorological (temperature, humidity, wind speed, pressure)	Albuquerque, NM 35.0545°N, -106.5401°W	1 year constructed from data collected from 2013-2015 at 5-minute resolution
Sandia	Global Horizontal Spectral Irradiance Data	[48]|[49]	Spectral Irradiance (direct, diffuse, global horizontal, global normal), meteorological (temperature, humidity, wind speed, pressure)	Albuquerque, NM 35.0545°N, -106.5401°W	1 year constructed from data collected from 2019-2021 at 5-minute resolution
NIST	PV and Weather Data	[50]|[51]	PV performance (power, IV metrics, module temperature,), meteorological (irradiance, ambient temperature, humidity, wind speed/direction, *etc.*), sky images	Gaithersburg, MD 39.1374°N, -77.2187°W	2015-2018, 1 second instantaneous 1 minute averages
NREL	Solar & Wind Integration Data	N/A |[52][53]	Multiple integration data sets (power, radiation, wind, synthetic)	Multiple US sites	Mutiple coverage, wind 2004-2006, radiation since 1981, *etc.*
OEDI	Extreme weather data	[54]|[22]	Performance ratio, OM tickets, extreme weather events (storm, hurricane, flooding, *etc.*)	837 US sites for OM data, 186 US sites for production data	2018-2020, daily

TABLE II
OVERVIEW OF NOTABLE REGIONAL AND SITE-SPECIFIC METEOROLOGICAL DATA SETS.

for direct download from their website [35]. In addition, SARAH-2 [43] can be accessed through the website [44] or “pvlib” API [45].

B. Regional & Site-specific weather data

This section reviews site-specific and small-scale data sets that support ML analyses for PV degradation. Although these localized data sets lack the expansive spatial coverage of the global data sets discussed earlier, they offer several key advantages. First, they are typically measurement-based, avoiding reliance on the modeling and data assimilation techniques used in larger-scale data sets. Second, they often provide higher-resolution insights into the meteorological characteristics of specific locations. In addition, these data sets can be paired with actual power output data from nearby solar arrays, creating valuable opportunities to investigate the relationships between weather conditions and PV performance or degradation.

1) Sandia Spectral Irradiance Data: The Sandia National Laboratories (SNL) provide global normal and horizontal spectral irradiance data, along with other meteorological variables measured at their facility in Albuquerque, NM. The first data set includes global normal spectral irradiance, broadband irradiance (GHI, DHI, DNI), temperature, humidity, wind speed, and atmospheric pressure recorded every 5 minutes from sunrise to sunset in 2014 [46]. The second data set contains global horizontal spectral irradiance, broadband irradiance (GHI, DHI, DNI), temperature, humidity, and wind speed, with data collected in 5-minute intervals during daylight hours from 2019 to 2021 [48]. The 2019-2021 data set is composed of twelve 15-day segments, each representing a different month. Both data sets are accessible through the DuraMAT Datahub [47], [49].

2) NIST PV and Weather Data: The National Institute of Standards and Technology (NIST) has created a data set from three PV solar arrays and a weather monitoring station located on its campus in Gaithersburg, MD. This data, spanning from 2015 to 2018, includes a range of meteorological variables, such as solar radiation, wind speed/direction, temperature, humidity, precipitation, and air pressure, paired with performance data from nearby solar arrays. All of these arrays use monocrystalline silicon modules but differ in their configurations [50]. The data set provides both second-by-second instantaneous values and one-minute averages. In addition, images of the PV arrays are taken every 5 minutes and sky images are captured every 5 seconds. The data is accessible through the NIST Photovoltaic Data website [51].

3) NREL Solar & Wind Integration Data: NREL provides comprehensive wind [52] and solar [53] integration data sets designed for energy system analysis and grid integration studies. The wind data includes the Eastern and Western Wind Integration Data Sets, offering time-series measurements from 2004 to 2006 in the US, specifically created for analyzing wind power plant performance and transmission planning. For solar resources, NREL offers multiple data collections, including the Baseline Measurement System [55] with live and historical solar radiation data since 1981. The Solar Integration National Dataset Toolkit [56] provides modeled subhourly solar power data with high temporal and spatial resolution, specifically designed for regional solar generation integration studies. This toolkit focuses on time-synchronized weather conditions, site-specific solar power output ramps, and appropriate spatial-temporal correlations for accurate power system modeling [57]. Additional specialized databases include the Spectral Solar Radiation Data Base [58], collected from 1986 to 1988 across multiple US locations, representing various atmospheric conditions and collector types [59].

4) Extreme weather data: The final data set reviewed focuses on extreme weather conditions and their effects on PV performance. These data were collected as part of a study investigating the impact of compound extreme weather events on PV systems [54]. It combines operations and maintenance (O&M) ticket data, PV performance data, and weather data from more than 800 sites in 24 states. The O&M and performance data are sourced from the PV Reliability Operations and Maintenance (PVROM) database, maintained by SNL [60]. Weather data is drawn from NOAA's Storm Events Database, the Global Historical Climate Network (GHCN), and the Parameter-elevation Regressions on Independent Slopes Model (PRISM) from Oregon State University. This data set spans the years 2008–2020 with daily resolution and is categorized by climate zone. It is available on the Open Energy Data Initiative (OEDI) website [22].

C. Data processing tools and models

Some of the aforementioned data sets can also be accessed through "pvlib" API [61].Global and site-specific data sets have been extensively used to train data-driven models, with most focusing on irradiance forecasting problems [62], [63], [64]. However, recent efforts have started exploring how these meteorological data sets can aid in investigating PV degradation. One key challenge is that degradation modeling often occurs under varying environmental conditions. For example, NSRDB data have been used to train and validate a spatio-temporal graph neural network model that estimates performance loss rates using typical meteorological year (TMY) data [65]. Similarly, ERA5 satellite data have been applied with a teaching-learning-based optimization algorithm to analyze degradation rates [66]. In contrast, some studies have employed machine learning techniques to isolate non-weather-related factors contributing to performance loss. For example, NSRDB weather data, accessed via API, was used to normalize weather impacts and isolate other drivers of degradation [67]. Site-specific data, such as the high-temporal-resolution NIST data set, has been utilized to train deep reinforcement learning models for fault detection in PV systems [68]. Similarly, site-specific data from organic PV modules have been used to forecast degradation rates using neural networks [69], although this data set is available only upon request.

D. Insights

The operational environment, encompassing solar radiation and meteorological factors such as temperature and humidity, plays a significant role in the performance, degradation, and overall lifetime of PV modules. Accurately characterizing these relationships is becoming increasingly important as PV is deployed in more diverse environments worldwide. While the data sets discussed above offer a broad view of the environmental conditions PV arrays might face, environmental data alone is insufficient for fully understanding degradation mechanisms. It must be paired with performance or degradation data from PV modules. Large-scale historical data sets, such as NSRDB and ERA5, can be combined with deployed PV module data (discussed later in this review) that share spatio-temporal overlap. Site-specific data sets often include both meteorological and performance data, providing an integrated approach to degradation research.

Recent studies have begun leveraging these combined data sources to explore predictive ML models that estimate degradation rates under varying environmental conditions. However, significant open questions remain, particularly regarding the relationship between different degradation pathways and specific environmental factors, as module materials and other array characteristics that affect degradation rates also often vary between locations. The development of explainable ML methods, which aim to provide interpretable insights into why certain factors contribute to degradation, could drive the design of more resilient PV modules. Moreover, connecting these insights with operational decision making could help improve maintenance practices and extend the lifespan of PV arrays.

III. PV LOCATION

This section reviews published data sets on PV installations, covering both utility-scale and rooftop systems across national and international facilities.

PV location data includes geospatial information such as geographic coordinates, land cover details, and attributes of solar installations, including capacity, operational status, and installation dates. Figure 3 illustrates an example of the global installation capacity distribution as of 2018. This kind of data is crucial for understanding the global distribution, capacity and impact of PV systems. Accurate PV location data assists in managing generation intermittency, assessing climate change mitigation efforts, and identifying potential land-use conflicts with biodiversity and conservation priorities [70].

Collecting PV location data primarily involves processing remote sensing imagery to detect and classify installations [71], [72]. This approach allows for the extraction of detailed geospatial data, including precise location, size, and installation type. Advances in machine learning, particularly deep learning models like UNet, have greatly enhanced the accuracy and scalability of PV inventories derived from satellite imagery [73].

A. Data Sources

This subsection reviews four key PV location data sets, including country-specific PV sites in the US and UK, global PV installations, and a rooftop PV database. Table III provides an overview of these data sets.

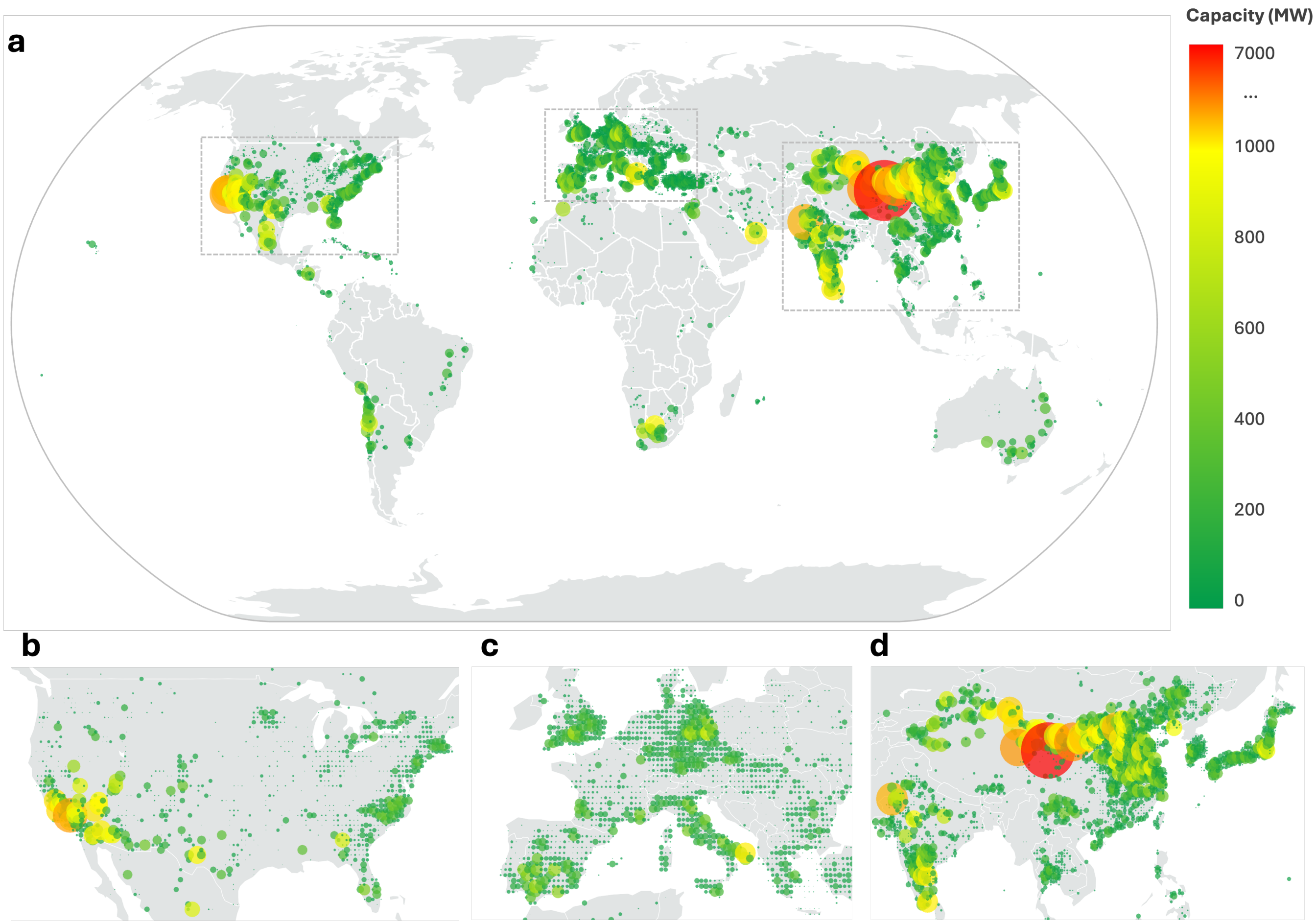

Fig. 3. Visualization map of global PV installation capacity by 2018. (a) Worldwide PV installation (b) US (c) Europe (d) Asia. Figure generated using the data from [74]

1) US PV sites: The Lawrence Berkeley National Laboratory (LBNL) and the United States Geological Survey (USGS) have published the "United States Large-Scale Solar Photovoltaic Database" (USPVDB) [75] which documents 3,699 ground-mounted large-scale solar photovoltaic (LSPV) installations in the United States, with a total rated capacity of 55,156 MWAC [76]. This data set covers facilities in 47 states and the District of Columbia, providing georectified polygons and detailed attributes such as geographic coordinates, land area, operational year, power technology, and capacity. The data is accessible in multiple formats (CSV, XML, GIS shapefile, GeoJSON) and through an interactive web application [77].

2) UK PV sites: Queen Mary University of London (QMUL) and the Alan Turing Institute, in collaboration with Open Climate Fix Ltd. and the University of Sheffield, have published a data set [78] comprising over 260,000 PV installations across the UK [79]. This data set covers an estimated 86% of the country's total capacity, with mapped installations totaling 10.66 GW or 13.93 GW when inferred capacities are included. The data provide geographic details and metadata for each installation, including coordinates, capacity, installation type, and surface area. The data is available in CSV and GeoJSON formats.

3) Global PV sites: The University of Oxford and Descartes Labs Inc., in collaboration with the World Resources Institute, have published a data set [70] of global PV installation sites [74]. This data set documents 68,661 commercial, industrial and utility-scale PV installations worldwide, each with a nameplate capacity exceeding 10 kW, estimating a global installed generating capacity of 423 GW as of the end of 2018. Each installation is associated with metadata including coordinates, capacity, installation date, land cover, and proximity to protected areas. The data is available in CSV and GeoJSON formats, as well as via an interactive web application [80].

4) PV Rooftop Database: NREL has published two databases of rooftop PV sites. The first data set, Photovoltaic Rooftop Database (PVRDB) [81], is a geospatially resolved lidar-derived data set of suitable roof surfaces and their PV technical potential for 128 metropolitan regions in the United States. The PVRDB data are organized by city and by year of lidar collection. The second one, PV Rooftop Database for Puerto Rico (PVRDB-PR) [82], collects additional data of PV suitable roof surfaces for 96% buildings in Puerto Rico.

Platform	Data set	Ref | URL	Size	Spatial coverage	Temporal coverage	Remark
LBNL	USPVDB	[76] |[75]	3,699 ground-mounted facilities; 55,156 MWAC capacity	47 US states plus the district of Columbia	1986 - current	Interactive web application available [77]
QMUL	UKPVGeo	[78] |[79]	over 260,000 PV installation sites	86% capacity across UK	N/A - 2020	
Oxford University	Global inventory	[74] |[70]	68,661 PV installation site	World wide	N/A - 2018	Interactive web application available [80]
NREL	PVRDB	N/A |[81]	N/A	128 metropolitan regions in the US	2006 - 2014	rooftop data
NREL	PVRDB-PR	N/A |[82]	N/A	PV suitable roof surfaces for 96% buildings in Puerto Rico	2015 - 2017	rooftop data

TABLE III
OVERVIEW OF PV LOCATION DATA SETS.

B. Data processing tools and models

Most research in this area focuses on constructing PV installation data sets rather than performing post-modeling analyzes on the data. The locations of PV installation can be determined from Google Maps [83] or OpenStreetMap [84]. To streamline the collection and processing of PV location data, researchers increasingly rely on automated processes that integrate remote sensing technologies, such as LiDAR and high-resolution satellite imagery, with computer vision models such as convolutional neural networks (CNNs) for detecting and classifying PV installations. This section reviews two automated pipelines for building large-scale data sets of rooftop and ground-mounted PV installations. The tools and models are summarized in Table XII. Other tools or models that are used to process satellite images, but are not designed primarily to construct location data sets, will be reviewed in Section V.

Kruitwagen, *et al.*, developed a machine learning pipeline to map commercial, industrial, and utility-scale PV installations worldwide using high-resolution satellite imagery [74]. The pipeline employs CNNs in the UNet architecture to process and locate PV installations from Sentinel-2 and SPOT6/7 imagery, effectively identifying the presence and spatial arrangement of PV panels. Recurrent neural networks (RNNs) are used in the second stage to analyze time-series data derived from the imagery, helping to filter out false positives and estimate installation dates. This pipeline is implemented in Python and is available online [85].

The second pipeline, "URSUS-PV," developed by Rodríguez-Gómez, *et al.*, uses LiDAR data to estimate rooftop solar installations in urban areas [86]. Semantic segmentation techniques are applied to extract roof characteristics such as orientation, inclination, and area from LiDAR images, which are then combined with meteorological data to predict PV energy production. The tool estimates both the next-day hourly production and the long-term average solar energy potential for selected urban areas. Written in R, this tool is available on GitHub [87], and while it also supports a web-based dashboard [88], the dashboard is currently nonfunctional.

C. Insights

Comprehensive PV location data sets are critical for understanding the global distribution, capacity, and future installation trends of PV systems. To accurately and automatically detect installation sites, large-scale annotated data sets are essential for training machine learning models. The data sets reviewed here provide detailed geospatial information on PV installations, which will be vital for developing more advanced models as computer vision technology evolves. Improved models will also facilitate timely updates of the installation distribution maps, addressing the issue of outdated data in the current maps. Future projects could leverage PV installation data to analyze correlations with socioeconomic factors or urban planning metrics, grid, distribution, and load planning, and to predict PV system degradation in specific areas. In addition, ML can be developed to directly predict solar energy potential from satellite or LiDAR images by integrating geospatial and meteorological data.

IV. PERFORMANCE DATA

PV performance data is used for evaluating the efficiency, reliability, and overall behavior of solar energy systems under various environmental conditions. This section reviews open data sets related to PV performance, focusing on two main types: production data and current-voltage (I-V) characteristics data.

Production data provides insights into the actual power output of PV systems over time, facilitating the assessment of performance degradation, shading impacts, soiling, and other real-world conditions. These data are essential for short-term

Owner	Data set	Ref |URL	Size	Spatial coverage	Temporal coverage	Remarks
NREL	PVDAQ	N/A |[89]	158 PV systems	158 sites in the U.S.	1 - 29 years	Experimental and commercial PV sites
Elia Group	Elia Open Data platform	N/A |[90]	12 PV production data sets	Belgium, Germany	2018 - 2024	Data platform
BOKU	Chile GCPV database	[91] |[92]	103 PV systems	Chile	2014 - 2018	Capacity factor
INESC TEC	solar power data sets	[93] |[94]	44 household PV units	Portugal	2011 - 2013	Hourly time series
UCSD	Micro grid data sets	[95] |[95]	26 PV plants	San Diego, US	2015 - 2020	Real AC power
CWRU/UCF	SunSmart	N/A |[96]	90 PV systems	Florida, US	2012 - 2016	Time series
CWRU/Sandia	DOE-RTC	N/A |[97]	8 PV Systems of c-Si	Albuquerque, Denver, Las Vegas, Orlando	2014 - 2019	Minute-by-Minute time-interval data, with Ground and Satellite based weather.
ARENA	Yulara Solar Project	N/A |[98]	5 PV systems (22.6 - 1058.4 kW)	Australia	2016 - 2024	used in [99]
DKASC	Alice Springs system	N/A |[100]	38 PV systems (2 - 10.5kW)	Australia	2009 - 2024	Higher resolution data may also be available on request, used in [101]
Standford	Stanford Benchmark	[102] |[103]	30.1 kW PV system	California, US	2017 - 2019	

TABLE IV
OVERVIEW OF PRODUCTION DATA SETS.

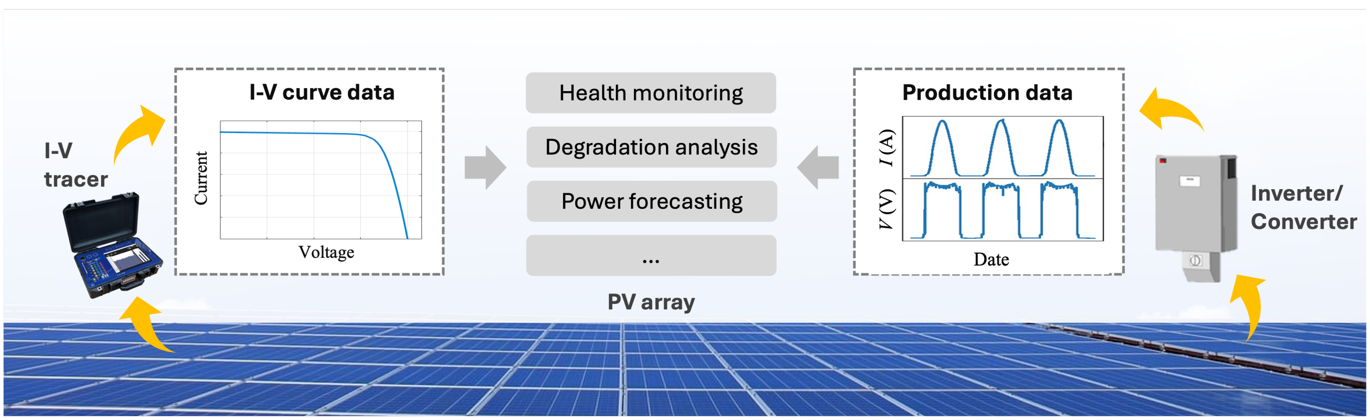

Fig. 4. Visualization of production and IV curve data and their applications.

performance analysis and long-term yield predictions, which are critical for optimizing system design and maintenance strategies.

I-V curve data offer a more detailed view of the PV module performance by depicting the relationship between current output and voltage under varying irradiance and temperature conditions. The analysis of the I-V curves provides valuable information about the maximum power point and other critical operating points of the module, as well as potential problems such as cell mismatch or partial shading.

A. Data sources for production data

PV production data, particularly related to electrical characteristics, involves monitoring current and voltage (DC or AC) throughout the system to evaluate performance. These measurements are critical for calculating the power output, which reflects the energy production capacity at any given time. Analyzing these data helps maintain optimal system operation, detect performance issues, and diagnose system faults.

A summary of the production data available on the major platforms and individual data sets is presented in Table IV.

1) Photovoltaic Data Acquisition (PVDAQ) Public Data sets: The NREL PVDAQ [89] is a large-scale time series database that contains system metadata and performance data from various experimental and commercial PV sites. These data sets are

Owner	Data set	Ref |URL	Size	Spatial coverage	Temperoal coverage	Remark
NIST	NIST PV testbed	N/A [51]	6 reference modules	Maryland, US	2015 - 2018	For roof, parkinglot, and groudn PV installations
Sandia	PV lifetime Modules	N/A |[106]	916 modules	NM, CO, FL, US	2016 - now	mono/multi-c-Si, HIT PV module
NREL	Performance model validation	[107], [108] |[109]	11 modules	CO, US	2011 - 2014	mono/multi-c-Si, CdTe, CIGS, a-Si, HIT PV module

TABLE V
OVERVIEW OF DATA SETS FOR IV CHARACTERISTICS.

commonly used for ongoing performance and degradation analyzes, addressing factors that impact PV performance, such as soiling and other environmental conditions.

2) Elia Open Data platform: The Elia Open Data platform [90], managed by the Elia Group, provides public access to extensive data on electricity grid and market operations in Belgium and parts of Germany. The data are available in multiple formats (CSV, JSON, XML). The platform offers tools for data access and visualization.

3) Chile GCPV database: The University of Natural Resources and Life Sciences, Vienna (BOKU) has published a database [91] that includes hourly capacity factors (values 0 to 1) for 103 large PV installations in Chile. These factors, derived from official sources and simulated using "pvlib" with ERA5-land and MERRA-2 reanalysis data, cover the period from 2014 to 2018. Simulations were performed under two system assumptions: a "fixed" system with optimal north-facing orientation and latitude-based inclination, and a horizontal single-axis tracking system with backtracking.

4) INESC TEC solar power data sets: The Institute for Systems and Computer Engineering, Technology and Science (INESC TEC) data set [93] contains hourly time-series data of electrical energy from 44 small household PV units, with installed capacities ranging from 1.1 to 3.7 kWp. This data, collected via smart meters at 15-minute intervals, was used in the "FP7 SuSTAINABLE" project.

5) UCSD micro grid data sets: The UC San Diego (UCSD) microgrid [95] consists of 26 PV plants on campus, where solar inverters operate at unity power factor according to the 2003 IEEE 1547 standard. The data set includes real AC power generation data for the PV plants, which can be downloaded from the supplementary materials of their paper [95].

6) CWRU/UCF OSF SunSmart Schools data set: This data set originates from Florida's SunSmart Schools program [104], which installed 10 to 12 kW PV + battery systems in public schools to reduce energy costs and enhance resiliency when used as emergency shelters. The data includes 30 systems, each equipped with three PV arrays, and provides PV and battery DC current and voltage measurements, along with other relevant metrics.

7) DOE RTC Baseline System: The Department of Energy funds several Regional Test Centers (RTCs) [105] across four locations in the US: Albuquerque, NM (Sandia), Denver, CO (NREL), Las Vegas, NV (Sandia) and Orlando, FL (UCF/FSEC). These sites provide high-resolution (1-minute interval) performance data from identical c-Si systems, which are maintained to isolate the effects of different climates. These high-quality data are useful for model validation and development.

B. Data sources for I-V characteristics

I-V curve data sets typically fall into two categories: laboratory (indoor) measurements and in-situ (outdoor) measurements. I-V curves demonstrate the diode behavior of PV modules and provide valuable insight into sources of power loss, such as internal degradation and external factors such as non-uniform shading. These data sets can also be used for production modeling to predict how a module will perform under specific irradiance and temperature conditions.

Since actual module performance may deviate from nameplate values, I-V curves measured at various irradiance and temperature conditions are used to generate PAN files for PVSyst or to fit the Sandia Array Performance Model (SAPM). These models enable for more accurate performance simulations for specific PV systems.

One of the most widely used databases for PV module performance characteristics is the California Energy Commission (CEC) PV Module Database [110], which lists nameplate values for key performance metrics under standard test conditions (STC) and nominal module operating temperature (NMOT), along with temperature coefficients. The CEC database is regularly updated and integrated with tools such as the System Advisor Model (SAM) and pvlib-python.

In some cases, only the I-V parameters are reported rather than full I-V curves. These parameters can include the full single-diode model variables (I_L, I_0, R_s, R_{sh}, nN_sV_{th}) or more basic parameters such as short-circuit current (I_{SC}), maximum power current (I_{MP}), open-circuit voltage (V_{OC}), maximum power voltage (V_{MP}), series resistance (R_s), and shunt resistance (R_{sh}). It is important to note that R_s and R_{sh} are typically fitted with the single-diode model, while in basic models they are estimated from the slopes of the I-V curve near V_{OC} and I_{SC}, respectively.

Here, we review publicly available I-V data sets, including full curves and parameter data, for both laboratory and in-situ measurements. These data sets are summarized in Table V.

1) Laboratory (indoor) I-V data sets: Although I-V curve measurements are commonly conducted in PV test labs, high-quality public data sets for laboratory I-V characterizations are relatively rare. Laboratory I-V measurements are typically conducted under STC (1000 W/m^2, 25 °C), or across a range of irradiance and temperature values. Most laboratory setups use flash testers, which emit a brief pulse of light at the desired intensity, and the I-V curve is recorded by modulating the panel voltage and measuring the current response during the pulse. I-V curves may also be derived from measurements such as Suns-Voc [111], which provides a curve that is independent of series resistance. This can be useful for measuring the impacts of accelerated aging on other parameters, although Suns-Voc measurements are not widely published.

Steady-state solar simulators, while less common, can also be used. Solar simulators are rated based on the IEC 60904-9:2020 standard, with classifications ranging from A to C, depending on their performance in spectral matching, spatial uniformity, and temporal stability.

Often I-V curves are taken to verify initial performance of modules compared to nameplate values, and/or at various time steps in the module's life to identify degradation mechanisms and quantify power loss.

Laboratory I-V curves are often used to verify initial performance relative to nameplate values or to monitor degradation mechanisms by periodically testing the same module over its lifespan, either for fielded modules or those undergoing accelerated aging (*e.g.*, damp heat or thermal cycling).

The Sandia National Laboratories PV Lifetime project [106] has released laboratory characterization data for 12 module types, including I-V curves and resulting PAN files, following the IEC 61853-1 and 61853-2 standards. However, few comprehensive flash-test I-V data sets exist online, limiting their availability for machine learning applications.

2) In-Situ (outdoor) I-V data sets: In-situ I-V measurements, conducted outdoors on PV modules and systems, are less common than lab-based tests, but advances in irradiance and performance monitoring tools make these data more accessible. Module-level I-V tracers are now marketed for bifacial single-axis tracking PV arrays, which enable more effective irradiance monitoring [112]. In-situ I-V data is collected under ambient temperature and plane-of-array (POA) irradiance, and accurate analysis requires corresponding meteorological data, including measured or modeled module temperature and global horizontal POA irradiance.

An example of the in-situ I-V data set is the "Data for Validating Models for PV Module Performance" [109] from NREL, which provides approximately one year of time-series I-V and meteorological data for 11 PV modules of various technologies across three locations: Cocoa, FL; Golden, CO; and Eugene, OR [107], [108]. NIST has also built up an open-source online data platform, where the time-series I-V curve data from 6 reference modules can be downloaded here [51].

3) Synthetic I-V data: Some researchers have also developed synthetic I-V data sets for training and testing machine learning algorithms [113], [114], [115]. These synthetic I-V curves are typically generated using the single-diode model or other electrical models to simulate performance under different degradation scenarios or system faults. However, these data sets are not widely published, with only the work by Hopwood [113] providing tools for generating synthetic data publicly.

C. Data processing tools and models

Common methods for processing PV performance data include time correction, clear-sky detection, outlier detection, and I-V curve correction. Time correction ensures consistency by adjusting for daylight savings, time zones, and offsets [116]. Clear-sky detection identifies cloud-free conditions using algorithms or sensors, allowing high-quality data to be filtered for analysis [117]. Outlier detection is critical for identifying abnormal data points that deviate from expected patterns, using statistical methods or machine learning algorithms to ensure data quality [118]. Time-series production data have been widely studied for power forecasting, with machine learning models such as artificial neural networks (ANNs) [119] and long short-term memory (LSTM) networks [120] commonly applied.

For field-measured I-V curves that are typically recorded under varying environmental conditions, correction to STC is essential for consistent analysis [121]. In addition, extracting key parameters from I-V curves is a common requirement. A summary of open-source tools used for these data processing tasks is presented in Table VI, where each tool may support multiple functions.

D. Insights

Performance data is one of the most readily available types of PV data, but several common issues persist in the published data sets. A major issue arises from maintenance activities and configuration changes that occur over the years during field operation, such as system shutdowns, sensor replacements, or panel removals. These events are frequently undocumented in the data set, leading to "outliers" that require manual verification and correction if analyzed under the assumption of the original system configuration.

Another significant challenge is the lack of essential metadata in some data sets. Without key information such as PV module specifications, system configuration details, or the types of weather data recorded, it becomes more difficult to accurately analyze and interpret the performance data. The absence of such metadata can hinder the ability to draw reliable conclusions.

In most cases, production data reflects the maximum power point (MPP) on the I-V curve, making it possible to correlate the two types of data. Tools like PV-Pro can reconstruct I-V curves from production data, enabling more detailed performance analysis.

Function	Tool name	Ref |URL	Data type	Remarks
Time correction	solar-data-tools	N/A|[122]	Time-series data	
Time correction	pvlib	[116]|[123]	Time-series data	
Clear sky detection	pvlib	[116]|[123]	Time-series power and weather data	
Clear sky detection	solar-data-tools	N/A|[122]	Time-series weather and production data	
Clear sky detection	Rdtools	N/A|[124]	Time-series power and weather data	Based on a multi-step clear sky workflow
Outlier detection	Rdtools	N/A|[124]	Time-series power and weather data	Remove outages and outliers
Outlier detection	PV-Pro	[3]|[125]	Time-series weather and production data (DC voltage, current)	Detect outliers based on current vs irradiance, voltage vs temperature
I-V curve correction	Pdynamic	[121]|[126]	I-V curves	Applicable for degraded PV modules
Parameter extraction	ddiv (Data-driven I-V feature extraction)	[127]|[127]	I-V curves	Quantify I-V features including steps
Parameter extraction	$Suns\text{-}V_{OC}$	[128]|[129]	Time-series I-V curves	Quantify degradation and power loss
Modeling	SolarSpatialTools	[130]|[131]	Cloud motion	Python package for spatial solar energy analysis
Power forecating	Analytics-project	[119]|[132]	Time-series power and NWP data	ANN model for power forecasting

TABLE VI

OPEN-SOURCE TOOLS FOR PROCESSING PERFORMANCE DATA (PRODUCTION OR I-V CURVE DATA).

V. IMAGE INSPECTION

This section reviews open data sources and processing tools for various imaging techniques used in PV system evaluation. We cover data sets containing visual / satellite, Electroluminescence (EL), Photoluminescence (PL), Infrared (IR) and Ultraviolet Fluorescence (UVF) images. Examples of these images are shown in Figure 5(a).

Early detection and replacement of PV modules with defective cells are essential to avoid significant energy losses and mitigate potential safety hazards. To maintain optimal performance, fast and reliable evaluation methods during and after module production are critical. Imaging techniques, often used alongside I-V curve measurements, are commonly employed to detect defects in PV modules. Figure 5(b) and Table VII summarize some common failure modes and related imaging techniques that can be applied [133], [134].

Imaging offers several advantages for PV system evaluation, such as non-destructive testing and rapid data acquisition. Techniques such as IR imaging allow for inspection even during normal PV system operation. Recent advances in machine learning and drone technology have further improved the efficiency and accuracy of these inspections, reducing the need for human intervention. For example, object detection models have been developed to automatically locate defective cells in images [6], [7], [135], [136]. Furthermore, quantitative information such as power output predictions can be derived from EL images, providing deeper insights into system performance [137].

A. Data sources for Visual/Satellite images

Visual inspection is a common method for detecting defective modules in the field, traditionally performed by personnel on the ground. In recent decades, this method has evolved with the use of unmanned drones, providing a bird's-eye view of installations [141]. While drones can be integrated with more sophisticated inspection methods, the most cost-effective approach is ordinary visual inspections, as they do not require special cameras. Visual images can detect many types of failures, as listed in Table VII. However, some faults are difficult to detect visually, making techniques such as EL, PL, or IR imaging more effective, although at a higher cost due to the need for specialized equipment or labor. For example, taking EL images requires disconnecting and polarizing the modules in a string. In addition, visual inspections can reveal issues beyond the modules themselves, such as racking or tracker failures. Currently, there are no known open-source databases of drone-based aerial images of solar fields to our knowledge.

A growing trend in visual inspections is the use of satellite imagery [142]. Satellite images have been employed in multiple studies [74], [143], [144], [145], [146], [147], enabled by freely available images from sources such as Google Maps, which can be accessed using the Maps static API [83]. These images have a sufficient resolution to detect solar modules [142], although the time resolution of the images is not consistent. Depending on the location, it ranges from three months to a year.

Annotated data sets are essential for detecting solar modules in satellite images using deep learning methods [142]. One such publicly available training set was obtained using Google Earth Engine and is described in [148], with the data accessible on Zenodo [140]. Other satellite image data sets used for training models include those for DeepSolar++ [143] and PV Identifier [73]. The DeepSolar++ data are available on GitHub [149], and the PV Identifier uses a labeled data set of solar

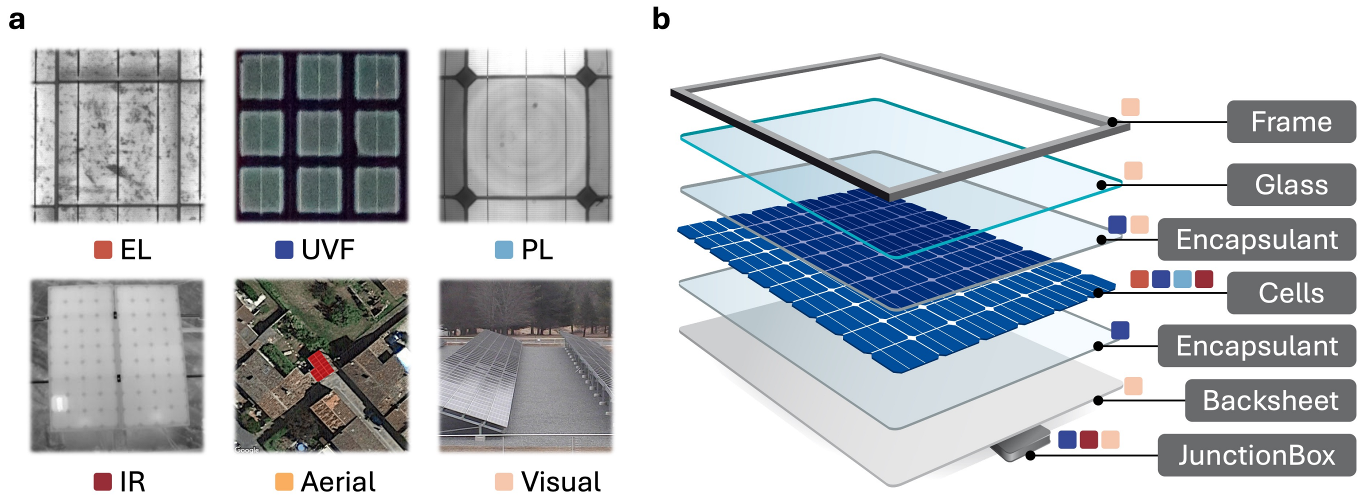

Fig. 5. Imaging techniques and module structure. (a) Example of photos taken from Various imaging techniques. (b) The layered structure of a solar module. The colored squares represent the common imaging techniques used to characterize these module components. EL and PL images are from the authors' own work. The UVF image is from [138]. The IR image is from [139]. The aerial image is from [140]. The visual image is from [51].

PV component	Failure mode	Imaging technique
Solar cell	Cracking	EL, PL, UVF
	Hotspot	EL, PL, IR, UVF
	Contact disconnection	EL
Encapsulant	Yellowing	Visual, UVF
	Delamination	Visual, UVF
Backsheet	Delamination	Visual
	Cracking	Visual
Front glass	Breakage	Visual
Frame	Breakage	Visual
Junction box	Breakage, hot, poor sealing	Visual
	Hot	IR
	Poor sealing	UVF
Connector	Breakage	Visual
	Hot	IR
Rack	Breakage (wind)	Visual
Tracker	Breakage (wind), stall	Visual

TABLE VII
OVERVIEW OF FAILURES AND IMAGE INSPECTION TECHNIQUES.

modules in 600 satellite images and are available here [150]. Training data for the Global PV Inventory [74] is also available on Zenodo [70], and a data set to detect mounting configurations from NREL [151], [145] is available on the DuraMat DataHub [152].

B. Data sources for EL images

PV cells primarily function by absorbing light and converting it into electricity. According to the reciprocity principle, the reverse process is also possible, allowing EL imaging. The EL technique involves applying a direct current to a PV module and capturing the emitted photons with an infrared-sensitive camera. EL imaging is highly effective for detecting microcracks and other defects in the cell material, helping to identify and address faults in PV modules.

To capture EL images, solar modules are typically set under forward bias in a dark environment, and an EL camera is used to collect luminescence. Cameras equipped with silicon charge-coupled device (CCD), complementary metal oxide-semiconductor (CMOS), or InGaAs sensors are positioned perpendicular to the module surface to prevent distortions such as perspective errors. Long lenses help minimize lens or fisheye distortion, and long-pass filters (usually 850–1050 nm) isolate the signal from background noise. Emerging technologies also allow for in-situ EL imaging during daylight [153].

Platform	Data set	Ref |URL	Size	Indoor/Outdoor	Module Info	Remark
LBNL	Crack segmentation	[154]|[155]	1,837 cell images with annotated masks; 400 x 400 pixels / image	Indoor	Mono/poly-c cells; Various busbars	Contain pixel-level annotations for crack segmentation
NREL	DMX	[156]|[157]	146,226 time-series cell images; 640 x 512 pixels / image	Indoor	Poly-c cells	Study crack evolution through millions of small cycles
EPRI	PREDICTS2	[158]|N/A	108 images from 36 modules with I-V; 3408x1704 pixels / module image; 284x284 pixels / cell image	Both	Poly-c cells	Data collected before and after aging test, and after field exposure; Contain flash test data at each time step
Sandia	EL for VAE	N/A |[159]	1,180 cell masks with annotated masks; 400 x 400 pixels / image	N/A	N/A	Crack mask only
PVEL-KIWA	N-type field aging	N/A |[160]	N/A	Outdoor	Half-cut cells; Multiple busbars	Cracked cells after field aging; Contain flash test data
FAU	ELPV	[161]|[162]	2,624 cell images with class labels; 300 x 300 pixels / image	Indoor	Mono/poly-c cells	Used for classifying damage levels
FAU	ELPVPOWER	[137]|[163]	718 module images with I-V; 900 x 1500 pixels / image	Indoor	Mono/poly-c cells	Used for predicting the operational maximum power
Beihang	PVEL-AD	[164]|[165]	36,543 cell images with annotated bbox; 1, 024 x 1, 024 pixels / image	Indoor	Poly-c cells	Contain bounding box annotations for object detection
Wits	Benchmark EL	[166]|[167]	1,266 cell images with annotated masks; 2,573 cell images without annotations; 512x512 pixels / image	Indoor	Mono/poly-c cells	Contain pixel-level annotations for 24 different objects
UCF	EL-Defect	[8]|[168]	17,064 cell images with annotated masks; 400 x 400 pixels / image	Indoor	Mono/poly-c cells	Contain pixel-level annotations for 5 objects
PERL	TF-PV	[169]|[170]	6,000 module images with I-V	Indoor	CIGS module	Contain flash test data
CWRU	Damp heat EL	N/A|[171]	1028 cell images with class labels 250 x 250 pixels / image	Indoor	Mono/poly-c cells	Study indoor damp heat exposure

TABLE VIII
OVERVIEW OF EL IMAGE DATA SETS.

EL image data sets typically include raw images, annotations (if used to develop computer vision models), and module information, such as silicon type and cell configuration. The EL data sets discussed in this section are summarized in Table VIII.

1) LBNL crack segmentation EL: LBNL has published a crack segmentation data set [172], [155] containing 1,837 annotated EL images of individual solar cells. This data set was used primarily to train semantic segmentation models [154]. Cell images, with a resolution of 400×400 pixels, include various types (monocrystalline and polycrystalline silicon) and busbar configurations. Pixel-level annotations classify images into five categories: “background”, “crack”, “cross crack”, “busbar”, and “degradation area”, stored as binary masks. The data set also includes 576 module images from which the cell images were cropped. All images and annotations are in JPEG format.

2) NREL time-series EL: NREL has published a time-series EL data set [157] that contains 146,226 cell images, in order to study the long-term degradation of cracked cells. These images were collected from a single cell in a polycrystalline module under dynamic mechanical acceleration (DMX) aging tests, simulating wind-driven pressure cycles [156]. The data is organized by four pressure levels (10 Pa, 30 Pa, 100 Pa, and 300 Pa), with one million cycles conducted at each level. EL images were taken throughout the testing process, with a resolution of 640×512 pixels.

3) EPRI accelerated aging test EL: The Electric Power Research Institute (EPRI) has released a data set [173], [158] that contains 108 module images collected from 36 polycrystalline modules at three time points in the project: “Physics of Reliability: Evaluating Design Insights for Component Technologies in Solar 2” (PREDICTS2) [174]. The data set includes EL images taken before and after the standard IEC 61215 aging tests [175] and after field exposure. Both original and processed (perspective-corrected) module images are available, with resolutions of 3408×1704 and 284×284 for individual cell images. The flash test I-V parameters corresponding to the EL images are also included.

Platform	Data set	Ref |URL	Size	Indoor/Outdoor	Module Info	Remark
Raptor Maps Inc.	InfraredSolarModules	[178] |[177]	20,000 cell images with class labels; 24 x 40 pixels / image	Indoor	Half-cut cells	Contains class labels of 11 different anomalies
University of Valle	Solar panel infrared images	[180] |[181]	67 module images with annotated bbox; 336 x 256 pixels / image	Outdoor	Mono-c cells	Contain bounding box annotations for object detection

TABLE IX
OVERVIEW OF IR IMAGE DATA SETS.

4) Sandia crack EL for VAE: SNL provides a data set [159] of 1,180 crack mask images, extracted from EL images. The images, with a resolution of 400×400 pixels, were segmented using a retrained version of the PV-Vision model [172], [154], and are stored in TIFF format.

5) University of Erlangen–Nuremberg (FAU): The University of Erlangen–Nuremberg (FAU) has published two EL data sets: "ELPV" and "ELPVPOWER". The ELPV data set [173] contains 2,624 grayscale images (300×300 pixels) of monocrystalline and polycrystalline cells from 44 different solar modules, classified into four damage levels [161], [4]. The ELPVPOWER data set [176] includes 718 module images (900×1500 pixels), with measured maximum power (P_{mp}) values for each module [137].

6) Beihang PVEL-AD: Beihang University has published a data set called "PVEL-AD" [165], developed during research on defective cell detection [164]. This data set contains 36,543 cell images, each with a resolution of 1024×1024 pixels. It includes annotations for 21,044 defects, classified into eight classes such as "linear crack", "star crack", "finger interruption" and "black core". Additionally, 4,148 images are labeled with two cell-level classifications: "misalignment" and "material". These cell images are stored in JPEG format, and the annotations are provided in XML files.

7) Wits Benchmark EL: The University of Witwatersrand, Johannesburg (Wits) has released a data set [167] containing 1,266 labeled and 2,573 unlabeled cell images for semantic segmentation [166]. These images, with a resolution of 512×512 pixels, are cropped from monocrystalline and polycrystalline silicon modules. The pixel-level annotation masks cover 24 different classes. The labeled data are provided in two batches, with 582 cell images in the first batch and 684 in the second. All images and corresponding annotations are in PNG format.

8) UCF-EL-Defect: The University of Central Florida has published a data set called "UCF-EL-Defect" [168], used for semantic segmentation [8]. This data set contains 17,064 cell images, cropped from monocrystalline and polycrystalline silicon modules, with annotation masks for ten categories. Due to class imbalance, some categories are grouped into five broader classes: "background", "crack", "contact", "interconnect", and "corrosion". The images are around 400×400 pixels, stored in JPEG format, and the annotation masks are provided in PNG format.

9) PERL TF-PV: The IEK5-Photovoltaik Forschungszentrum Jülich has published a data set [170] with 6,000 CIGS module images and corresponding I-V data [169].

10) CWRU: The SDLE Research Center at CWRU has published a data set [171] entitled "Electroluminescent (EL) Image Dataset of PV Module Under Step-wise Damp Heat Exposures". This data set is part of an indoor damp heat exposure project, where modules were cut into cells and labeled as "Good", "Cracked", or "Corroded". EL images were taken every 500 hours during a 3,000-hour exposure, with 60 cell images captured at each step. The primary failure mode in this data set is heavy corrosion, with some minor cracking in a few cells.

C. Data sources for IR images

IR thermal imaging detects radiation emitted by a surface in the infrared wavelength, providing a temperature distribution across the surface. IR images are commonly used for detecting hotspots, electrical disconnections, and short-circuited cells in PV modules.

Data sets containing IR images are summarized in Table IX.

1) Raptor Maps: Raptor Maps provides one of the largest open-source infrared image data sets for solar module inspection [177]. These images, captured by unmanned aerial vehicles, are low-resolution (24×40 pixels) with a pixel size that varies from 3 to 15 cm per pixel. The images are labeled into 12 categories, including cracked cells, soiling, and vegetation blockage, among others [177], [178]. The data is available for download as a zip file, with accompanying metadata in a JSON file. The JSON structure and instructions for data access are provided in the Raptor Maps repository [177], [178]. This data set is well suited for developing image classification pipelines using CNNs to detect anomalies in solar panels.

2) University of Valle: The University of Valle in Colombia has released a data set titled "Solar Panel Infrared Images", consisting of 67 grayscale infrared images of solar panels [179]. The images, which include defects such as snail trails and hotspots, are available for download [180]. The data set is organized by acquisition date, with each folder containing the IR images collected that day. An annotated version of this data set, with bounding boxes marking detected defects, is also available [181].

Platform	Data set	Ref |URL	Size	Indoor/Outdoor	Module Info	Remark
KIT	Perovskite PL	[183] |[184]	1,129 cell images with I-V; 65 $\times$ 56 pixels / image	Indoor	Perovskite cells	Contain flash test results
NREL	Field PL	N/A |[185]	856 module images; 4056 $\times$ 3040 pixels / image	Outdoor	Half-cut cells	

TABLE X
OVERVIEW OF PL IMAGE DATA SETS.

Platform	Data set	Ref |URL	Size	Indoor/Outdoor	Module Info	Remark
Southern Company	OpenUVF	[190] |[138]	50 labeled module images; 100 unlabeled module images; 3,000 labeled cell images; 21,600 unlabeled cell images	Outdoor	Over 20 distinct module models from over 8 manufacturers	Contain bounding box annotation for defect detection
BrightSpot Automation	Inspection	N/A |[191]	33 entries	Outdoor	Different modules of 30 unique make-model combinatoins	See table at the bottom of the webpage
UCF	UVF in Field	[188] |[192]	2,989 module images	Outdoor	7,190 modules of 26 unique make-model combinations	Covers 3 climate regions

TABLE XI
OVERVIEW OF UVF IMAGE DATA SETS.

D. Data sources for PL images

PL imaging uses optical excitation to capture images without requiring electrical connectivity to the module. Cameras with InGaAs detectors are commonly used to capture emitted light. PL imaging is effective for detecting cracks, snail trails, and EVA degradation in PV modules [182]. In addition, PL can help determine the power output of PV modules.

Data sets containing PL images are summarized in Table X.

1) Karlsruhe Institute of Technology (KIT): The Karlsruhe Institute of Technology (KIT) has published a PL image data set of perovskite cells to address the scalability challenges in perovskite fabrication [183]. This data set is hosted on Zenodo [184] and can be downloaded in HDF format from [186]. The data set includes 780 solar cell images for training and 349 images for testing. The target variables for training include metrics such as power conversion efficiency, open-circuit voltage, short-circuit efficiency, and fill factor.

2) NREL field PL: NREL has published a PL image data set on DuraMat Datahub [185] for field-aged modules, which includes 856 images of half-cut cells. The images are stored in JPEG format, with a resolution of 4056$\times$3040 pixels.

E. Data sources for UVF images

UVF imaging provides a high-throughput, non-intrusive method to inspect PV modules in the field. UVF images are typically captured at night to minimize interference from solar radiation and are often collected using unmanned aerial vehicles (UAV) or locally mounted monopod systems.

PV modules contain polymer encapsulants with some additives fluorescing under UV light. EVA, the most commonly used encapsulant material, releases fluorophores as it degrades. UVF imaging can reveal cell defects (*e.g.*, fractures, delamination, cracks beneath busbars), differences in materials, and issues with edge and junction box sealing [187], [188].

The UVF data sets introduced in this section are summarized in Table XI.

1) Southern Company OpenUVF: Southern Company has developed a large UVF data set as part of the DuraMAT initiative, available alongside the OpenUVF software on GitHub [189], with the data set hosted on Dropbox [138]. This data set includes 50 labeled module images, 100 unlabeled module images, 3,000 labeled cell images, and 21,600 unlabeled cell images, all in PNG format. The primary goal of this data set is crack detection in individual cells, and detailed methods are described in [190].

2) UCF, CWRU and Brightspot: UCF, CWRU, and BrightSpot Automation have jointly published a UVF image database containing 2,989 images representing 7,190 modules across 26 unique make-model combinations [192]. These modules are installed in Florida, Ohio, and Massachusetts, with all images provided in JPG format.

F. Data processing tools and models

Image data provide valuable information on solar module health, which can be harnessed with appropriate processing methods. The tools for pre-processing and analyzing these images enable enhanced visualization and facilitate automated tasks, such as perspective transformation [193], [194]. Key tasks in processing these images include cell classification, defect localization, object segmentation, and power prediction. This section introduces tools and models designed for processing the image data sets mentioned above. A summary of these tools is provided in Table XII.

Function	Tool / Model	Ref |URL	Data I/O	Remark
EL - image preprocessing	PV-Vision	N/A |[193]	Input EL images of original modules; Output well-aligned module images or cropped cell images	Automate background removal, module alignment and cell cropping; Handle ML models
EL - image preprocessing	PVimage	N/A|[194]	Input EL images of original modules; Output well-aligned module images or cropped cell images	Remove barrel distortion, reduce noise, remove background, align modules and crop cells
EL - defect localization	PV-Vision (YOLO v3)	[136] |[193]	Input EL images of aligned modules; Output defective cell position and class with bbox	Can detect cracks, solder disconnection, striation rings, and intra-cell defects
EL - defect segmentation	PV-Vision (UNet)	[154] |[193]	Input EL images of cropped cells; Output object masks in pixel level	Can segment crack masks and extract quantitative crack descriptors
EL - defect segmentation	Deeplab v3 + ResNeT-50	[8] |[168]	Input EL images of cropped cells; Output object masks in pixel level	Can segment masks of nine objects including different types of cracks, interruptions, interconnections, contact corrosion
EL - power prediction	ELPVPOWER	[137] |[176]	Input EL images of aligned modules; Output maximum operation power	
PL - defect	K-means	[183] |[186]	Input PL images of cells; Output I-V parameters	
UVF - defect	OpenUVF	[190] |[189]	Input UVF images of original modules; Output well-aligned module image and defects inside bbox	Enable end-to-end processing that takes UVF photos of original modules and generate analysis report
Visual - module detection	DeepSolar	[144] |[195]	Input satellite images; Output solar adoption trajectories (via segmented module masks)	
Visual - module detection	DeepSolar++	[143] |[149]	Input satellite images; Output solar adoption trajectories (via segmented module masks)	Compared to DeepSolar, DeepSolar++ also uses a diffusion model to understand PV adoption over time
Visual - module detection	3D-PV-Locator	[196] |[197]	Input satellite and 3D building images; Output segmented module masks	Extract the module information from both satellite image and 3D building image
Visual - module detection	PV global inventory	[74] |[85]	Input satellite images; Output segmented module masks	Data collected until 2018; Also contains the installation capacity
Visual - module detection	PV identifier	[73] |[73]	Input satellite images; Output segmented module masks	
Visual - module detection	URSUS-PV	[86] |[87]	Input LiDAR images; Output roof characteristic	For rooftop module image processing

TABLE XII
OVERVIEW OF TOOLS/MODELS FOR IMAGING DATA. SOME PREVIOUS STUDIES THAT DID NOT PUBLISH THE RELATED TOOLS/MODELS ARE NOT SUMMARIZED IN THE TABLE BUT STILL REVIEWED IN THIS PAPER.

1) EL image preprocessing: The quality of solar module images taken in the field or laboratory is often affected by factors such as lens and perspective distortion. In addition, the presence of vegetation, racks, or neighboring modules in the background can complicate the localization of the target module. PV-Vision [193] is a software tool developed to automatically preprocess EL images, preparing them for analysis and model training. This tool includes features for background removal, module alignment, and cropping, along with functions for model training and evaluation. The model weights retrained for cells with thin busbars and multiclassification can be accessed here [159], with retraining and validation code available here [198].

Another tool, PVimage [194], corrects barrel distortion, reduces noise, and removes irrelevant background data. After detecting module edges, it applies a perspective transformation to uniformly orient and planarize the module image, which can then be segmented into individual cells. Unlike PV-Vision, PVimage relies on classical image processing techniques rather than machine learning. Although these tools were initially developed for EL images, their core algorithms can be adapted for use with other imaging techniques, such as PL images.

2) EL defect identification: Numerous studies have focused on automatic defect identification using EL images, covering cell classification, semantic segmentation, and defect detection. Beyond image preprocessing, PV-Vision also supports pipelines for defective cell localization through object detection [136] and crack extraction via semantic segmentation [154]. The model developed by Fioresi *et al.* [8] can differentiate between cracks, contact interruptions, cell interconnect failures, and contact corrosion, using a Deeplabv3 model with a ResNet-50 backbone. This model is available on Github [168]. Another publication [137] comprises both EL images of the modules and the corresponding measured maximum power. This data set was used to develop a fully connected neural network [176] that can predict the maximum power of the module based on the EL images of the modules.

Additional studies are also developing models for power prediction from EL images. Although these studies have yet to publish related data or code, some notable projects include IEK5-Photovoltaik Forschungszentrum Jülich's encoder-decoder deep neural network for detecting shunts and "droplets" in CIGS module images, trained on 142 and 14 annotated images for each defect [169]. The code is available upon request. SNL is working on a variational autoencoder (VAE) for predicting power loss from solar cell crack masks, with the code being released on GitHub [159].

3) IR tools: Currently we cannot find public models or tools for processing IR images. Some previous studies shown in the following proposed workflows, such as IR image classification or object detection, but their codes are not publicly available.

Pierdicca *et al.* [199] analyzed 3,336 thermal IR images using CNN. Specifically, they trained images using the VGG-16 network to categorize images as damaged or normal. To enhance model performance, they applied data augmentation techniques, including rotating and flipping images. Niazi [200] analyzed 375 thermal IR images using a Naive Bayesian classifier to categorize images into three classes. Feature extraction was conducted in Matlab, followed by principal component analysis and classification using pattern recognition. Neither study has published its data or code.

Object detection techniques have also been explored in IR image processing. For example, Shihavuddin *et al.* combined optical and IR images from their published data set [139] to detect defects using EfficientDet (D0-D5) and YOLO (v3-v5) architectures [201]. However, model overfitting was noted due to the small data set size, and while the data set itself is publicly available, the models are not

4) PL tools: Several tools have been applied to analyze PL images in PV research. For example, Laufer *et al.* used K-means clustering to predict I-V parameters based on an open-source PL data set [184], with a tutorial provided as a Jupyter notebook on GitHub [186].

Gao *et al.* utilized a modified YOLOv5 model to localize 13 defect types in silicon solar cells, though the code and model weights are not publicly available [202].

Beyond defect detection, deep learning models have been employed to predict device parameters and losses in silicon cells. Huang *et al.* used a neural network model with PL and EL images as input to estimate voltage and fill factor losses, showing comparable results to Griddler mapping [203].

In another study, Hoffmann *et al.* captured 54 PL images with 2048-pixel resolution to investigate power loss in cells by comparing them with reference EL images [137]. Using a CNN with ResNet18 architecture, they treated power loss estimation as a linear regression problem and observed slight performance improvements when the model was pretrained on EL images rather than on ImageNet data. However, the related code has not been published.

5) Visual tools: Deep learning applications for geolocating solar installations have been prominent in recent studies. For example, Stanford's DeepSolar [144], [204] has been used to analyze residential solar adoption trajectories, with model weights accessible on GitHub [195]. While the training and testing sets for DeepSolar are restricted, DeepSolar++[143] provides time-lapse data and code on GitHub[149]. The same team has worked on validating 3D orientation detection of PV installations using aerial imagery and the German PV registry [196], with code available on GitHub [197].

Kruitwagen *et al.* developed a global PV inventory of installations exceeding 10kW, using a similar approach for data processing and analysis [74], with code published here [85]. Additionally, the PV Identifier [73] includes code as supplementary material. PVNet is another segmentation model for PV sites from satellite images. Although PVNet data are on GitHub [205], the model code itself has not been shared. The uses of these tools include energy prediction, planning, and land assessment, as outlined in related publications [146]. Studies also address pre-installation land use analysis [206] and specific installation types such as floating PV [147], though without shared code or data.

Some other models detect building rooftops for potential solar capacity studies instead of detecting solar modules in satellite imagery [207], [208]. Their codes are not available.

6) UVF tools: UVF imaging in field inspections typically captures multiple modules per image. The low contrast from the UV flash source complicates edge detection, making segmentation models essential for machine learning analysis.

OpenUVF, developed in TensorFlow, segments modules and cells and applies a segmentation model for crack detection using TensorFlow's object detection API and pretrained models [190]. The pipeline involves MATLAB for segmentation and Python for crack detection, with the workflow accessible on GitHub [189]. Another approach utilizes a Mask-RCNN model in Detectron2 for module segmentation in UVF images [188]; however, this project is ongoing, and the tool is not yet publicly available.

7) Image Data Augmentation Tools: Image sets are often unbalanced or lack sufficient samples for effectively training ML models. Image data augmentation encompasses techniques aimed at increasing data set variation and size, categorized broadly into geometric transformations and feature space augmentations. Geometric transformations directly manipulate images through methods such as flipping, cropping, color space adjustment, and blurring. While these transformations are relatively straightforward to implement, it is essential to apply them carefully to avoid obscuring critical features necessary for accurate model training.

Feature space augmentations, in contrast, use generative AI to produce synthetic training samples. Techniques include style transfer, generative adversarial networks (GANs), and VAEs [209].

Romero *et al.* [210] successfully utilized a custom GAN to generate synthetic images of cell defects. Their GAN was trained on 602 cell-level EL images to create EL images with diverse cell defects from random seeds. The original and synthetic images are available at [211]. Each synthetic image was then assigned a maximum power value using a random forest regression model trained on real images. This enriched data set was subsequently used to train a CNN to predict the maximum power of cells based on EL images in [212]. However, the authors had not published their code at the time of this review.

Similarly, Luo *et al.* [213] employed a custom GAN to generate synthetic EL images of defective cells. The GAN was trained on a geometrically augmented EL image data set that included defect classification labels. This geometric augmentation involved image flipping, rotation, and color jittering, increasing the original data set size 11 times. The GAN-generated images created a balanced data set with evenly distributed cell defect types, which was then used to train a CNN classifier for cell defect detection in EL images. Luo *et al.* reported up to a 14% improvement in classification accuracy using this feature space augmented data set. However, the data and code have not been made public.

G. Insights

Most of the models developed for automatic inspection of solar cells are based on the researchers' own data sets. A public and universal data set for solar cell inspection is still missing to benchmark the performance of different models. In addition, many studies do not publish the corresponding code or integrate models into user-friendly tools, limiting accessibility and reproducibility.

For EL image automation, most available data sets contain high-resolution images, while field images, especially those captured in daylight [214] or with InGaAs cameras, are typically lower resolution and have higher noise levels. This discrepancy raises concerns that current pretrained models may perform poorly on these noisier, lower-quality images. Although there are denoising techniques for EL images [215], it remains an open question whether these methods retain sufficient diagnostic information. In addition, the variance of images due to different camera settings requires higher generalization of ML models.

Quantitative prediction of PV system power output based on EL images also remains an unresolved challenge. Future work could focus on developing models capable of predicting I-V parameters directly from EL images, advancing predictive capabilities in PV diagnostics. This approach is currently being pursued for specific PV degradation mechanisms such as cell fracture. However a universal model would require a highly varied dataset in both module/cell type and observed degradation mechanism, with correlated EL and I-V or other performance data.

For visual images, research is still in its early stages. Most current efforts involve extracting geolocations from satellite images and providing the corresponding data and code for model training [143], [74]. As satellite imagery becomes more widely available and offers higher resolutions through subscription services, future work may enable failure detection and potentially continuous monitoring of PV sites from space. High-resolution satellite images could allow image processing techniques, similar to those used in drone imagery [216], to be adapted for fault detection, such as cracked glass or structural damage from environmental impacts such as wind or stalled trackers.

Additional failure types, such as damaged connectors or junction boxes, could also be identified by applying deep learning to ground-based visual images. However, since these components are located on the back of modules, their detection is challenging; in such cases, IR imaging from drones may be better suited to identify overheated components.

VI. PV MATERIALS

In this section, we review bill-of-materials (BOM) data sets for silicon (Si)-based PV modules, which dominate the current market, as well as data sets for perovskite solar cells, a promising next-generation PV technology. Here, the term "PV material" refers to commercialized Si-based modules, unless the term "perovskite" is explicitly mentioned.

Material durability is essential for the long-term reliable operation of PV modules. Installed PV modules are exposed to various environmental stressors, including sunlight, temperature fluctuations, humidity, precipitation, wind, and snow. Over time, these conditions can degrade materials such as encapsulants, polymer backsheets, or even the solar cells themselves. Mechanical stresses, such as wind, snow load or impacts such as hail, also contribute to degradation. These stresses can cause microcracks in cells, damage to the protective glass, or structural issues in the module frame, potentially reducing efficiency, or even leading to electrical failures if not promptly addressed. Comprehensive analysis and modeling of PV materials and degradation mechanisms are crucial for replacing problematic materials and optimizing module design. However, published data sets detailing each component material of PV modules remain limited.

A. Data sources for PV materials

A PV module is composed of several components designed to convert sunlight into electricity efficiently and reliably, as shown in Figure 5(b). A typical silicon photovoltaic module consists primarily of interconnected c-Si solar cells encased in a polymer encapsulant and protected by glass or polymer front- and back-sheets. These components are housed within an aluminum frame for easy structural support and installation. Electrical connections are facilitated by busbars and interconnects, which collect the generated electricity and deliver it to a junction box located on the backside of the module. The junction box contains diodes and connectors to safely manage the electrical output. This multi-component structure of PV modules makes the analysis of material degradation even more complex due to the synergistic effects of different degradation pathways of each component materials. For example, moisture ingress causes the hydrolysis of polymers [217], which releases acetic acid that can corrode metal contacts [218] and embrittle the backsheet [219].

Platform	Data set	Ref | URL	Size	Remarks
NREL	REMPD	[220] |[224]	531 records of materials required for solar power plants	Material inputs needed to build solar power plants of different sizes
PVEL	BOM	N/A |[221]	1094 (general) records of module BOM and test performance	Compiled results from annual PV Module Reliability Scorecards published by PV Evolution Labs including cell and module technologies
LBNL	BOM-TC	[225] |[222]	251 (detailed) records of module BOM and thermal cycling degradation	
NREL	PVDeg	N/A |[223]	small collection of material and degradation properties	Software tools and materials database for typical degradation mechanisms found in PV modules

TABLE XIII
OVERVIEW OF DATA SETS FOR PV MATERIALS

Platform	Data set	Ref |URL	Size	Remarks
Lund University	Perovskite Database Project	[229] |[230]	42,400 perovskite solar cell devices with up to 100 parameters per device.	Materials, fabrication processes, and performance metrics of perovskite solar cells; FAIR data principles, interactive analysis tools.
Boğaziçi University	Long term stability	[226] |[226]	404 cells, around 15 parameters/cell	Data set containing long-term stability data of perovskite solar cells, including materials, deposition methods, and storage conditions; Analysis using ML to investigate stability factors

TABLE XIV
OVERVIEW OF DATA SETS FOR PEROVSKITE MATERIALS

Public access to comprehensive BOM data sets for PV module technologies remains limited, leading to uncertainty for module buyers and PV system owners about the specific materials used. This information is critical for understanding a module's long-term performance, durability, and reliability, as PV modules and their components are subject to various forms of degradation and potential failure over their lifespan.

In the following, we review data sources that provide information on commonly used materials in PV modules, along with sources detailing typical degradation mechanisms and failure modes observed in PV modules over time. Table XIII summarizes the data sets reviewed in this section.

1) Renewable Energy Materials Properties Database: The Renewable Energy Materials Properties Database (REMPD) is a comprehensive repository of materials data relevant to renewable energy applications, including those in wind and solar energy systems [220]. It documents material requirements for four types of solar PV systems: residential, commercial, and utility-scale PV (UPV) systems with c-Si or cadmium telluride (CdTe) modules. The REMPD data are designed to represent PV technologies as of 2022.

2) PV Module BOM and Test Data: The DuraMAT consortium hosts two data sets related to BOM and module degradation. The PV module and test data [221] summarizes the BOM information and accelerated test performance of the modules tested by KIWA PV Evolution Labs (PVEL). PVEL publishes the best performing modules in their annual scorecard reports. Although detailed material information is missing from the data set, the summary spreadsheet provides an overview of different cell and module technologies, as well as scorecard information of the reliability testing.

The second data set was also published in a work that correlated BOM and thermal cycling degradation [222]. This data set contains additional material information as the encapsulate material.

3) PVDeg – Data: PV Degradation Tools (PVDeg) is an open-source software toolkit that allows users to analyze PV module performance degradation and durability comparing laboratory data with field exposure [223]. It also contains a library of material parameters suitable for estimating the durability and reliability of components used in PV panels.

B. Data sources for Perovskite materials

While extensive studies on the stability and reliability of perovskite solar cells are rare, performance data are more readily available [226], [227], [228]. However, large data sets linking materials, processing parameters, and device architectures to stability can be created by mining research literature and compiling results from multiple studies. This section highlights two such data sources that enable insights into the stability of perovskite solar cells through data analysis and machine learning. Table XIV summarizes the data sets reviewed in this section.

1) Lund University (Perovskite Database Project): The Perovskite Database Project is a comprehensive initiative to organize and consolidate data on halide perovskite solar cells, a field that is rapidly advancing [231]. This project analyzes over 15,000 peer-reviewed publications, identified through Web of Science searches for "perovskite solar", with contributions from approximately 100 volunteers and an estimated 5,000 to 10,000 person-hours for manual data extraction [229]. The data set includes 42,000 devices, each characterized by up to 400 device parameters, and is accessible on Zenodo [230] and the project website [232].

Function	Tool / Model	Ref |URL	Remark
Materials degradation	PVDeg	N/A |[223]	Simplify common foundational computational operations for obtaining meteorological data, running the degradation analysis, performing Monte-Carlo simulations, as well as post processing and plotting of the results
BOM analysis	BOM-TC	[225] |[234]	A workflow to correlate BOM and degradation with ML and study the predominant factors with post-hoc interpretation

TABLE XV
OVERVIEW OF TOOLS/MODELS FOR PROCESSING DATA OF SOLAR MODULE MATERIALS.

The data set encompasses materials, synthesis methods, deposition techniques, and performance indicators such as power conversion efficiency (PCE) for perovskite solar cells. It details 5,500 unique device stacks, covering over 400 perovskite compositions, 194 substrates, 1,957 hole transport layers (HTLs), 1,443 electron transport layers (ETLs), and 288 back contact configurations. Stability data are available for 7,400 entries, with approximately 5,500 measurements representing shelf life in the dark and 550 under operational conditions. This rich data set enables analysis of performance and stability across various stack types and allows researchers to investigate the effects of synthesis variations within similarly structured devices.

The Perovskite Database Project adheres to the FAIR data principles [233], Findability, Accessibility, Interoperability, and Reusability, to ensure data utility for the broader research community. Ongoing updates are supported by user contributions [229].

2) Boğaziçi University (Long-term stability data set): Researchers at Boğaziçi University have developed a perovskite solar cell database by compiling data from the literature, similar to the Perovskite Database Project [226]. This data set was manually curated from papers published between 2016 and August 2018, with keywords such as "perovskite solar" and "stability". It includes degradation profiles (*e.g.*, changes in PCE over time) for 404 cells, with 10-15 categorical variables describing each cell. Key parameters include perovskite material type, deposition method, solvents, ETLs, HTLs, back-contact materials, and storage conditions (humidity, temperature, and illumination). The data set is provided as an Excel file in the Supplementary Information of the related paper [226].

C. Data processing tools and models

Currently, very few open-source tools or models are available for processing solar module material data. We summarize two resources in Table XV. Of the reviewed PV material data sets, only PVDeg provides readily usable tools. The PVDeg repository [223] includes functions for calculating material degradation in PV modules under both accelerated and field-testing conditions. It is designed to streamline essential computational tasks, such as obtaining meteorological data, conducting degradation analysis, performing Monte Carlo simulations, and post-processing results for visualization.

The BOM-TC project [225] also offers relevant code and models on GitHub [234]. It utilizes machine learning models to correlate BOM characteristics with degradation and applies post-hoc interpretation techniques to evaluate the impact of primary design factors on durability. However, these codes are provided as "Jupyter notebooks" rather than as a full software package.

To our knowledge, no standard tools exist for analyzing perovskite solar stability as a function of synthesis and processing parameters. Current perovskite data sets are generally intended for users to download in raw formats (*e.g.*, Excel) for customized preprocessing and analysis using their own tools (*e.g.*, R, Python / Pandas, Mathematica). However, the Perovskite Database Project provides some online tools for exploring its data, enabling users to generate interactive, publication-quality plots for various parameter combinations. Examples include swarm plots for open-circuit voltage by HTL choice, scatter plots correlating open-circuit voltage with band gap, and T80 (time to 80% efficiency) stability as a function of publication date [229].

D. Insights

The reviewed PV material data sets were generally created for specific project purposes, and a unified collection of commonly used PV material properties has yet to be developed. The wide range of applications, from high-efficiency material discovery to durable and reliable packaging materials, complicates the creation of a comprehensive database. In addition, new materials often introduce unknown degradation pathways and failure mechanisms. A large, standardized database containing both PV module and array materials and historical performance data could facilitate studies on life cycle and technoeconomic analyses, survival analysis, and even the discovery of durable materials and combinations thereof that withstand the operating environmental conditions of PV modules.

The stability data set from Boğaziçi University was analyzed using various machine learning and statistical learning techniques (association rule mining, decision trees, *etc.*) to determine factors influencing the stability (*i.e.*, T80 value) of perovskite solar cells [235]. Furthermore, a decision tree was trained to classify the degradation results based on various factors such as stored humidity and HTL additive. The results were broken down into recommendations for both regular n-i-p and inverted p-i-n architectures. A subsequent study linked stability to other factors such as hysteresis, reproducibility, and PCE, finding that factors leading to high stability correlate with these other desired outcomes [236].

VII. Discussion

A. Synergistic utilization of various data types

The effective evaluation of PV systems requires an integrated approach that combines diverse data types. By leveraging environmental data, operational metrics, imaging techniques, and materials information, researchers and industry can gain a comprehensive understanding of PV module durability, thus enhancing long-term system reliability.

Environmental data provide a critical context for understanding the external stressors that affect PV systems. By correlating these environmental factors with performance data, such as power production metrics, researchers can assess how conditions accelerate degradation processes, improving predictive models for system lifetimes and optimizing maintenance schedules. PVPRO, for example, has explored this by applying a physical model to environmental factors [3].

Imaging data provides visual insights into internal defects, such as microcracks, corrosion, and hotspots. Combined with performance and environmental data, imaging serves as a powerful diagnostic tool. For instance, EL images can be linked with I-V curve data to quantify the impact of microcracks on power loss, enabling targeted maintenance and improving model predictions for system degradation. Examples of such correlation are found in [237], [158].

Integrating material data with performance and environmental data allows for pinpointing specific degradation pathways across PV module designs. Material characteristics can act as inputs of predictive models that simulate long-term performance under varying conditions. For instance, linking polymer degradation data with performance metrics in humid environments can clarify how specific materials deteriorate over time, aiding the development of models such as PVDeg [223] and supporting more robust material choices in module design.

Furthermore, these diverse data types can also be harnessed to build large-scale multi-agent systems based on large language models (LLMs) which has been explored on energy data in a previous work [238]. By integrating weather data, production metrics and image processing tools, LLMs enhanced with retrieval-augmented generation (RAG) and function calling can potentially process complex PV system data to perform root cause analysis and perform autonomously monitoring, diagnosing and recommending maintenance actions.

B. Data sparsity and fragmentation

The success of machine learning models is heavily dependent on large, high-quality datasets, which are essential for both training and benchmarking. In computer vision, data sets such as ImageNet [239] and CIFAR [240] have been crucial to advancing deep learning models in object recognition, classification, and segmentation. In particular, ImageNet, with more than 14 million images spanning more than 21,000 categories, has become a cornerstone for training and benchmarking cutting-edge vision models.

In contrast, solar module degradation analysis lacks similarly comprehensive standardized data, which hinders machine learning progress. Taking image data as an example, existing data sets, such as those based on EL, UVF, or IR imaging, are limited in both quantity and diversity. While some annotated data sets are available [154], [159], [161], [137], [164], [166], [8], [169], they are relatively small compared to benchmark datasets in computer vision domains. In addition, these data sets often do not cover the wide range of degradation modes, environmental conditions, and module designs that are present in real-world solar installations.

This issue is exacerbated by fragmented data sources. Data sets are typically collected independently by various research groups and companies, leading to inconsistent metadata, such as camera settings and environmental parameters. This variability complicates data integration, limiting the potential for large-scale unified repositories that could serve as community benchmarks. Furthermore, many data sets are distributed across disparate platforms, making them difficult to access and consolidate. As a result, solar degradation research often relies on isolated data sets specific to the original collectors, limiting the generalizability and robustness of machine learning models in tasks such as defect detection and prediction of power output [5], [241], [8].

C. Suggestions for future data publication

Advancing machine learning applications in solar module degradation analysis will require the development of large, diverse, and well-annotated data sets that cover various PV module designs, degradation modes, and testing methods. Certain data types, such as I-V curve data and PL data, remain especially scarce, and we encourage researchers to make these data publicly available for the PV community.

For future data publications, we recommend following the FAIR principles (Findability, Accessibility, Interoperability, and Reuse of digital assets) [242], which provide guidelines for improving the reusability of scientific data. To establish standardized data sets, we suggest the following steps: First, thoroughly clean the data set to remove corrupt entries and ensure data integrity. Second, include an overview table summarizing data set contents for ease of access. Third, clearly label any missing data, such as absent timestamps or missing data pairs (*e.g.*, a missing EL image corresponding to an IV curve) to enhance transparency. Finally, include comprehensive metadata to facilitate data reuse and understanding, while respecting any constraints imposed by nondisclosure agreements (NDAs).

Table XVI summarizes some recommended platforms for publishing data.

Resource	Website	Ownership
DuraMAT Datahub	https://datahub.duramat.org/	Public – DOE/NREL
Open Science Framework	https://help.osf.io/article/218-sharing-data	Public - Center for Open Science
Nature Scientific Data	https://www.nature.com/sdata/	Private - Springer
Data.gov	https://data.gov/	Public - DOE
Open energy data initiative	https://data.openei.org/	Public - DOE
Zenodo	https://zenodo.org/	Public – European Commission
Github large file storage	https://docs.github.com/en/repositories/working-with-files/managing-large-files/about-git-large-file-storage	Private - Github
OpenDataElia	https://opendata.elia.be/pages/home/	Private – Elia Group
Open Data AWS	https://registry.opendata.aws/	Private - AWS

TABLE XVI
OVERVIEW OF DATAHUBS FOR CONTRIBUTING DATA

VIII. CONCLUSION

Comprehensive data sources for PV systems are essential for understanding various degradation mechanisms, modeling the performance of PV systems and optimizing their operational reliability. This work reviews available data sources for PV systems, covering a wide scope of topics including operation environment, PV location, performance data, image inspection, and PV materials. We explore what data are provided, how they were obtained, how they can be accessed, and what tools or models can be utilized to process them. We also provide insights on how they could be used in future data-driven research. We hope that this work can improve the issue of data fragmentation and guide the community to leverage the available data to advance research in the field of PV durability.

ACKNOWLEDGMENT

The authors would like to thank Cliff Hansen, Will Hobbs, Colleen O'Brien, Young Lee and Emma Cattley Cooper for their contributions to the discussions during the initial phase of this project.

Some schematic illustrations and graphical elements of Figure 1, Figure 2 and Figure 5 were sourced from Adobe Stock under an educational license. The authors acknowledge Adobe Stock for providing these visual assets.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work the author(s) used Writefull & Chatgpt in order to check grammar and improve readability. The author(s) did not use this tool/service to directly generate the manuscript but only to polish the manuscript written by the author(s). After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

REFERENCES

[1] I. E. A. (IEA), "Renewables 2023: Analysis and forecasts to 2028," International Energy Agency (IEA), Tech. Rep., 2024. [Online]. Available: https://www.iea.org/reports/renewables-2023
[2] I. R. E. A. (IRENA), "Renewable power generation costs in 2022," International Renewable Energy Agency (IRENA), Tech. Rep., 2023.
[3] B. Li, T. Karin, B. E. Meyers, X. Chen, D. C. Jordan, C. W. Hansen, B. H. King, M. G. Deceglie, and A. Jain, "Determining circuit model parameters from operation data for PV system degradation analysis: PVPRO," *Solar Energy*, vol. 254, pp. 168–181, Apr. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0038092X23001573
[4] S. Deitsch, V. Christlein, S. Berger, C. Buerhop-Lutz, A. Maier, F. Gallwitz, and C. Riess, "Automatic classification of defective photovoltaic module cells in electroluminescence images," *Solar Energy*, vol. 185, pp. 455–468, 2019.
[5] A. M. Karimi, J. S. Fada, M. A. Hossain, S. Yang, T. J. Peshek, J. L. Braid, and R. H. French, "Automated Pipeline for Photovoltaic Module Electroluminescence Image Processing and Degradation Feature Classification," *IEEE Journal of Photovoltaics*, vol. 9, no. 5, pp. 1324–1335, Sep. 2019. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8744467
[6] X. Zhang, Y. Hao, H. Shangguan, P. Zhang, and A. Wang, "Detection of surface defects on solar cells by fusing Multi-channel convolution neural networks," *Infrared Physics & Technology*, vol. 108, p. 103334, 2020.
[7] Y. Zhao, K. Zhan, Z. Wang, and W. Shen, "Deep learning-based automatic detection of multitype defects in photovoltaic modules and application in real production line," *Progress in Photovoltaics: Research and Applications*, vol. 29, no. 4, pp. 471–484, 2021.
[8] J. Fioresi, D. J. Colvin, R. Frota, R. Gupta, M. Li, H. P. Seigneur, S. Vyas, S. Oliveira, M. Shah, and K. O. Davis, "Automated Defect Detection and Localization in Photovoltaic Cells Using Semantic Segmentation of Electroluminescence Images," *IEEE Journal of Photovoltaics*, vol. 12, no. 1, pp. 53–61, Jan. 2022.
[9] J. F. Gaviria, G. Narváez, C. Guillen, L. F. Giraldo, and M. Bressan, "Machine learning in photovoltaic systems: A review," *Renewable Energy*, vol. 196, pp. 298–318, Aug. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960148122009454

[10] A. Mellit and S. A. Kalogirou, "Artificial intelligence techniques for photovoltaic applications: A review," *Progress in Energy and Combustion Science*, vol. 34, no. 5, pp. 574–632, Oct. 2008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360128508000026

[11] A. Youssef, M. El-Telbany, and A. Zekry, "The role of artificial intelligence in photo-voltaic systems design and control: A review," *Renewable and Sustainable Energy Reviews*, vol. 78, pp. 72–79, Oct. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364032117305555

[12] R. Ahmed, V. Sreeram, Y. Mishra, and M. D. Arif, "A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization," *Renewable and Sustainable Energy Reviews*, vol. 124, p. 109792, May 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364032120300885

[13] K. S. Garud, S. Jayaraj, and M.-Y. Lee, "A review on modeling of solar photovoltaic systems using artificial neural networks, fuzzy logic, genetic algorithm and hybrid models," *International Journal of Energy Research*, vol. 45, no. 1, pp. 6–35, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/er.5608

[14] Y. Buratti, G. M. N. Javier, Z. Abdullah-Vetter, P. Dwivedi, and Z. Hameiri, "Machine learning for advanced characterisation of silicon photovoltaics: A comprehensive review of techniques and applications," *Renewable and Sustainable Energy Reviews*, vol. 202, p. 114617, Sep. 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364032124003435

[15] Z. U. Khan, A. D. Khan, K. Khan, S. A. K. AlKhatib, S. Khan, M. Q. Khan, and A. Ullah, "A Review of Degradation and Reliability Analysis of a Solar PV Module," *IEEE Access*, pp. 1–1, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10606239/?arnumber=10606239

[16] K. AbdulMawjood, S. S. Refaat, and W. G. Morsi, "Detection and prediction of faults in photovoltaic arrays: A review," in *2018 IEEE 12th International Conference on Compatibility, Power Electronics and Power Engineering (CPE-POWERENG 2018)*, Apr. 2018, pp. 1–8. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8372609

[17] D. C. Jordan, C. Deline, S. R. Kurtz, G. M. Kimball, and M. Anderson, "Robust PV Degradation Methodology and Application," *IEEE Journal of Photovoltaics*, vol. 8, no. 2, pp. 525–531, Mar. 2018. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8233204

[18] R. Pan, J. Kuitche, and G. Tamizhmani, "Degradation analysis of solar photovoltaic modules: Influence of environmental factor," in *2011 Proceedings - Annual Reliability and Maintainability Symposium*, Jan. 2011, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/5754514

[19] Y. Lyu, A. Fairbrother, M. Gong, J. H. Kim, X. Gu, M. Kempe, S. Julien, K.-T. Wan, S. Napoli, A. Hauser, G. O'Brien, Y. Wang, R. French, L. Bruckman, L. Ji, and K. Boyce, "Impact of environmental variables on the degradation of photovoltaic components and perspectives for the reliability assessment methodology," *Solar Energy*, vol. 199, pp. 425–436, Mar. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0038092X20301237

[20] E. Young, X. He, R. King, and D. Corbus, "A fluid-structure interaction solver for investigating torsional galloping in solar-tracking photovoltaic panel arrays | Journal of Renewable and Sustainable Energy | AIP Publishing," *Journal of Renewable and Sustainable Energy*, vol. 12, no. 6, p. 063503, Nov. 2020. [Online]. Available: https://pubs.aip.org/aip/jrse/article/12/6/063503/824118

[21] A. Bala Subramaniyan, R. Pan, J. Kuitche, and G. TamizhMani, "Quantification of Environmental Effects on PV Module Degradation: A Physics-Based Data-Driven Modeling Method," *IEEE Journal of Photovoltaics*, vol. 8, no. 5, pp. 1289–1296, Sep. 2018. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8410667

[22] T. Gunda and N. Jackson, "Dataset for Evaluation of Extreme Weather Impacts on Utility-Scale Photovoltaic Plant Performance in the United States," 2021. [Online]. Available: https://data.openei.org/submissions/4055

[23] M. Sengupta, Y. Xie, A. Lopez, A. Habte, G. Maclaurin, and J. Shelby, "The national solar radiation data base (NSRDB)," *Renewable and Sustainable Energy Reviews*, vol. 89, pp. 51–60, 2018.

[24] M. Crespi and G. Ferrari, "Reanalysis, or retrospective analsyis, in meteorology," *Open Meteorology Notebook*, vol. 2, 2020.

[25] "NSRDB: National Solar Radiation Database." [Online]. Available: https://nsrdb.nrel.gov/

[26] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellan, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, G. De Chiara, P. Dahlgren, D. Dee, M. Diamantakis, R. Dragani, J. Flemming, R. Forbes, M. Fuentes, A. Geer, L. Haimberger, S. Healy, R. J. Hogan, E. Hólm, M. Janisková, S. Keeley, P. Laloyaux, P. Lopez, C. Lupu, G. Radnoti, P. de Rosnay, I. Rozum, F. Vamborg, S. Villaume, and J.-N. Thépaut, "The ERA5 global reanalysis," *Quarterly Journal of the Royal Meteorological Society*, vol. 146, no. 730, pp. 1999–2049, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803

[27] "ECMWF Reanalysis v5 (ERA5)," Feb. 2020. [Online]. Available: https://www.ecmwf.int/en/forecasts/dataset/ecmwf-reanalysis-v5

[28] R. Gelaro, W. McCarty, M. J. Suárez, R. Todling, A. Molod, L. Takacs, C. A. Randles, A. Darmenov, M. G. Bosilovich, R. Reichle, K. Wargan, L. Coy, R. Cullather, C. Draper, S. Akella, V. Buchard, A. Conaty, A. M. da Silva, W. Gu, G.-K. Kim, R. Koster, R. Lucchesi, D. Merkova, J. E. Nielsen, G. Partyka, S. Pawson, W. Putman, M. Rienecker, S. D. Schubert, M. Sienkiewicz, and B. Zhao, "The modern-era retrospective analysis for research and applications, version 2 (MERRA-2)," *Journal of Climate*, vol. 30, no. 14, pp. 5419 – 5454, 2017. [Online]. Available: https://journals.ametsoc.org/view/journals/clim/30/14/jcli-d-16-0758.1.xml

[29] "Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2)." [Online]. Available: https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/

[30] M. Kanamitsu, W. Ebisuzaki, J. Woollen, S.-K. Yang, J. Hnilo, M. Fiorino, and G. Potter, "NCEP–DOE AMIP-II Reanalysis (R-2)," *Bulletin of the American Meteorological Society*, vol. 83, no. 11, pp. 1631–1644, 2002.

[31] "NCEP/DOE Reanalysis II: NOAA Physical Sciences Laboratory NCEP/DOE Reanalysis II." [Online]. Available: https://psl.noaa.gov/data/gridded/data.ncep.reanalysis2.html

[32] S. Kobayashi, Y. Ota, Y. Harada, A. Ebita, M. Moriya, H. Onoda, K. Onogi, H. Kamahori, C. Kobayashi, H. Endo, Miyaoka, Kengo, and K. Takahashi, "The JRA-55 reanalysis: General specifications and basic characteristics," *Journal of the Meteorological Society of Japan. Ser. II*, vol. 93, no. 1, pp. 5–48, 2015. [Online]. Available: https://www.jstage.jst.go.jp/article/jmsj/93/1/93_2015-001/_article/-char/ja/

[33] "JRA-55 – the Japanese 55-year Reanalysis." [Online]. Available: https://jra.kishou.go.jp/JRA-55/index_en.html#jra-55

[34] U. Pfeifroth, J. Drücke, S. Kothe, J. Trentmann, M. Schröder, and R. Hollmann, "SARAH-3 - satellite-based climate data records of surface solar radiation," *Earth System Science Data Discussions*, pp. 1–28, Apr. 2024. [Online]. Available: https://essd.copernicus.org/preprints/essd-2024-91/

[35] "Surface Radiation Data Set - Heliosat (SARAH) - Edition 3." [Online]. Available: https://navigator.eumetsat.int/product/EO:EUM:DAT:0863

[36] M. Sengupta, A. Habte, Y. Xie, A. Lopez, and G. Buster, "National Solar Radiation Database (NSRDB)," 2018. [Online]. Available: https://data.openei.org/submissions/1

[37] A. K. Heidinger, M. J. Foster, A. Walther, and X. T. Zhao, "The pathfinder atmospheres–extended AVHRR climate dataset," *Bulletin of the American Meteorological Society*, vol. 95, no. 6, pp. 909–922, 2014.

[38] Y. Xie, M. Sengupta, and J. Dudhia, "A Fast All-sky Radiation Model for Solar applications (FARMS): Algorithm and performance evaluation," *Solar Energy*, vol. 135, pp. 435–445, 2016.

[39] European Centre For Medium-Range Weather Forecasts, "ECMWF IFS CY41r2 High-Resolution Operational Forecasts," 2016. [Online]. Available: http://rda.ucar.edu/datasets/ds113.1/

[40] M. Bonavita, E. Hólm, L. Isaksen, and M. Fisher, "The evolution of the ECMWF hybrid data assimilation system," *Quarterly Journal of the Royal Meteorological Society*, vol. 142, no. 694, pp. 287–303, 2016. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.2652

[41] L. Isaksen, M. Bonavita, R. Buizza, M. Fisher, J. Haseler, M. Leutbecher, and L. Raynaud, "Ensemble of data assimilations at ECMWF," Technical Memorandum 636, ECMWF, Reading, UK, Tech. Rep., 2010.

[42] J.-N. Thépaut, D. Dee, R. Engelen, and B. Pinty, "The Copernicus Programme and its Climate Change Service," in *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, Jul. 2018, pp. 1591–1593. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8518067

[43] A. G. Amillo, N. Taylor, A. Fernandez, E. Dunlop, P. Mavrogiorgios, F. Fahl, G. Arcaro, and I. Pinedo, "Adapting pvgis to trends in climate, technology and user needs," in *Proceedings of the 38th European Photovoltaic Solar Energy Conference and Exhibition, Online*, 2021, pp. 6–10.
[44] ——, "Pvgis-sarah2." [Online]. Available: https://joint-research-centre.ec.europa.eu/photovoltaic-geographical-information-system-pvgis/pvgis-data-download/sarah-2-solar-radiation-data_en
[45] "pvlib-pvgis." [Online]. Available: https://pvlib-python.readthedocs.io/en/stable/_modules/pvlib/iotools/pvgis.html
[46] A. Driesse and J. S. Stein, "Global normal spectral irradiance in Albuquerque: a one-year open dataset for PV research," Sandia National Lab. (SNL-NM), Albuquerque, NM (United States), Tech. Rep. SAND2020-12693, Nov. 2020. [Online]. Available: https://www.osti.gov/biblio/1814068
[47] a. driesse, "Global normal spectral irradiance in Albuquerque." [Online]. Available: https://datahub.duramat.org/dataset/gni-spectra-abq
[48] A. Driesse, M. Theristis, and J. S. Stein, "Global horizontal spectral irradiance and module spectral response measurements: an open dataset for PV research," Sandia National Lab. (SNL-CA), Livermore, CA (United States); Sandia National Lab. (SNL-NM), Albuquerque, NM (United States), Tech. Rep. SAND-2023-02045, Apr. 2023. [Online]. Available: https://www.osti.gov/biblio/2293575
[49] A. Driesse, "Global horizontal spectral irradiance in Albuquerque." [Online]. Available: https://datahub.duramat.org/en/dataset/metadata/ghi-spectra-abq
[50] M. Boyd, "Performance Data from the NIST Photovoltaic Arrays and Weather Station," *Journal of Research of the National Institute of Standards and Technology*, vol. 122, pp. 1–4, Nov. 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7339723/
[51] M. Boyd, Tzong Hao Chen, and B. Dougherty, "NIST Campus Photovoltaic (PV) Arrays and Weather Station Data Sets," 2017. [Online]. Available: https://pvdata.nist.gov
[52] NREL, "Nrel wind integration data and tools." [Online]. Available: https://www.nrel.gov/grid/wind-integration-data.html
[53] ——, "Nrel solar integration data and tools." [Online]. Available: https://www.nrel.gov/grid/solar-integration-data.html
[54] N. D. Jackson and T. Gunda, "Evaluation of extreme weather impacts on utility-scale photovoltaic plant performance in the United States," *Applied Energy*, vol. 302, p. 117508, 2021.
[55] NREL, "Nrel baseline measurement system." [Online]. Available: https://midcdmz.nrel.gov/apps/sitehome.pl?site=BMS
[56] ——, "Nrel solar integration national dataset (sind) toolkit." [Online]. Available: https://www.nrel.gov/grid/sind-toolkit.html
[57] M. Hummon, E. Ibanez, G. Brinkman, and D. Lew, "Sub-hour solar data for power system modeling from static spatial variability analysis," National Renewable Energy Lab.(NREL), Golden, CO (United States), Tech. Rep., 2012.
[58] NREL, "Nrel spectral solar radiation data base." [Online]. Available: https://www.nrel.gov/grid/solar-resource/spectral-solar.html
[59] C. J. Riordan, D. R. Myers, and R. L. Hulstrom, "Spectral solar radiation data base documentation," Solar Energy Research Inst.(SERI), Golden, CO (United States), Tech. Rep., 1990.
[60] G. T. Klise, P. H. Kobos, R. R. Hill, C. J. Hamman, V. P. Gupta, B. B.-Y. Yang, and N. Enbar, "PV Reliability Operations and Maintenance (PVROM) Database Initiative: 2014 Project Report," Sandia National Lab. (SNL-NM), Albuquerque, NM (United States); Electric Power Research Institute (EPRI), Palo Alto, CA (United States), Tech. Rep. SAND-2014-20612, Dec. 2014. [Online]. Available: https://www.osti.gov/biblio/1504104
[61] "pvlib-iotools." [Online]. Available: https://pvlib-python.readthedocs.io/en/stable/reference/iotools.html#module-pvlib.iotools
[62] G. M. Yagli, D. Yang, and D. Srinivasan, "Automatic hourly solar forecasting using machine learning models," *Renewable and Sustainable Energy Reviews*, vol. 105, pp. 487–498, 2019.
[63] A. Dairi, F. Harrou, Y. Sun, and S. Khadraoui, "Short-Term Forecasting of Photovoltaic Solar Power Production Using Variational Auto-Encoder Driven Deep Learning Approach," *Applied Sciences*, vol. 10, no. 23, p. 8400, Jan. 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/23/8400
[64] Y. Hatanaka, Y. Glaser, G. Galgon, G. Torri, and P. Sadowski, "Diffusion Models for High-Resolution Solar Forecasts," Jan. 2023. [Online]. Available: http://arxiv.org/abs/2302.00170
[65] Y. Fan, R. Wieser, X. Yu, Y. Wu, L. S. Bruckman, and R. H. French, "Using spatio-temporal graph neural networks to estimate fleet-wide photovoltaic performance degradation patterns," *PLOS ONE*, vol. 19, no. 2, p. e0297445, Feb. 2024. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0297445
[66] J. Chakar, M. Pavlov, Y. Bonnassieux, and J. Badosa, "Determining solar cell parameters and degradation rates from power production data," *Energy Conversion and Management: X*, vol. 15, p. 100270, 2022.
[67] D. Fregosi, R. Dhakal, and D. Widrick, "Evaluating the Use of Satellite Data and Machine Learning Models for PV Performance Monitoring," in *2023 IEEE 50th Photovoltaic Specialists Conference (PVSC)*, Jun. 2023, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10359946
[68] G. Seo, S. Yoon, J. Song, E. Srivastava, and E. Hwang, "Label-Free Fault Detection Scheme for Inverters of PV Systems: Deep Reinforcement Learning-Based Dynamic Threshold," *Applied Sciences*, vol. 13, no. 4, p. 2470, Jan. 2023. [Online]. Available: https://www.mdpi.com/2076-3417/13/4/2470
[69] T. W. David, G. A. Soares, N. Bristow, D. Bagnis, and J. Kettle, "Predicting diurnal outdoor performance and degradation of organic photovoltaics via machine learning; relating degradation to outdoor stress conditions," *Progress in Photovoltaics: Research and Applications*, vol. 29, no. 12, pp. 1274–1284, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/pip.3453
[70] L. Kruitwagen, K. Story, J. Friedrich, L. Byers, S. Skillman, and C. Hepburn, "A global inventory of solar photovoltaic generating units - dataset," Oct. 2021. [Online]. Available: https://zenodo.org/records/5005868
[71] S. Ansari, A. Ayob, M. S. H. Lipu, M. H. M. Saad, and A. Hussain, "A Review of Monitoring Technologies for Solar PV Systems Using Data Processing Modules and Transmission Protocols: Progress, Challenges and Prospects," *Sustainability*, vol. 13, no. 15, p. 8120, Jul. 2021. [Online]. Available: https://www.mdpi.com/2071-1050/13/15/8120
[72] G. R. Venkatakrishnan, R. Rengaraj, S. Tamilselvi, J. Harshini, A. Sahoo, C. A. Saleel, M. Abbas, E. Cuce, C. Jazlyn, S. Shaik, P. M. Cuce, and S. Riffat, "Detection, location, and diagnosis of different faults in large solar PV system—a review," *International Journal of Low-Carbon Technologies*, vol. 18, pp. 659–674, Feb. 2023. [Online]. Available: https://academic.oup.com/ijlct/article/doi/10.1093/ijlct/ctad018/7083000
[73] N. Lu, L. Li, and J. Qin, "PV Identifier: Extraction of small-scale distributed photovoltaics in complex environments from high spatial resolution remote sensing images," *Applied Energy*, vol. 365, p. 123311, Jul. 2024. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306261924006949
[74] L. Kruitwagen, K. T. Story, J. Friedrich, L. Byers, S. Skillman, and C. Hepburn, "A global inventory of photovoltaic solar energy generating units," *Nature*, vol. 598, no. 7882, pp. 604–610, Oct. 2021. [Online]. Available: https://www.nature.com/articles/s41586-021-03957-7
[75] "U.S. Photovoltaic Database." [Online]. Available: https://eerscmap.usgs.gov/uspvdb/
[76] K. S. Fujita, Z. H. Ancona, L. A. Kramer, M. Straka, T. E. Gautreau, D. Robson, C. Garrity, B. Hoen, and J. E. Diffendorfer, "Georectified polygon database of ground-mounted large-scale solar photovoltaic sites in the United States," *Scientific Data*, vol. 10, no. 1, p. 760, Nov. 2023. [Online]. Available: https://www.nature.com/articles/s41597-023-02644-8
[77] "Viewer | USPVDB." [Online]. Available: https://eerscmap.usgs.gov/uspvdb/viewer/#3/37.25/-96.25
[78] D. Stowell, J. Kelly, D. Tanner, J. Taylor, E. Jones, J. Geddes, E. Chalstrey, G. Williams, and J. Clough, "Solar panels and solar farms in the UK - geographic open data (UKPVGeo)," Sep. 2020. [Online]. Available: https://zenodo.org/records/4059881
[79] D. Stowell, J. Kelly, D. Tanner, J. Taylor, E. Jones, J. Geddes, and E. Chalstrey, "A harmonised, high-coverage, open dataset of solar photovoltaic installations in the UK," *Scientific Data*, vol. 7, no. 1, p. 394, Nov. 2020. [Online]. Available: https://www.nature.com/articles/s41597-020-00739-0
[80] "Global Inventory of Solar Energy Installations." [Online]. Available: https://resourcewatch.org/data/explore/ene032-Solar-Plants_1
[81] M. Mooney, "PV rooftop database," Jan. 2016. [Online]. Available: https://data.openei.org/submissions/4
[82] W. Mooney, Meghan, "PV rooftop database for puerto rico (PVRDB-PR)," Dec. 2020. [Online]. Available: https://data.openei.org/submissions/2862
[83] Google, "Google Maps Platform Documentation | Maps Static API." [Online]. Available: https://developers.google.com/maps/documentation/maps-static
[84] "Openstreetmap." [Online]. Available: https://www.openstreetmap.org
[85] L. Kruitwagen, K. Story, J. Frierich, L. Byers, S. Skillman, and C. Hepburn, "A global inventory of solar photovoltaic generating units - code," Oct. 2021. [Online]. Available: https://zenodo.org/records/5045001

[86] F. Rodríguez-Gómez, J. del Campo-Ávila, M. Ferrer-Cuesta, and L. Mora-López, "Data driven tools to assess the location of photovoltaic facilities in urban areas," *Expert Systems with Applications*, vol. 203, p. 117349, Oct. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417422007035
[87] "ursusdm/ursusdm_pv," Apr. 2024. [Online]. Available: https://github.com/ursusdm/ursusdm_pv
[88] "URSUS_pv." [Online]. Available: http://ursus-shiny.uma.es/ursusdm_pv/
[89] "Open Energy Data Initiative (OEDI)." [Online]. Available: https://data.openei.org/submissions/4568
[90] "Explore — Elia Open Data Portal." [Online]. Available: https://opendata.elia.be/explore
[91] L. R. Camargo and J. Schmidt, "Simulation of multi-annual time series of solar photovoltaic power: Is the ERA5-land reanalysis the next big step?" *SUSTAINABLE ENERGY TECHNOLOGIES AND ASSESSMENTS*, vol. 42, p. 100829, Dec. 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S221313882031256X
[92] L. Ramirez Camargo and J. Schmidt, "Time series of electricity output for large grid connected photovoltaic installations in Chile," Jul. 2020. [Online]. Available: https://zenodo.org/records/3939047
[93] C. Gonçalves, R. J. Bessa, and P. Pinson, "A critical overview of privacy-preserving approaches for collaborative forecasting," *International Journal of Forecasting*, vol. 37, no. 1, pp. 322–342, Jan. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016920702030090X
[94] "Geographically distributed solar power time series - CKAN." [Online]. Available: https://rdm.inesctec.pt/dataset/pe-2020-003
[95] S. Silwal, C. Mullican, Y.-A. Chen, A. Ghosh, J. Dilliott, and J. Kleissl, "Open-source multi-year power generation, consumption, and storage data in a microgrid," *Journal of Renewable and Sustainable Energy*, vol. 13, no. 2, p. 025301, Mar. 2021. [Online]. Available: https://doi.org/10.1063/5.0038650
[96] P. Tripathi, R. H. French, E. Barcelos, R. Wieser, L. Bruckman, T. Ciardi, H. H. Aung, K. Hernandez, M. Li, and K. O. Davis, "Data Driven Digital Twin," Jun. 2024. [Online]. Available: https://osf.io/dvrc7/
[97] A. Curran, C. B. Jones, J. L. Braid, J. S. Stein, R. H. French, K. Rath, and W. Oltjen, "US DOE-RTC-BaselineSystems," Mar. 2019. [Online]. Available: https://osf.io/yvzhk/
[98] "DKA Solar Centre." [Online]. Available: https://dkasolarcentre.com.au/download?location=yulara
[99] Z. Xiao, X. Huang, J. Liu, C. Li, and Y. Tai, "A novel method based on time series ensemble model for hourly photovoltaic power prediction," *Energy*, vol. 276, p. 127542, Aug. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360544223009362
[100] "DKA Solar Centre." [Online]. Available: https://dkasolarcentre.com.au/download?location=alice-springs
[101] Z.-F. Liu, X.-R. Chen, Y.-H. Huang, X.-F. Luo, S.-R. Zhang, G.-D. You, X.-Y. Qiang, and Q. Kang, "A novel bimodal feature fusion network-based deep learning model with intelligent fusion gate mechanism for short-term photovoltaic power point-interval forecasting," *Energy*, vol. 303, p. 131947, Sep. 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360544224017201
[102] Y. Nie, X. Li, A. Scott, Y. Sun, V. Venugopal, and A. Brandt, "SKIPP'D: A SKy Images and Photovoltaic Power Generation Dataset for short-term solar forecasting," *Solar Energy*, vol. 255, pp. 171–179, May 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0038092X23002037
[103] Y. Nie, X. Li, and A. Scott, "2017-2019 Sky Images and Photovoltaic Power Generation Dataset for Short-term Solar Forecasting (Stanford Benchmark)," 2022. [Online]. Available: https://purl.stanford.edu/dj417rh1007
[104] "SunSmart Schools – Bringing Solar Power to Florida Schools." [Online]. Available: https://www.sunsmartschools.org/
[105] "Regional test centers for solar technologies." [Online]. Available: https://www.energy.gov/eere/solar/regional-test-centers-solar-technologies
[106] "PV Lifetime Modules." [Online]. Available: https://pvpmc.sandia.gov/pv-research/pv-lifetime-project/pv-lifetime-modules/
[107] B. Marion, A. Anderberg, C. Deline, J. del Cueto, M. Muller, G. Perrin, J. Rodriguez, S. Rummel, T. J. Silverman, F. Vignola, R. Kessler, J. Peterson, S. Barkaszi, M. Jacobs, N. Riedel, L. Pratt, and B. King, "New data set for validating PV module performance models," in *2014 IEEE 40th Photovoltaic Specialist Conference (PVSC)*, Jun. 2014, pp. 1362–1366. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6925171
[108] B. Marion, M. G. Deceglie, and T. J. Silverman, "Analysis of measured photovoltaic module performance for Florida, Oregon, and Colorado locations," *Solar Energy*, vol. 110, pp. 736–744, Dec. 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0038092X14005039
[109] "Data for Validating Models for PV Module Performance - DuraMAT Data Hub." [Online]. Available: https://datahub.duramat.org/dataset/data-for-validating-models-for-pv-module-performance
[110] C. E. Commission, "PV Module List." [Online]. Available: https://www.energy.ca.gov/media/2367
[111] A. C. Killam, J. F. Karas, A. Augusto, and S. G. Bowden, "Monitoring of photovoltaic system performance using outdoor suns-voc," *Joule*, vol. 5, no. 1, pp. 210–227, 2021.
[112] J. L. Braid, J. S. Stein, B. H. King, C. Raupp, J. Mallineni, J. Robinson, and S. Knapp, "Effective irradiance monitoring using reference modules," in *2022 IEEE 49th Photovoltaics Specialists Conference (PVSC)*. IEEE, 2022, pp. 1073–1078.
[113] M. W. Hopwood, J. S. Stein, J. L. Braid, and H. P. Seigneur, "Physics-Based Method for Generating Fully Synthetic IV Curve Training Datasets for Machine Learning Classification of PV Failures," *Energies*, vol. 15, no. 14, p. 5085, Jan. 2022. [Online]. Available: https://www.mdpi.com/1996-1073/15/14/5085
[114] M. Laurino, M. Piliougine, and G. Spagnuolo, "Artificial neural network based photovoltaic module diagnosis by current–voltage curve classification," *Solar Energy*, vol. 236, pp. 383–392, Apr. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0038092X2200144X
[115] B. Li, C. Delpha, A. Migan-Dubois, and D. Diallo, "Fault diagnosis of photovoltaic panels using full I–V characteristics and machine learning techniques," *Energy Conversion and Management*, vol. 248, p. 114785, Nov. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0196890421009614
[116] K. S. Anderson, C. W. Hansen, W. F. Holmgren, A. R. Jensen, M. A. Mikofski, and A. Driesse, "pvlib python: 2023 project update," *Journal of Open Source Software*, vol. 8, no. 92, p. 5994, Dec. 2023. [Online]. Available: https://joss.theoj.org/papers/10.21105/joss.05994
[117] C. A. Gueymard, J. M. Bright, D. Lingfors, A. Habte, and M. Sengupta, "A posteriori clear-sky identification methods in solar irradiance time series: Review and preliminary validation using sky imagers," *RENEWABLE & SUSTAINABLE ENERGY REVIEWS*, vol. 109, pp. 412–427, Jul. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364032119302382
[118] A. Mellit, G. M. Tina, and S. A. Kalogirou, "Fault detection and diagnosis methods for photovoltaic systems: A review," *RENEWABLE & SUSTAINABLE ENERGY REVIEWS*, vol. 91, pp. 1–17, Aug. 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364032118301370
[119] C. Brester, V. Kallio-Myers, A. V. Lindfors, M. Kolehmainen, and H. Niska, "Evaluating neural network models in site-specific solar pv forecasting using numerical weather prediction data and weather observations," *Renewable Energy*, vol. 207, pp. 266–274, 2023.
[120] T. Limouni, R. Yaagoubi, K. Bouziane, K. Guissi, and E. H. Baali, "Accurate one step and multistep forecasting of very short-term pv power using lstm-tcn model," *Renewable Energy*, vol. 205, pp. 1010–1024, 2023.
[121] B. Li, C. W. Hansen, X. Chen, D. Diallo, A. Migan-Dubois, C. Delpha, and A. Jain, "A robust I–V curve correction procedure for degraded photovoltaic modules," *Renewable Energy*, p. 120108, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960148124001733
[122] "solar-data-tools," Jul. 2024. [Online]. Available: https://github.com/slacgismo/solar-data-tools
[123] "pvlib-python," Jul. 2024. [Online]. Available: https://github.com/pvlib/pvlib-python
[124] "rdtools," Jul. 2024. [Online]. Available: https://github.com/NREL/rdtools
[125] "pvpro," Jun. 2024. [Online]. Available: https://github.com/DuraMAT/pvpro
[126] B. Li, "IVcorrection," May 2024. [Online]. Available: https://github.com/lbj2011/IVcorrection
[127] X. Ma, W. Huang, E. Schnabel, M. Köhl, J. Brynjarsdóttir, J. L. Braid, and R. H. French, "Data-Driven I–V Feature Extraction for Photovoltaic Modules," *IEEE Journal of Photovoltaics*, vol. 9, no. 5, pp. 1405–1412, Sep. 2019.
[128] M. Wang, J. Liu, T. J. Burleyson, E. J. Schneller, K. O. Davis, R. H. French, and J. L. Braid, "Analytic Isc–Voc Method and Power Loss Modes From Outdoor Time-Series I–V Curves," *IEEE Journal of Photovoltaics*, vol. 10, no. 5, pp. 1379–1388, Sep. 2020.

[129] "SunsVoc." [Online]. Available: https://cran.r-project.org/web/packages/netSEM/vignettes/IVfeature.html
[130] J. Ranalli and W. B. Hobbs, "Pv plant equipment labels and layouts can be validated by analyzing cloud motion in existing plant measurements," *IEEE Journal of Photovoltaics*, vol. 14, no. 3, pp. 538–548, 2024.
[131] J. Ranalli and W. Hobbs, "Solarspatialtools: A python package for spatial solar energy analyses," *Journal of Open Source Software*, vol. 9, no. 101, p. 6984, 2024.
[132] "Analytics project 2021." [Online]. Available: https://github.com/christinabrester/Analytics-project
[133] M. W. Akram, "Failures of Photovoltaic modules and their Detection: A Review," *Applied Energy*, 2022.
[134] M. Köntges, S. Kurtz, C. E. Packard, U. Jahn, K. A. Berger, K. Kato, T. Friesen, H. Liu, M. Van Iseghem, J. Wohlgemuth, D. Miller, M. Kempe, P. Hacke, F. Reil, N. Bogdanski, W. Herrmann, C. Buerhop-Lutz, G. Razongles, and G. Friesen, "Review of Failures of Photovoltaic Modules," IEA International Energy Agency, Report, 2014. [Online]. Available: https://repository.supsi.ch/9645/
[135] Z. Meng, S. Xu, L. Wang, Y. Gong, X. Zhang, and Y. Zhao, "Defect object detection algorithm for electroluminescence image defects of photovoltaic modules based on deep learning," *Energy Science & Engineering*, vol. 10, no. 3, pp. 800–813, 2022.
[136] X. Chen, T. Karin, and A. Jain, "Automated defect identification in electroluminescence images of solar modules," *Solar Energy*, vol. 242, pp. 20–29, 2022.
[137] M. Hoffmann, C. Buerhop-Lutz, L. Reeb, T. Pickel, T. Winkler, B. Doll, T. Würfl, I. Marius Peters, C. Brabec, A. Maier, and V. Christlein, "Deep-learning-based pipeline for module power prediction from electroluminescense measurements," *Progress in Photovoltaics: Research and Applications*, vol. 29, no. 8, pp. 920–935, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/pip.3416
[138] "OpenUVF-site A images." [Online]. Available: https://www.dropbox.com/sh/tx211rja0uapaf1/AAAD_zD4IF_LZB4fVEZM9LBma?dl=0&e=2
[139] https://datasetninja.com/power-plants-damage detection, "Damage Detection of Power Plants." [Online]. Available: https://datasetninja.com/power-plants-damage-detection
[140] G. Kasmi, Y.-M. Saint-Drenan, D. Trebosc, R. Jolivet, J. Leloux, B. Sarr, and L. Dubus, "A crowdsourced dataset of aerial images with annotated solar photovoltaic arrays and installation metadata," Jul. 2022. [Online]. Available: https://zenodo.org/records/7358126
[141] A. Michail, A. Livera, G. Tziolis, J. L. C. Candás, A. Fernandez, E. A. Yudego, D. F. Martínez, A. Antonopoulos, A. Tripolitsiotis, P. Partsinevelos *et al.*, "A comprehensive review of unmanned aerial vehicle-based approaches to support photovoltaic plant diagnosis," *Heliyon*, 2024.
[142] V. Golovko, S. Bezobrazov, A. Kroshchanka, A. Sachenko, M. Komar, and A. Karachka, "Convolutional neural network based solar photovoltaic panel detection in satellite photos," in *2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, vol. 1. IEEE, 2017, pp. 14–19.
[143] Z. Wang, M.-L. Arlt, C. Zanocco, A. Majumdar, and R. Rajagopal, "DeepSolar++: Understanding residential solar adoption trajectories with computer vision and technology diffusion models," *Joule*, vol. 6, no. 11, pp. 2611–2625, Nov. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2542435122004779
[144] J. Yu, Z. Wang, A. Majumdar, and R. Rajagopal, "DeepSolar: A Machine Learning Framework to Efficiently Construct a Solar Deployment Database in the United States," *Joule*, vol. 2, no. 12, pp. 2605–2617, Dec. 2018. [Online]. Available: https://www.cell.com/joule/abstract/S2542-4351(18)30570-1
[145] K. Perry and C. Campos, "Panel Segmentation: A Python Package for Automated Solar Array Metadata Extraction Using Satellite Imagery," *IEEE Journal of Photovoltaics*, vol. 13, no. 2, pp. 208–212, 2023.
[146] J. Wang, "PVNet: A novel semantic segmentation model for extracting high-quality photovoltaic panels in large-scale systems from high-resolution remote sensing imagery," *International Journal of Applied Earth Observation and Geoinformation*, 2023.
[147] Z. Xia, "Mapping global water-surface photovoltaics with satellite images," *Renewable and Sustainable Energy Reviews*, 2023.
[148] G. Kasmi, Y.-M. Saint-Drenan, D. Trebosc, R. Jolivet, J. Leloux, B. Sarr, and L. Dubus, "A crowdsourced dataset of aerial images with annotated solar photovoltaic arrays and installation metadata," *Scientific Data*, vol. 10, no. 1, p. 59, Jan. 2023. [Online]. Available: https://www.nature.com/articles/s41597-023-01951-4
[149] Z. Wang, "DeepSolar_timelapse," May 2024. [Online]. Available: https://github.com/wangzhecheng/DeepSolar_timelapse
[150] A. Streltsov, "Full Collection: Distributed Solar Photovoltaic Array Location and Extent Data Set for Remote Sensing Object Identification," Jul. 2020. [Online]. Available: https://figshare.com/collections/Full_Collection_Distributed_Solar_Photovoltaic_Array_Location_and_Extent_Data_Set_for_Remote_Sensing_Object_Identification/3255643
[151] K. Perry and C. Campos, "Solar Mounting Configuration Satellite Imagery Data," Jul. 2022.
[152] K. Perry, "satellite_imagery_mounting_config_data_set." [Online]. Available: https://datahub.duramat.org/dataset/solar-mounting-configuration-satellite-imagery-data/resource/620c4cf0-84fa-4f09-a2c5-157a15a7ae3e
[153] M. Guada, Á. Moretón, S. Rodríguez-Conde, L. A. Sánchez, M. Martínez, M. Á. González, J. Jiménez, L. Pérez, V. Parra, and O. Martínez, "Daylight luminescence system for silicon solar panels based on a bias switching method," *Energy Science & Engineering*, vol. 8, no. 11, pp. 3839–3853, 2020.
[154] X. Chen, T. Karin, C. Libby, M. Deceglie, P. Hacke, T. J. Silverman, and A. Jain, "Automatic crack segmentation and feature extraction in electroluminescence images of solar modules," *IEEE Journal of Photovoltaics*, pp. 334–342, 2023.
[155] "LBNL crack segmentation." [Online]. Available: https://datahub.duramat.org/dataset/crack-segmentation
[156] T. J. Silverman, N. Bosco, M. Owen-Bellini, C. Libby, and M. G. Deceglie, "Millions of small pressure cycles drive damage in cracked solar cells," *IEEE Journal of Photovoltaics*, vol. 12, no. 4, pp. 1090–1093, 2022.
[157] "Electroluminescence images from one cell undergoing millions of mechanical cycles - DuraMAT Data Hub." [Online]. Available: https://datahub.duramat.org/dataset/electroluminescence-images-from-one-cell-undergoing-millions-of-mechanical-cycles
[158] C. Libby, B. Paudyal, X. Chen, W. B. Hobbs, D. Fregosi, and A. Jain, "Analysis of PV module power loss and cell crack effects due to accelerated aging tests and field exposure," *IEEE Journal of Photovoltaics*, vol. 13, no. 1, pp. 165–173, 2022.
[159] J. Norman, "pvcracks: crack masks for VAE - DuraMAT Data Hub." [Online]. Available: https://github.com/sandialabs/pvcracks/tree/main/vae
[160] "cracks and n-type field aging." [Online]. Available: https://datahub.duramat.org/project/cracks-and-n-type-field-aging
[161] C. Buerhop-Lutz, S. Deitsch, A. Maier, F. Gallwitz, S. Berger, B. Doll, J. Hauch, C. Camus, and C. Brabec, "A Benchmark for Visual Identification of Defective Solar Cells in Electroluminescence Imagery," in *35th European Photovoltaic Solar Energy Conference and Exhibition; 1287-1289*, 2018. [Online]. Available: https://userarea.eupvsec.org/proceedings/35th-EU-PVSEC-2018/5CV.3.15/
[162] "zae-bayern/elpv-dataset," May 2024. [Online]. Available: https://github.com/zae-bayern/elpv-dataset
[163] M. Hoffmann, C. Buerhop-Lutz, L. Reeb, T. Pickel, T. Winkler, B. Doll, T. Würfl, I. M. Peters, C. J. Brabec, A. Maier, and V. Christlein, "ELPVPower: A dataset for large scale PV power prediction using EL images of cells," Mar. 2022. [Online]. Available: https://data.fz-juelich.de/dataset.xhtml?persistentId=doi:10.26165/JUELICH-DATA/TVWUUP
[164] B. Su, Z. Zhou, and H. Chen, "PVEL-AD: A large-scale open-world dataset for photovoltaic cell anomaly detection," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 1, pp. 404–413, 2022.
[165] B. Su, "Photovoltaic cell anomaly detection dataset," Mar. 2022. [Online]. Available: https://ieee-dataport.org/documents/photovoltaic-cell-anomaly-detection-dataset
[166] L. Pratt, J. Mattheus, and R. Klein, "A benchmark dataset for defect detection and classification in electroluminescence images of PV modules using semantic segmentation," *Systems and Soft Computing*, vol. 5, p. 200048, 2023.
[167] L. Pratt, "TheMakiran/BenchmarkELimages," Mar. 2024. [Online]. Available: https://github.com/TheMakiran/BenchmarkELimages
[168] "UCF-EL-Defect," Apr. 2024. [Online]. Available: https://github.com/ucf-photovoltaics/UCF-EL-Defect

[169] E. Sovetkin, E. J. Achterberg, T. Weber, and B. E. Pieters, "Encoder–Decoder Semantic Segmentation Models for Electroluminescence Images of Thin-Film Photovoltaic Modules," *IEEE Journal of Photovoltaics*, vol. 11, no. 2, pp. 444–452, Mar. 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9296290/
[170] E. J. Achterberg, E. Sovetkin, B. E. Pieters, and T. Weber, "PEARL TF-PV, thin-film modules, shunts and droplets," 2020. [Online]. Available: https://b2share.fz-juelich.de/records/e19fe0f0bc0e4f1b8e05b28e977934c9
[171] R. H. French, A. M. Karimi, and J. L. Braid, "Electroluminescent (EL) Image Dataset of PV Module Under Step-wise Damp Heat Exposures," Jun. 2019. [Online]. Available: https://osf.io/4qrtv/
[172] X. Chen, T. Karin, C. Libby, M. Deceglie, P. Hacke, T. Silverman, A. Gabor, and A. Jain, "A benchmark for crack segmentation in electroluminescence images," 2022.
[173] "predicts2 - DuraMAT Data Hub." [Online]. Available: https://datahub.duramat.org/dataset/predicts2
[174] C. Libby, C. B. Jones, W. Hobbs, and B. Hamzavy, "Improving Solar PV Degradation Prediction Certainty: PREDICTS2 (Final Technical Report)," Electric Power Research Inst.(EPRI), Knoxville, TN (United States), Tech. Rep., 2019.
[175] I. E. Commission, "Terrestrial photovoltaic (PV) modules - Design qualification and type approval - Part 2: Test procedures," International Electrotechnical Commission, Tech. Rep., 2016.
[176] ma0ho, "ma0ho/elpvpower," May 2024. [Online]. Available: https://github.com/ma0ho/elpvpower
[177] "RaptorMaps/InfraredSolarModules," May 2024. [Online]. Available: https://github.com/RaptorMaps/InfraredSolarModules
[178] M. Millendorf, E. Obropta, and N. Vadhavkar, "Infrared solar module dataset for anomaly detection," in *Proc. Int. Conf. Learn. Represent*, 2020.
[179] E. Alfaro-Mejía, H. Loaiza-Correa, E. Franco-Mejía, A. D. Restrepo-Girón, and S. E. Nope-Rodríguez, "Dataset for recognition of snail trails and hot spot failures in monocrystalline Si solar panels," *Data in Brief*, vol. 26, p. 104441, Oct. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352340919307966
[180] E. Alfaro Mejia, H. Loaiza Correa, E. Franco Mejia, A. D. Restrepo Giron, and S. E. Nope Rodriguez, "Photovoltaic system thermal images," Aug. 2019, publisher: Mendeley Data. [Online]. Available: https://data.mendeley.com/datasets/82vzccxb6y/2
[181] "Visualization tools for damage detection of power plants dataset," https://datasetninja.com/power-plants-damage-detection, sep 2024, visited on 2024-09-10. [Online]. Available: https://datasetninja.com/power-plants-damage-detection
[182] B. Doll, J. Hepp, M. Hoffmann, R. Schuler, C. Buerhop-Lutz, I. M. Peters, J. A. Hauch, A. Maier, and C. J. Brabec, "Photoluminescence for Defect Detection on Full-Sized Photovoltaic Modules," *IEEE Journal of Photovoltaics*, vol. 11, no. 6, pp. 1419–1429, Nov. 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9520381/
[183] F. Laufer, S. Ziegler, F. Schackmar, E. A. Moreno Viteri, M. Götz, C. Debus, F. Isensee, and U. W. Paetzold, "Process Insights into Perovskite Thin-Film Photovoltaics from Machine Learning with In Situ Luminescence Data," *Solar RRL*, vol. 7, no. 7, p. 2201114, 2023. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/solr.202201114
[184] ——, "In Situ Photoluminescence Imaging Dataset of Blade-Coated Perovskite Photovoltaics," Jan. 2023. [Online]. Available: https://doi.org/10.5281/zenodo.7503391
[185] T. Silverman, "Pl imagery on 100 fielded modules." [Online]. Available: https://datahub.duramat.org/dataset/metadata/pl-imagery-on-100-fielded-modules
[186] F. Laufer, "AI-InSu-Pero/ML-PerovskitePV-InSituLuminescene," May 2024. [Online]. Available: https://github.com/AI-InSu-Pero/ML-PerovskitePV-InSituLuminescene
[187] M. Kontges, A. Morlier, G. Eder, E. Fleis, B. Kubicek, and J. Lin, "Review: Ultraviolet Fluorescence as Assessment Tool for Photovoltaic Modules," *IEEE Journal of Photovoltaics*, vol. 10, no. 2, pp. 616–633, Mar. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8959389/
[188] D. Colvin, A. M. Gabor, W. C. Oltjen, P. J. Knodle, A. D. Yao, B. A. Thompson, N. Khan, S. Lotfian, J. Raby, A. Jojo, X. Yu, M. Liggett, H. P. Seigneur, R. H. French, L. S. Bruckman, M. Li, and K. O. Davis, "Ultraviolet Fluorescence Imaging for Photovoltaic Module Metrology: Best Practices and Survey of Features Observed in Fielded Modules," Apr. 2024. [Online]. Available: https://www.techrxiv.org/users/773082/articles/860566-ultraviolet-fluorescence-imaging-for-photovoltaic-module-metrology-best-practices-and-survey-of-features-observed-in-fielded-modules?commit=8569ba6e72076b3a865b350f33c4d3419c398660
[189] "southern-company-r-d/OpenUVF," Jun. 2024. [Online]. Available: https://github.com/southern-company-r-d/OpenUVF
[190] B. Gilleland, W. B. Hobbs, and J. B. Richardson, "High Throughput Detection of Cracks and Other Faults in Solar PV Modules Using a High-Power Ultraviolet Fluorescence Imaging System," in *2019 IEEE 46th Photovoltaic Specialists Conference (PVSC)*. Chicago, IL, USA: IEEE, Jun. 2019, pp. 2575–2582. [Online]. Available: https://ieeexplore.ieee.org/document/8981262/
[191] "Publications." [Online]. Available: https://brightspotautomation.com/publications/
[192] "OSF | Ultraviolet Fluorescence Images of Modules in the Field." [Online]. Available: https://osf.io/t8x9k/
[193] X. Chen, "pv-vision," 2022. [Online]. Available: https://github.com/hackingmaterials/pv-vision
[194] "cwru-sdle/pvimage," Jun. 2024. [Online]. Available: https://github.com/cwru-sdle/pvimage
[195] Z. Wang, "wangzhecheng/DeepSolar," Jul. 2024. [Online]. Available: https://github.com/wangzhecheng/DeepSolar
[196] K. Mayer, B. Rausch, M.-L. Arlt, G. Gust, Z. Wang, D. Neumann, and R. Rajagopal, "3D-PV-Locator: Large-scale detection of rooftop-mounted photovoltaic systems in 3D," *Applied Energy*, vol. 310, p. 118469, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261921016937
[197] Kevin, "kdmayer/3D-PV-Locator," Mar. 2024. [Online]. Available: https://github.com/kdmayer/3D-PV-Locator
[198] "sandialabs/pvcracks." [Online]. Available: https://github.com/sandialabs/pvcracks/tree/main/retrain
[199] R. Pierdicca, E. S. Malinverni, F. Piccinini, M. Paolanti, A. Felicetti, and P. Zingaretti, "Deep Convolutional Neural Network for Automatic Detection of Damaged Photovoltaic Cells," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLII-2, pp. 893–900, May 2018. [Online]. Available: https://isprs-archives.copernicus.org/articles/XLII-2/893/2018/isprs-archives-XLII-2-893-2018.html
[200] K. A. K. Niazi, W. Akhtar, H. A. Khan, Y. Yang, and S. Athar, "Hotspot diagnosis for solar photovoltaic modules using a Naive Bayes classifier," *Solar Energy*, vol. 190, pp. 34–43, Sep. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0038092X19307340
[201] A. Shihavuddin, M. R. A. Rashid, M. H. Maruf, M. A. Hasan, M. A. u. Haq, R. H. Ashique, and A. A. Mansur, "Image based surface damage detection of renewable energy installations using a unified deep learning approach," *Energy Reports*, vol. 7, pp. 4566–4576, Nov. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352484721005102
[202] M. Gao, Y. Xie, P. Song, J. Qian, X. Sun, and J. Liu, "A Definition Rule for Defect Classification and Grading of Solar Cells Photoluminescence Feature Images and Estimation of CNN-Based Automatic Defect Detection Method," *Crystals*, vol. 13, no. 5, p. 819, May 2023. [Online]. Available: https://www.mdpi.com/2073-4352/13/5/819
[203] Z. Huang, J. W. Ho, K. B. Choi, A. Danner, and K. S. Chan, "Predicting Loss Analysis from Luminescence Images in Si Solar Cells with Convolutional Neural Networks," *Solar RRL*, vol. 7, no. 23, p. 2300396, 2023. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/solr.202300396
[204] "Stanford DeepSolar." [Online]. Available: https://deepsolar.web.app/
[205] Wang_RS, "RS-Wangjx/PVP-Dataset," Jul. 2024. [Online]. Available: https://github.com/RS-Wangjx/PVP-Dataset
[206] Z. Xia, Y. Li, R. Chen, D. Sengupta, X. Guo, B. Xiong, and Y. Niu, "Mapping the rapid development of photovoltaic power stations in northwestern China using remote sensing," *Energy Reports*, 2022.
[207] H. Ren, C. Xu, Z. Ma, and Y. Sun, "A novel 3D-geographic information system and deep learning integrated approach for high-accuracy building rooftop solar energy potential characterization of high-density cities," *Applied Energy*, vol. 306, p. 117985, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261921012885

[208] T. Zhong, Z. Zhang, M. Chen, K. Zhang, Z. Zhou, R. Zhu, Y. Wang, G. Lü, and J. Yan, "A city-scale estimation of rooftop solar photovoltaic potential based on deep learning," *Applied Energy*, vol. 298, p. 117132, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261921005729
[209] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.
[210] H. F. M. Romero, L. Hernández-Callejo, M. Á. G. Rebollo, V. Cardeñoso-Payo, V. A. Gómez, H. J. Bello, R. T. Moyo, and J. I. M. Aragonés, "Synthetic dataset of electroluminescence images of photovoltaic cells by deep convolutional generative adversarial networks," *Sustainability*, vol. 15, no. 9, p. 7175, 2023.
[211] hectorfelipe98, "hectorfelipe98/Synthetic-PV-cell-dataset," May 2024. [Online]. Available: https://github.com/hectorfelipe98/Synthetic-PV-cell-dataset
[212] H. F. M. Romero, L. Hernández-Callejo, M. Á. G. Rebollo, V. Cardeñoso-Payo, V. A. Gómez, J. I. M. Aragonés, and R. T. Moyo, "Optimized estimator of the output power of pv cells using el images and i–v curves," *Solar Energy*, vol. 265, p. 112089, 2023.
[213] Z. Luo, S. Cheng, and Q. Zheng, "GAN-based augmentation for improving CNN performance of classification of defective photovoltaic module cells in electroluminescence images," in *IOP conference series: Earth and environmental science*, vol. 354. IOP Publishing, 2019, p. 012106.
[214] G. Alves dos Reis Benatto, C. Mantel, S. Spataru, A. A. Santamaria Lancia, N. Riedel, S. Thorsteinsson, P. B. Poulsen, H. Parikh, S. Forchhammer, and D. Sera, "Drone-Based Daylight Electroluminescence Imaging of PV Modules," *IEEE Journal of Photovoltaics*, vol. 10, no. 3, pp. 872–877, May 2020. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9042865
[215] G. Liu, P. Dwivedi, T. Trupke, and Z. Hameiri, "Deep Learning Model to Denoise Luminescence Images of Silicon Solar Cells," *Advanced Science*, vol. 10, no. 18, p. 2300206, 2023. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/advs.202300206
[216] X. Li, Q. Yang, Z. Chen, X. Luo, and W. Yan, "Visible defects detection based on UAV-based inspection in large-scale photovoltaic systems," *IET Renewable Power Generation*, vol. 11, no. 10, pp. 1234–1244, 2017. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-rpg.2017.0001
[217] O. K. Segbefia, A. G. Imenes, and T. O. Saetre, "Moisture ingress in photovoltaic modules: A review," *Solar Energy*, vol. 224, pp. 889–906, 2021.
[218] J. Karas, L. Michaelson, K. Munoz, M. Jobayer Hossain, E. Schneller, K. O. Davis, S. Bowden, and A. Augusto, "Degradation of copper-plated silicon solar cells with damp heat stress," *Progress in Photovoltaics: Research and Applications*, vol. 28, no. 11, pp. 1175–1186, 2020.
[219] Y. Lyu, A. Fairbrother, M. Gong, J. H. Kim, A. Hauser, G. O'Brien, and X. Gu, "Drivers for the cracking of multilayer polyamide-based backsheets in field photovoltaic modules: in-depth degradation mapping analysis," *Progress in Photovoltaics: Research and Applications*, vol. 28, no. 7, pp. 704–716, 2020.
[220] A. Cooperman, A. Eberle, D. Hettinger, M. Marquis, B. Smith, R. F. Tusing, and J. Walzberg, "Renewable energy materials properties database: Summary," Aug. 2023. [Online]. Available: https://www.osti.gov/biblio/1995804
[221] J. Karas, "PV module BOM and test data," Dec. 2023. [Online]. Available: https://www.osti.gov/biblio/2229021
[222] "BOM thermal cycling degradation - DuraMAT Data Hub." [Online]. Available: https://datahub.duramat.org/dataset/bom_thermal_cycling_degradation
[223] M. Springer, M. Brown, S. Ovaitt, M. D. Kempe, T. Ford, J. Karas, M. Campanelli, D. M. Holsapple, and K. Anderson, "NREL/PVDegradationTools: 0.3.2," May 2024. [Online]. Available: https://doi.org/10.5281/zenodo.11123249
[224] "REMPD Background Overview." [Online]. Available: https://apps.openei.org/REMPD/overview/background
[225] X. Chen, T. Karin, and A. Jain, "Analyzing the Impact of Design Factors on Solar Module Thermomechanical Durability Using Interpretable Machine Learning Techniques," May 2024. [Online]. Available: http://arxiv.org/abs/2402.11911
[226] Ç. Odabaşı and R. Yıldırım, "Performance analysis of perovskite solar cells in 2013–2018 using machine-learning tools," *Nano Energy*, vol. 56, pp. 770–791, Feb. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2211285518308802
[227] B. Yılmaz and R. Yıldırım, "Critical review of machine learning applications in perovskite solar research," *Nano Energy*, vol. 80, p. 105546, Feb. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2211285520311204
[228] Y. Lu, D. Wei, W. Liu, J. Meng, X. Huo, Y. Zhang, Z. Liang, B. Qiao, S. Zhao, D. Song, and Z. Xu, "Predicting the device performance of the perovskite solar cells from the experimental parameters through machine learning of existing experimental results," *Journal of Energy Chemistry*, vol. 77, pp. 200–208, Feb. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2095495622005642
[229] E. Unger and T. J. Jacobsson, "The Perovskite Database Project: A Perspective on Collective Data Sharing," *ACS Energy Letters*, vol. 7, no. 3, pp. 1240–1245, Mar. 2022. [Online]. Available: https://doi.org/10.1021/acsenergylett.2c00330
[230] T. J. Jacobsson, A. Hultqvist, A. García-Fernández, A. Anand, A. Al-Ashouri, A. Hagfeldt, A. Crovetto, A. Abate, A. G. Ricciardulli, A. Vijayan, A. Kulkarni, A. Y. Anderson, B. P. Darwich, B. Yang, B. L. Coles, C. A. R. Perini, C. Rehermann, D. Ramirez, D. Fairen-Jimenez, D. Di Girolamo, D. Jia, E. Avila, E. J. Juarez-Perez, F. Baumann, F. Mathies, G. S. A. González, G. Boschloo, G. Nasti, G. Paramasivam, G. Martínez-Denegri, H. Näsström, H. Michaels, H. Köbler, H. Wu, I. Benesperi, M. I. Dar, I. Bayrak Pehlivan, I. E. Gould, J. N. Vagott, J. Dagar, J. Kettle, J. Yang, J. Li, J. A. Smith, J. Pascual, J. J. Jerónimo-Rendón, J. F. Montoya, J.-P. Correa-Baena, J. Qiu, J. Wang, K. Sveinbjörnsson, K. Hirselandt, K. Dey, K. Frohna, L. Mathies, L. A. Castriotta, M. H. Aldamasy, M. Vasquez-Montoya, M. A. Ruiz-Preciado, M. A. Flatken, M. V. Khenkin, M. Grischek, M. Kedia, M. Saliba, M. Anaya, M. Veldhoen, N. Arora, O. Shargaieva, O. Maus, O. S. Game, O. Yudilevich, P. Fassl, Q. Zhou, R. Betancur, R. Munir, R. Patidar, S. D. Stranks, S. Alam, S. Kar, T. Unold, T. Abzieher, T. Edvinsson, T. W. David, U. W. Paetzold, W. Zia, W. Fu, W. Zuo, V. R. F. Schröder, W. Tress, X. Zhang, Y.-H. Chiang, Z. Iqbal, Z. Xie, and E. Unger, "The Perovskite Database (Archive)," Jan. 2022. [Online]. Available: https://zenodo.org/records/5837035
[231] ——, "An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles," *Nature Energy*, vol. 7, no. 1, pp. 107–115, Jan. 2022. [Online]. Available: https://www.nature.com/articles/s41560-021-00941-3
[232] "The Perovskite Database." [Online]. Available: https://www.perovskitedatabase.com/
[233] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data*, vol. 3, no. 1, p. 160018, Mar. 2016. [Online]. Available: https://www.nature.com/articles/sdata201618
[234] "DuraMAT/bom_analysis," Feb. 2024. [Online]. Available: https://github.com/DuraMAT/bom_analysis
[235] Ç. Odabaşı and R. Yıldırım, "Machine learning analysis on stability of perovskite solar cells," *Solar Energy Materials and Solar Cells*, vol. 205, p. 110284, Feb. 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0927024819306130
[236] ——, "Assessment of reproducibility, hysteresis, and stability relations in perovskite solar cells using machine learning," *Energy Technology*, vol. 8, no. 12, p. 1901449, 2020.
[237] J. S. Fada, A. J. Loach, A. J. Curran, J. L. Braid, S. Yang, T. J. Peshek, and R. H. French, "Correlation of iv curve parameters with module-level electroluminescent image data over 3000 hours damp-heat exposure," in *2017 IEEE 44th Photovoltaic Specialist Conference (PVSC)*. IEEE, 2017, pp. 2697–2701.
[238] S. Majumder, L. Dong, F. Doudi, Y. Cai, C. Tian, D. Kalathil, K. Ding, A. A. Thatte, N. Li, and L. Xie, "Exploring the capabilities and limitations of large language models in the electric energy sector," *Joule*, vol. 8, no. 6, pp. 1544–1549, 2024.
[239] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 248–255. [Online]. Available: https://ieeexplore.ieee.org/document/5206848
[240] K. Alex, "Learning multiple layers of features from tiny images," *https://www. cs. toronto. edu/kriz/learning-features-2009-TR. pdf*, 2009.

[241] M. Turek and M. Meusel, “Automated classification of electroluminescence images using artificial neural networks in correlation to solar cell performance parameters,” *Solar Energy Materials and Solar Cells*, vol. 260, p. 112483, 2023.

[242] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne *et al.*, “The fair guiding principles for scientific data management and stewardship,” *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.