IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 30, 2020, accepted November 15, 2020, date of publication November 26, 2020,
date of current version December 8, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3039182

A Machine Learning Evaluation of Maintenance Records for Common Failure Modes in PV Inverters

THUSHARA GUNDA[1], SEAN HACKETT[2], LAURA KRAUS[3], CHRISTOPHER DOWNS[4], RYAN JONES[5], CHRISTOPHER MCNALLEY[6], MICHAEL BOLEN[2], AND ANDY WALKER[7]

[1]Sandia National Laboratories, Albuquerque, NM 87123, USA
[2]Electric Power Research Institute, Palo Alto, CA 94304, USA
[3]Strata Solar, Durham, NC 27701, USA
[4]Cypress Creek Renewables, Durham, NC 27713, USA
[5]CD Arevon, Scottsdale, AZ 85258, USA
[6]Consumers Energy Renewable Generation Operations, Saginaw, MI 48601, USA
[7]National Renewable Energy Laboratory, Golden, CO 80401, USA

Corresponding author: Thushara Gunda (tgunda@sandia.gov)

This material is based upon work supported by the Electric Power Research Institute and the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy - Solar Energy Technologies Office under Agreement Number 34172 and as part of the Durable Modules Consortium (DuraMAT), an Energy Materials Network Consortium.

ABSTRACT Inverters are a leading source of hardware failures and contribute to significant energy losses at photovoltaic (PV) sites. An understanding of failure modes within inverters requires evaluation of a dataset that captures insights from multiple characterization techniques (including field diagnostics, production data analysis, and current-voltage curves). One readily available dataset that can be leveraged to support such an evaluation are maintenance records, which are used to log all site-related technician activities, but vary in structuring of information. Using machine learning, this analysis evaluated a database of 55,000 maintenance records across 800+ sites to identify inverter-related records and consistently categorize them to gain insight into common failure modes within this critical asset. Communications, ground faults, heat management systems, and insulated gate bipolar transistors emerge as the most frequently discussed inverter subsystems. Further evaluation of these failure modes identified distinct variations in failure frequencies over time and across inverter types, with communication failures occurring more frequently in early years. Increased understanding of these failure patterns can inform ongoing PV system reliability activities, including simulation analyses, spare parts inventory management, cost estimates for operations and maintenance, and development of standards for inverter testing. Advanced implementations of machine learning techniques coupled with standardization of asset labels and descriptions can extend these insights into actionable information that can support development of algorithms for condition-based maintenance, which could further reduce failures and associated energy losses at PV sites.

INDEX TERMS Inverters, machine learning, natural language processing, photovoltaics, failures, weibull.

I. INTRODUCTION

Renewable energy has demonstrated significant growth over the last decade, with quarterly utility-scale production – from wind, solar, and hydro – topping coal-fired generation for the first time in the U.S. in 2020 [1]. Solar capacity deployment comprised 30% of all new builds in the U.S. over the last 5 years and as of 2020, is generating 2.6% of annual U.S. electricity on average [2]. With increasing penetration onto the grid, reliability of these renewable systems (i.e., maintenance of functionality over time [3]) becomes increasing important for ensuring affordability and grid stability [4]. The cost of maintenance balanced against energy production – or lack thereof – is a trade-off the plant owner must consider for achieving the targeted levelized cost of electricity for the system, which comprises lifetime costs (such as upfront capital and maintenance) divided by lifetime energy [5].

There have been marked improvements in the reliability of photovoltaic (PV) modules but balance of systems components warrant further attention [6]. A review of current literature has identified that inverters, in particular, have been associated with a large fraction of PV system repair events.

The associate editor coordinating the review of this manuscript and approving it for publication was Enamul Haque.

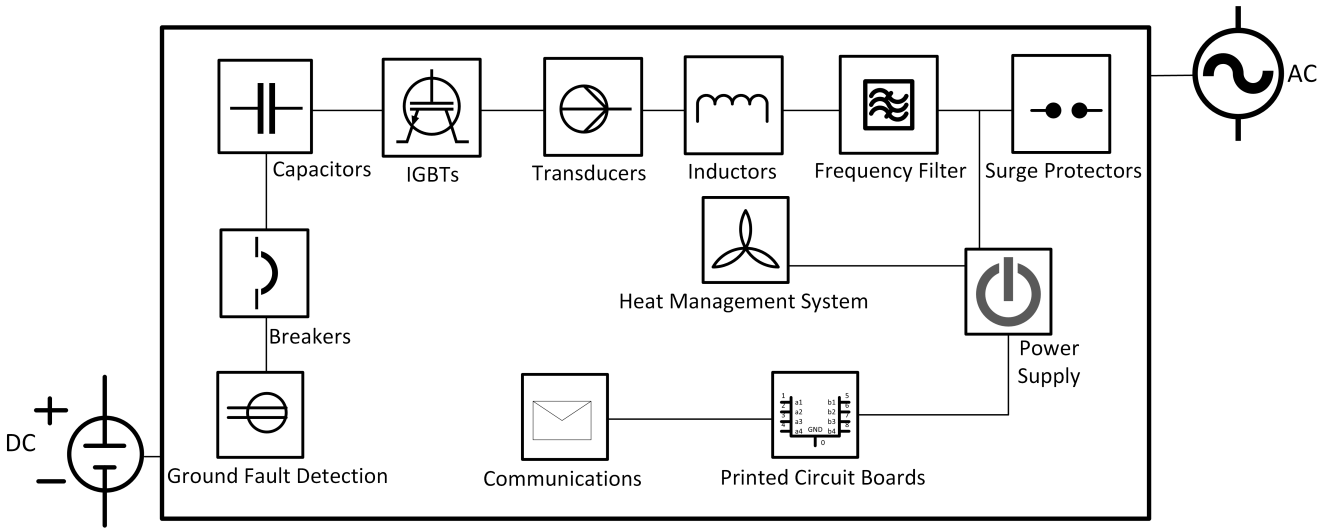


FIGURE 1. **Inverter Subsystems. IGBTs are insulated gate bipolar transistors.**

In an analysis of 3500 computerized maintenance management system (CMMS) records, inverters accounted for 43% of the records and 36% of the energy loss between Jan 2010-Mar 2012 for 350 systems [7]. Similarly, a review of annual performance reports from 100,000 PV systems installed as part of the U.S. Department of the Treasury Section 1603 Program identified inverters as leading hardware failures [4]. Inverter reliability plays a critical role in PV plant profitability since inverter failures lead to either reduced or no energy production; a recent industry report attributed 25% of lost revenue to inverter availability [8].

The high frequency of inverter failures is attributed to the multiple subsystems (with little redundancy in power electronics) that support a multitude of functions in harsh environmental conditions. Inverters set the voltage to maximize power from the PV collector field, convert direct current (DC) to alternating current (AC), interface with the local utility grid, measure and communicate energy production data, and shut down PV systems during unsafe conditions [3], [9]. These functions are supported by multiple subsystems within inverters - such as breakers, capacitors, heat management, ground fault detectors, power supply, and many others (Figure 1) - each of which is subject to failures.

Reliability analyses of PV inverters have, thus far, evaluated the impact of array sizing on inverter lifetime [10], the impact of different maintenance strategies and frequencies on economic return [11], [12], and development of fault-tolerant topologies [13]. Field-based failure information has also been used to generate failure rates or probability indicators [14]–[16]. The sensitivity of PV performance and reliability analyses to data fidelity has also prompted researchers to improve data processing methodologies [17] and data quality activities [18]. Knowing the underlying source of field data issues (e.g., communications outage vs. production outage) can help inform how invalid values or dataset reconstruction activities are handled [19].

However, integration of field failure insights into production assessments is challenging since there is no single way to detect faults at PV sites [20]. Instead, multiple datasets and techniques are leveraged to understand inverter failures, including statistical evaluation of one-diode models [21], wavelet analyses [22], [23], Fourier image reconstruction of electroluminescene images [24] and neural network-based classification of specific datasets (such as current-voltage curves [25]) or components (such as circuits [26]). Information from text-based sources (such as maintenance records) can also provide a lot of insight into failure patterns [4], [7], [19] but the diversity in these records has led to limited utilization of this information for reliability analyses. Thus far, such analyses have been limited to either individual plants [14], [15] or for an individual company's fleet [7], [16].

This analysis addresses this knowledge gap by leveraging a large database of 55,000 corrective maintenance records across 880 sites from multiple PV owners and operators within the United States to gain insight into common failure modes within inverters. The novelty of this work lies in the utilization of machine learning (ML) to consistently classify inverter-related records across multiple CMMS and identify patterns in relative failure rates across inverter subsystems. These findings help inform a number of ongoing activities focused on further reducing the LCOE of PV electricity, including inverter reliability simulation analyses, spare parts inventory management, cost model estimates for operations and maintenance (O&M), and development of standards for inverter testing. Further, this work demonstrates the utility of ML for diversifying the types of data - beyond numerical data [27]–[29] to text-based information - considered for reliability and failure analyses.

II. METHODOLOGY

Maintenance records were collected from multiple sites across the PV industry. These records were then processed using ML in two ways: 1) to consistently identify records related to inverters and 2) to group inverter records into subcategories (Figure 2). In this analysis, we consider an inverter failure to be any event that triggers a ticket in the

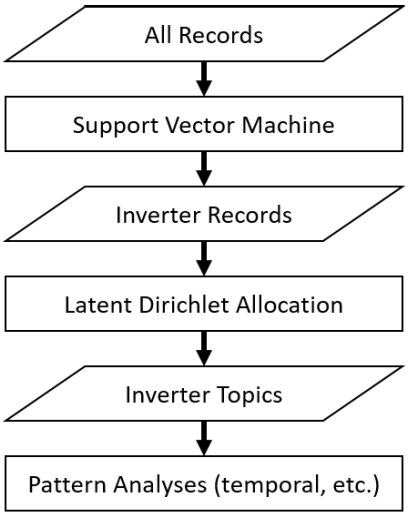


FIGURE 2. Dataset Processing and Analysis.

TABLE 1. Summary of Corrective Maintenance Dataset.

Attribute	All Records	Inverter Records
Number of Partners	6	5
Number of Records	54,909	18,014
Number of Sites	880	677
Aggregate DC capacity (GW)	4.9	4.8
Aggregate AC capacity (GW)	3.8	3.7
COD Range	2008-2019	2008-2019
Number of States	26	25
Number of Climate Zones	4	4
Records Date Range	Feb 2011 - Feb 2020	Feb 2011 - Feb 2020

CMMS [30]. The resulting grouped inverter-related records were then analyzed for patterns, including trends in time, space, and failure rates. The following subsections provide additional details about the dataset, ML implementations, and subsequent pattern analyses.

A. DATASET

The analyzed dataset consists of 55,000 corrective (or reactive) maintenance (CM) records collected from 880 sites owned-operated by 6 industry partners across the U.S. (Table 1). CM records capture details about repair needs associated with unplanned events, such as troubleshooting communications, replacing fuses, resetting inverters, and replacing inverter subcomponents [31]. The sites within the database range in commercial operation dates (COD), from 2008 to 2019 (Figure A1). The sites represent a total capacity of 4.9 gigawatts (GW) in DC and 3.8 GW in AC, with DC:AC ratios ranging between 1 to 1.5.

Geographically, these sites span 26 U.S. states, with a significant portion of the sites (based on capacity) located in North Carolina, California, and Texas (Figure 3). The sites span four climate regions (arid, tropical, temperate, and cold regions) per the Köppen Climate Classifications [32], with a majority of the sites in the temperate non-dry climate zone (Figure A2). The CM records contain varying levels of detail, but generally, all contain information about the specific site as well as the time and description of the event. Site details also included information about the types of inverters present at each site: central (strings of modules aggregated in [re]combiner boxes before entering the inverter), string (strings of modules feed the inverter directly without use of combiner boxes), micro (inverters installed on each module), or some mixed combination of central and string inverters used at a single site (Figure A3).

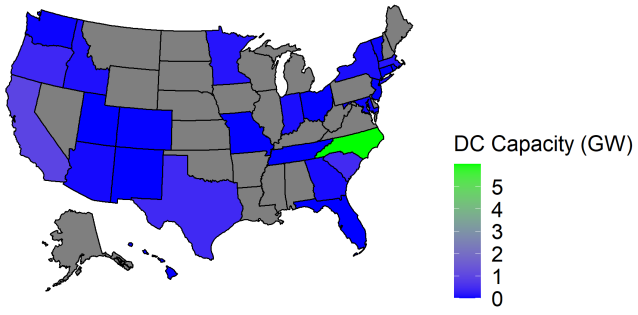


FIGURE 3. Location of PV Sites Represented within the Dataset. The dataset contains a high concentration of sites (based on capacity) within North Carolina, California, and Texas.

B. MACHINE LEARNING

Understanding the specific failures and patterns in inverter subsystems has been challenging due to diverse (non-standard) event capture practices within CMMSs. Each PV fleet owner records and categorizes failures in whatever manner and level of detail they feel is appropriate. Furthermore, there can be inconsistency in type and detail of information recorded amongst maintenance technicians within the same company. To date, CMMS analysis either depended on existing classifications made by technicians [7] or used manual categorization of entries (using either key term identification or classification of individual entries) [4], [33]. However, these approaches are time consuming and not reproducible. Instead in this work, ML techniques were introduced to the analysis to enable a more efficient and consistent identification and classification approach of these maintenance records. The supervised algorithm executed quickly (<3 seconds) while the unsupervised algorithm took 4673 seconds to execute; both algorithms were executed on a machine with an Intel Xeon CPU E5-1630 v3 processer with 48 GB RAM.

1) IDENTIFICATION OF INVERTER FAILURE RECORDS

The CM records generally contain details about the specific asset (or equipment) associated with a failure, such as inverters, trackers, transformers, or the overall facility (Figure A4). Most of the collected records contained a label indicating the type of asset to which a CM record pertains. However, 15% of the records were missing these specific details. These entries were gap-filled using a supervised ML approach, where the variable being predicted is the asset and the variable being used for prediction is the description of the event from the maintenance records (Table 2). The implementation involved: 1) converting text-based description information into a numerical representation using term frequency-inverse document frequency (TF-IDF) and then 2) applying a support vector machine (SVM) algorithm to generate the output of

TABLE 2. Example Records for ML. For the supervised SVM implementation, descriptions were transformed using TF-IDF to predict missing assets. For the unsupervised LDA implementation, inverter descriptions were analyzed to generate topics.

Alg.	Asset	Description
SVM	Inverter	Inverter offline due to failure
	Tracker	Many trackers time drifted, leading to shading during backtracking
	Facility	C4 could not remotely access SCADA via the remote desktop connection...troubleshooting a cell modem issue
	Transformer	Transformer offline due to internal failure
	Combiner	CB 2.3 went offline around 2:00 PM on 25-Jan
	Other	Comms - Contact overdue since 3 days (5/14/2016 9:03:34 AM)

Alg.	Topic(s)	Description
LDA	IGBT	The alarm was showing a PEBB 2 IGBT Failure.
	Coolant	Recharged the cooling systems to a static 30psi and monitored with the pump on.
	Ground Faults, Fuses	Inverter is offline with Array Fault: GFDI has tripped fault. Field Wiring Repair
	Communications, Unknown/cycle, Offline	Inverter down to fix communications problem on inverter 2.1. Power Cycle.

interest. The TF-IDF transformations, which weight words in a given CM record based on the relative frequency of word occurrence in the overall dataset, were used to predict the asset labels in the ML algorithm [34], [35].

A SVM algorithm, which trains a separate classifier for each pair of labels, was used in this study since it had one of the highest classification accuracies for the records [36]. The records with existing asset labels were split into a training (80%) and testing (20%) set for the SVM algorithm. The results were evaluated based on an accuracy score generated by comparing the ''predicted'' asset from the trained algorithm with the technician-labeled entry in the testing dataset. The scikit-learn library in Python was used to process the data and apply the SVM algorithm using the C-Support Vector Classification function [37]. Subsequent analysis was conducted for all records tagged with ''inverter'' as the asset.

2) CATEGORIZATION OF INVERTER FAILURE RECORDS

Structural topic modeling (STM) is an unsupervised ML approach that uses Latent Dirichlet Allocation (LDA) to identify ''topics'' or groups of words that occur frequently and exclusively together [38], [39]. Topic modeling has been effectively used to understand patterns in text in diverse fields, including sociology and law [40], [41]. STM is a robust approach for identifying related records since it does not rely on user specification of key terms and is not sensitive to misspelled words. The ''stm'' package in R was used to conduct the structural topic modeling on the inverter-related CM records.

STM was used to group the inverter CM records into those covering similar topics and help identify patterns among them. Each topic encompasses a different combination of words found within the text and can be used to understand the general category being discussed; the sum of all word probabilities for a given topic is one. These topics are then mapped onto the CM records to help identify the proportion of a topic within each record; for a given CM record, the topic proportions sum to one [42]. The resulting model helps contextualize the data, including the most frequent topics and the top words associated with each topic (Figure 2). These topics are manually assigned a label based on the most frequent words within the topics (Table 2). Similar to other unsupervised ML techniques, however, the number of topics used by STMs is user-specified; the ''searchK'' function within the stm package in R was used as a diagnostic tool to identify the number of topics for grouping the maintenance records [43]. In this study, the CM records were clustered into 66 topics.

C. DATA ANALYSIS

Failure rates over time were estimated using a Kaplan-Meier estimator, following the methodologies outlined by Gunda and Homan [44]. A non-parametric approach, the Kaplan-Meier estimator generates a conditional probability that estimates the probability of a site not yet having experienced a failure by a given point in time (for sites that experienced that failure) [45]. The values range between zero to one, indicating the probability that a failure will occur within a site by time t. Inverters in the field can vary in the amount of time they have been placed in service, with some inverters not experiencing a particular failure within the period observed [46]. The Kaplan-Meier estimator is able to account for this censorship within the data, where censorship is defined as records ending before a site experiences failure (either because the site went offline, or observations at the site ended) [47]. Since the sites all began operations at different times, failure analysis is conducted as a function of time since commercial operation started. Given the limited information regarding specific assets within the database, the analysis is limited to first occurrence of a given failure at a site (i.e., subsequent failures of the same type are not considered). While a non-parametric approach allows for more flexibility, Weibull distribution parameters are also useful for informing cost modeling efforts [48]. These distribution parameters were derived using the ''survival'' package in R [49].

A population of inverters in California may not see similar stress to inverters in New York or inverters in Florida due to various operational differences, including grid and/or climatic conditions [46]. To understand these variations in space and time within the inverter-related records, data visualizations and STM evaluations were conducted. Time series charts and pareto charts (which contain a rank order summary of items based on frequency) were used to identify the general patterns in records. Specific patterns within topics were also evaluated (using the ''estimateEffects'' function in the ''stm'' package in R [43]) to understand variations in topic coverage as a function of the different metadata characteristics, such as local climate, inverter type, time of

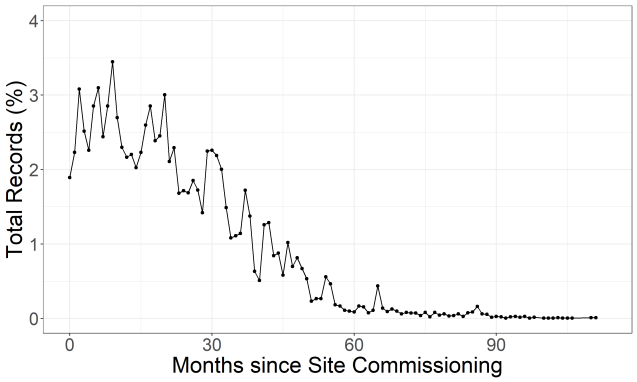


FIGURE 4. Frequency of Inverter Records over Time. Generally, a higher proportion of the records are present within the first 3 years of the site commissioning. Fewer records after five years likely reflects limited data since most sites are relatively young.

commissioning, and month of failure; these variations were estimated using regression where the topic proportions serve as the outcome while the metadata serve as covariates. Correlations between the topics were also used to understand patterns between associated records [50].

III. RESULTS AND DISCUSSION

A. MACHINE LEARNING IMPLEMENTATIONS

The SVM algorithm had an accuracy of 90%, indicating the predicted entries matched existing labels 90% of the time. The gap-filled assets dataset indicated that inverter-related records represent 33% of all records within the CM database (Table 1; Figure A4). These records are present for sites with central, string, and mixed inverter types; no inverter-related records were identified at sites with microinverters (Figure A3). Consistent with [4], a significant percentage of inverter records - underscoring high failures and associated maintenance activities - are concentrated in the first few years of operations (Figure 4). The lack of a significant spike (reflecting infant mortality issues) in the first few months of site operations (Figure 4) likely reflects discovery of certain issues (e.g., faulty or loose connections) during site construction and commissioning, which happen before commercial operations commence and are not captured in CMMSs [7], [46] while the presence of few records after 5 years likely reflects the young age of the sites (Figure A1), consistent with larger industry trends [4].

Unlike the SVM implementation, the LDA implementation does not contain an accuracy score; rather its utility is dependent on the coherency and interpretability of findings. Results from STM indicate multiple interpretable topics emerge from the CM records (Figure 5). The most common subsystems discussed within the records are communications (7% of records), heat management systems (i.e., coolants, fans, and filters; 4% of records), ground faults (3% of records), and IGBTs (2% of records); multiple heat management systems-related topics are present in the dataset, regarding coolants (Topic 45), contactors/fans (Topic 54), and filters (Topic 15). Issues related to PCBs, capacitors, breakers, fuses, and enclosures are also discussed within the records (Figure 5). Given the central role of the inverter within the PV system, issues plaguing neighboring equipment, such as combiners and transformers are also discussed within the inverter CM records. In addition to identifying specific subsystems, the ML analysis identified three common troubleshooting activities: power cycling of the inverters (Topic 7), remote reset (Topic 11), and making repairs (Topic 30) (Figure 5). A complete listing of top words within each topic can be found in the Appendix (Figure A6). These failure modes are generally consistent with findings from industry surveys and maintenance record reviews [3], [4], [6], [7], [9], [33], [51].

Correlation analysis identified varying associations between topics (Figure A7). The correlation between the transformer-related topics (Topics 10 and 42) is higher than the correlation between the communication-related topics (Topics 19, 43, and 55) (Figure A7). The low correlations between communication-related topics indicates notable diversity in underlying causes. For example, if an inverter were to fault resulting in a loss of communication, that fault may or may not have been observed prior to the inverter stopping communication. Ground fault-related maintenance records (Topic 53) are highly correlated to maintenance records discussing remote reset (Topic 11) indicating nuisance tripping of these systems (Figure A7). Topics associated with cabling (Topic 26), fuses (Topic 63), IGBTs (Topic 23), and coolants (Topic 45) also have strong associations with each other (Figure A7).

B. PATTERN ANALYSES

The four dominant subsystems from the STM analysis (communications, heat management systems, ground faults, and IGBTs) were further evaluated to identify patterns in failures. Implementation of the Kaplan-Meier estimator indicates that the likelihood of first failure as a function of time is relatively high in the first year of site operations across all four of these subsystems (Figure 6). Communication systems have the highest likelihood of failure until 1500 days (∼year 4), after which heat management systems dominate (Figure 6). The probability of a failure having occurred in communication systems reaches 50% at 685 days (∼1.9 years), while ground faults and heat management systems reach a 50% probability between 2-3 years (856 days and 908 days respectively), and IGBTs fail at a slower rate, reaching 50% of failure at 1514 days (∼4.1 years) (Figure 6). The maximum failure probabilities observed for these systems after 3200 days of operation (∼8.8 years, the longest period observed in the dataset), is 100% for heat management systems, 94% for ground faults, 92% for communications, and 84% for IGBTs (Figure 6). Associated Weibull parameters fitting these failure rates (provided in Table 3) indicate a decrease in failure rates over time, since the shape factors are all less than 1 [44]. These parameters can be used to inform PV reliability simulations and cost model planning estimates [5], [6], [30], [48].

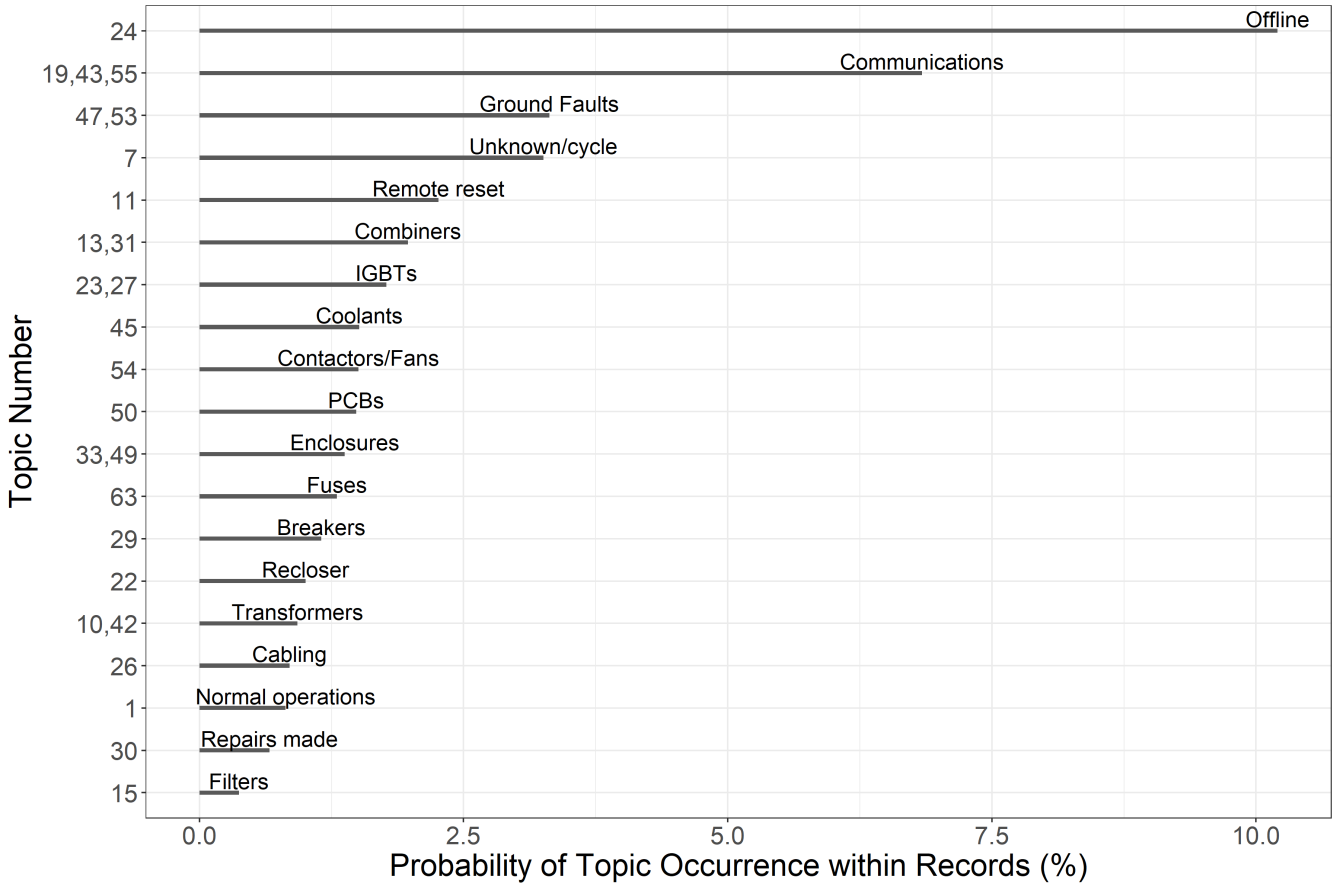


FIGURE 5. Select Topics within Inverter Records. Y-axis indicates that topic number(s) while x-axis indicates the average probability of that topic being present within a given ticket; probability values for subsystems with multiple topics (e.g., communications) were summed. Labels reflect the most frequent words associated with the topics.

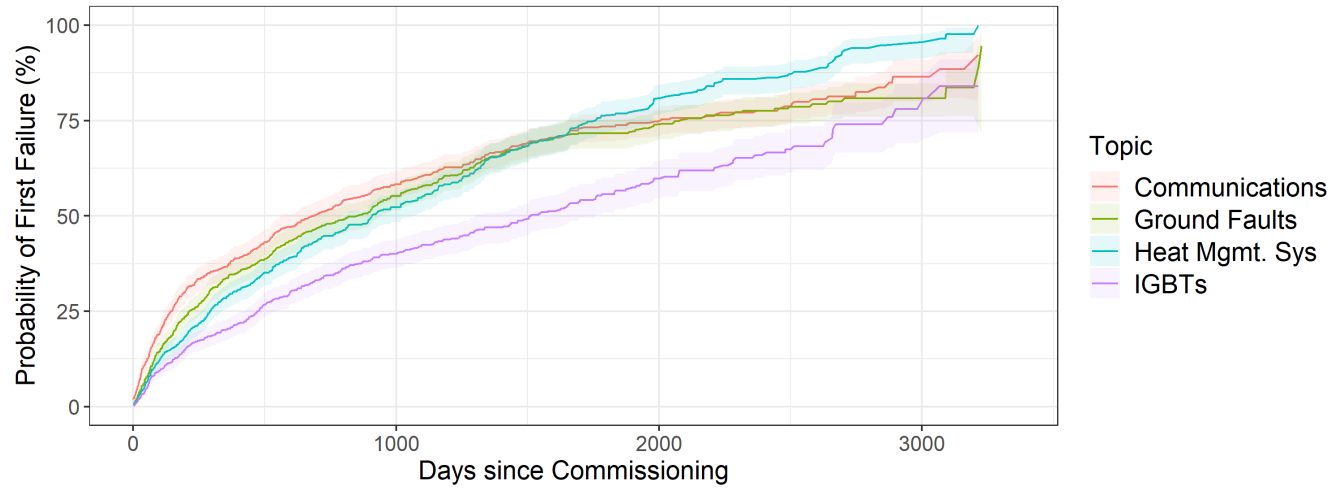


FIGURE 6. Probability of First Failures. Solid lines indicate probability of failure for the different topics with associated 95% confidence intervals in shaded regions. Generally, communication systems have a higher probability of failure until 1500 days (~year 4), after which heat management systems dominate. IGBTs have a lower likelihood of failure than the other systems.

Although inverter technology has changed over time, the number of communication- and IGBT-related records seem to be relatively stable on an annual basis (Figure 7). In contrast, the number of ground fault-related records have significantly increased over the last decade while the number of records related to heat management systems (i.e., coolants, fans, and filters) has decreased over time (Figure 7). These patterns could reflect changes in both technology implementation as well as performance. For example, code changes prompted the installation of ground

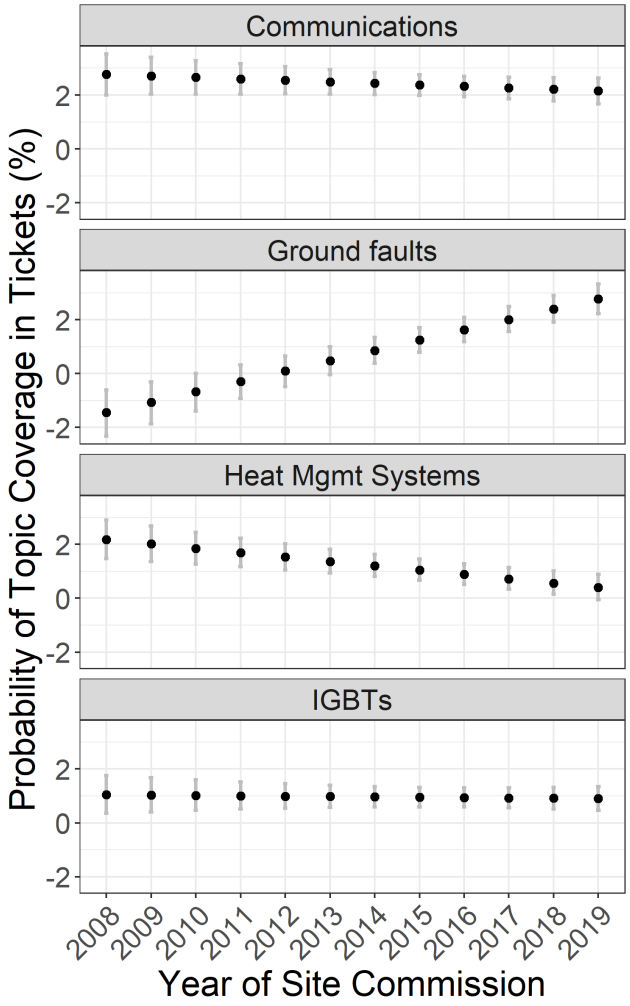


FIGURE 7. Topic Variations for Different Commissioning Dates. Points indicate mean values while gray bars indicate 95% confidence intervals; negative probabilities emerge from the interpolated nature of the spline function in the estimateEffects function. Number of records pertaining to ground faults has significantly increased in recent years while number of records related to heat management systems has decreased. Communications and IGBT-related records have been relatively stable over the last decade.

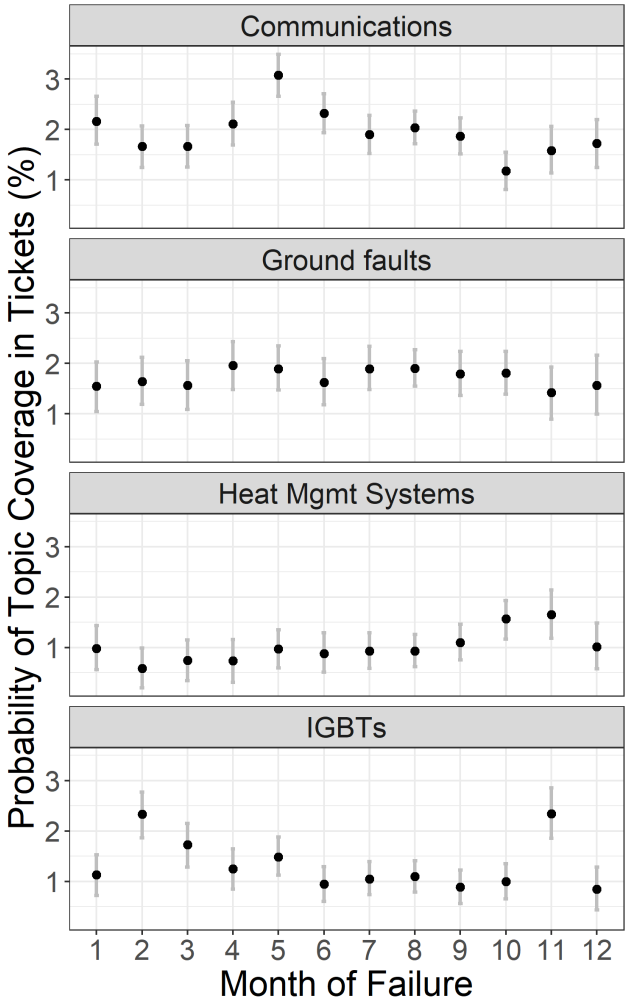


FIGURE 8. Seasonal Patterns in Topic Variations. Communications and ground fault-related records are generally greater in the spring and summer months while heat management-related records peak in the fall. IGBT-related records are highest in Feb and Nov.

TABLE 3. Weibull Parameters for Annual Failure Rates.

Inverter Subsystem	Shape Factor (α)	Scale Factor (β)
Communications	0.69	3.29
Ground Faults	0.77	3.60
Heat Mgmt. Systems	0.93	3.35
IGBTs	0.81	6.01

fault detection devices more consistently within PV systems, which has led to an increase in recorded observations (many of which are false positives) [52].

Temporal patterns are also present within a given year with inverter-related records being relatively frequent in the spring and summer months (Figure A5). However, different seasonal patterns emerge when looking at specific topics (Figure 8). Both communication and ground fault issues have greater occurrences in the spring and summer months (likely due to increased moisture conditions) while heat management issues peak in the fall months (Figure 8), which could reflect refilling of coolants when these liquids contract in the cooler months. IGBTs also exhibit a different seasonal pattern, with probabilities of these tickets highest in the late winter and spring months with another significant increase observed in late fall (Figure 8). The higher prevalence of IGBT issues in Feb and Nov could indicate stress induced by higher array voltage produced in lower temperatures when irradiance may still be high, leading to maximum power flows that could stress internal electronics.

Different patterns emerge as a function of site characteristics as well. For example, sites with string inverters report fewer communications issues while sites with central inverters report ground fault-related issues more often (Figure A8). The highest probability of failure is attributed to the sites with mixed inverter types (i.e., both central and string) but the dataset contains a relatively small percentage of these sites, so more data is needed to validate this pattern (Figure A3). The lower number of records related to communication and ground fault-related records at sites with string-level inverters could reflect the level of detail captured within CMMS records since string inverters may often be replaced in their

entirety (versus specific components within central inverters). These variations have important implications given the increasing installation of string inverters in current fleets (Figure A1).

Arid regions appear to have a slightly higher likelihood for ground faults than sites in other conditions but otherwise, no significant variability due to climate was observed in this analysis (Figure A9). This is likely due to the limited number of sites in non-temperate regions within the database (Figure A2). Since other studies have indicated a correlation of inverter failures with climate zones [4], additional data is needed to validate patterns found in this analysis and associated drivers, such as wire management issues driving the increased prevalence of ground faults in arid regions (Figure A9).

C. FUTURE WORK

Future work should consider extending the ML implementations to extract additional details from the text descriptions, such as cause and response activities associated with failures. This would extend the insights gained from the maintenance records, beyond those indicated by the correlation of topics (e.g., high nuisance alarms for ground faults). Additional text-based ML techniques that focus on relationship extraction and sentence sequence patterns within the text descriptions can help support the development of these capabilities [53], [54]. As more grid-forming inverters and smart inverters come online [55], creating standardized approaches for data collection and analysis can help improve the performance of the implemented ML algorithms, including SVM and LDA, as well as support benchmarking analyses, which can vary based on fleet size, technology, location, scope, labor rates, local energy prices, and available incentives [7], [33], [56].

Pattern analyses could also be extended to consider production and financial information. Industry reports highlight that inverter maintenance could be up to 75% of an overall annual site O&M budget, with inverter replacement encompassing an additional 60% [57]; however, little information is available regarding cost differences among specific inverter failure modes. For example, IGBTs have been identified as an expensive component [30] but details regarding the exact number of labor hours and equipment costs associated with these repairs were not readily available within the database. Expanding the database and improving data collection practices would support an evaluation of differences in failure rates due to inverter size, manufacturer, warranty coverage, and inverter hours as well as quantify associated cost and energy impacts of the observed failures in future work [6], [33]. Fusion of production information with maintenance records, in particular, would help quantify the information implicitly captured in inverter topics, which range from impairment (degraded power output and loss of major functionalities, such as communications) to outages (Figure 5). Research is currently underway to develop a text-to-timeseries toolkit that can leverage these ML techniques to develop consistent labels in O&M datasets and support further evaluation of failure and degradation patterns.

IV. CONCLUSION

This analysis highlights how ML techniques can aid in text-based data preparation and curation for reliability analyses, even when there's a lack of standardization in O&M data collection and management practices. Specifically, the implementation of supervised and unsupervised algorithms enables an efficient approach for identifying records of interest and categorizing them into consistent categories. These capabilities for extracting insights from text-based information can be extended to other energy sectors, such as wind [58] and nuclear [59], with similar challenges and interests in improving maintenance capabilities.

The data-driven evaluation of the maintenance records in this study indicates that inverters continue to dominate reported CM activities at PV sites and that inverter subsystems emerge as a strong commonality for categorizing failure modes across multiple CMMSs. Strong associations between PV inverter topics also indicates the opportunities provided by ML to identify co-occurring subsystems within a single maintenance record, which might not be readily apparent if these tickets were pre-labeled by technicians. Variations in subsystem failures were also extended to demonstrate patterns across time, space, and site features, which provide important insights into potential root causes of these failures. For example, the high prevalence of communication issues during wet conditions (spring/summer peaks and in tropical regions) indicates the importance of managing moisture conditions for these components.

However, consistent with larger industry trends [4], most of the sites within the database have only been online since 2015 (Figure A1). Thus, continued data collection is needed to more robustly evaluate failure rates over time, particularly for informing and developing cost-effective post-warranty maintenance activities and frequency [5]. Improved data collection to capture the specific component involved in a failure would enable evaluation of recurrent patterns. An expanded database would also support validation of patterns observed, including whether few records for micro-inverters and string inverters indicate greater reliability of these technologies or poor data collection (Figure A3) as well as underlying cause for the relatively low frequency of inverter-related records in tropical climate regions (Figure A2). Understanding these geographic and seasonal variations in inverter failures can inform the design of tailored control strategies that can improve reliability in performance [60].

In addition to expanding the data analyzed and standardizing labels, future work can consider additional text-based ML techniques and fusion of text data with production and financial data to develop actionable insights. Such increased understanding of failure modes can inform ongoing reliability activities as well as the development of new monitoring activities that shift the industry from reactive to condition-based

maintenance activities, which could reduce overall system downtime.

ACKNOWLEDGMENT

The authors would like to thank Sig Gonzales, Peter Hacke, Steve Hanawalt, Yang Hu, Nicole Jackson, Birk Jones, Hector Mendoza, Russ Morris, Fernando Rodriguez, David Petrie, and Josh Stein for their insights and feedback throughout this analysis. 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-NA-0003525. The views expressed in the article do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

REFERENCES

[1] S. Feaster and D. Wamsted. (2020). *IEEFA U.S.: Utility-Scale Renewables Top Coal for the First Quarter of 2020*. [Online]. Available: https://ieefa.org/ieefa-u-s-utility-scale-renewables-top-coal-for-the-f%irst-quarter-of-2020/

[2] *SETO in 2020: A Decade of Progress, A Promising Future*, DOE, New York, NY, USA, 2010.

[3] L. Cristaldi, M. Faifer, M. Lazzaroni, M. M. A. F. Khalil, M. Catelani, and L. Ciani, "Diagnostic architecture: A procedure based on the analysis of the failure causes applied to photovoltaic plants," *Measurement*, vol. 67, pp. 99–107, May 2015.

[4] D. C. Jordan, B. Marion, C. Deline, T. Barnes, and M. Bolinger, "PV field reliability status—Analysis of 100 000 solar systems," *Prog. Photovoltaics, Res. Appl.*, vol. 28, no. 8, pp. 739–754, Aug. 2020.

[5] A. Ristow, M. Begovic, A. Pregelj, and A. Rohatgi, "Development of a methodology for improving photovoltaic inverter reliability," *IEEE Trans. Ind. Electron.*, vol. 55, no. 7, pp. 2581–2592, Jul. 2008.

[6] P. Hacke, S. Lokanath, P. Williams, A. Vasan, P. Sochor, G. TamizhMani, H. Shinohara, and S. Kurtz, "A status review of photovoltaic power conversion equipment reliability, safety, and quality assurance protocols," *Renew. Sustain. Energy Rev.*, vol. 82, pp. 1097–1112, Feb. 2018.

[7] A. Golnas, "PV system reliability: An operator's perspective," in *Proc. IEEE 38th Photovoltaic Specialists Conf. (PVSC)*, Jun. 2012, pp. 1–6.

[8] B. Solar. (2019). *Thoughtful Inverter Procurement Can Prevent 25% of Lost Revenue: Inverter Warranty Management*. [Online]. Available: https://www.kwhanalytics.com/blog-archive/solar-risk-assessment-2019-qu%antitative-insights-from-the-industry-experts

[9] *2019 PV Inverter Scorecard*, PVEL, Berkeley, CA, USA, 2011.

[10] A. Sangwongwanich, Y. Yang, D. Sera, F. Blaabjerg, and D. Zhou, "On the impacts of PV array sizing on the inverter reliability and lifetime," *IEEE Trans. Ind. Appl.*, vol. 54, no. 4, pp. 3656–3667, Jul. 2018.

[11] L. Peters and R. Madlener, "Economic evaluation of maintenance strategies for ground-mounted solar photovoltaic plants," *Appl. Energy*, vol. 199, pp. 264–280, Aug. 2017.

[12] L. H. Stember, W. R. Huss, and M. S. Bridgman, "A methodology for photovoltaic system reliability & economic analysis," *IEEE Trans. Rel.*, vols. R–31, no. 3, pp. 296–303, Aug. 1982.

[13] V. V. S. Pradeep Kumar and B. G. Fernandes, "A fault-tolerant single-phase grid-connected inverter topology with enhanced reliability for solar PV applications," *IEEE J. Emerg. Sel. Topics Power Electron.*, vol. 5, no. 3, pp. 1254–1262, Sep. 2017.

[14] B. Dumnic, E. Liivik, D. Milicevic, B. Popadic, V. Katic, and F. Blaabjerg, "Fault analysis and field experiences of central inverter based 2 MW PV plant," in *Proc. 20th Eur. Conf. Power Electron. Appl.*, 2018, pp. 1–5.

[15] S.-V. Oprea, A. Bâra, D. Preoţescu, and L. Elefterescu, "Photovoltaic power plants (pv-pp) reliability indicators for improving operation and maintenance activities. A case study of PV-PP Agigea located in Romania," *IEEE Access*, vol. 7, pp. 39142–39157, 2019.

[16] F. Spertino, E. Chiodo, A. Ciocia, G. Malgaroli, and A. Ratclif, "Maintenance activity, reliability analysis and related energy losses in five operating photovoltaic plants," in *Proc. IEEE Int. Conf. Environ. Electr. Eng. IEEE Ind. Commercial Power Syst. Eur.*, Jun. 2019, pp. 1–6.

[17] A. Livera, M. Theristis, E. Koumpli, S. Theocharides, G. Makrides, J. Sutterlueti, J. S. Stein, and G. E. Georghiou, "Data processing and quality verification for improved photovoltaic performance and reliability analytics," *Prog. Photovoltaics, Res. Appl.*, Oct. 2020.

[18] K. A. Klise, "Performance monitoring using pecos," Sandia National Lab.(SNL-NM), Albuquerque, NM, USA, Tech. Rep. SAND2016-4303C, 2016.

[19] T. Gunda and C. B. Jones, "Data-driven analysis of PV failures from O&M records," in *Proc. Renewables O&M Innov. Workshop*, Charlotte, NC, USA: Sandia National Lab. (SNL-NM), 2019.

[20] M. K. Alam, F. H. Khan, J. Johnson, and J. Flicker, "PV faults: Overview, modeling, prevention and detection techniques," in *Proc. IEEE 14th Workshop Control Model. for Power Electron.*, Jun. 2013, pp. 1–7.

[21] E. Garoudja, F. Harrou, Y. Sun, K. Kara, A. Chouder, and S. Silvestre, "Statistical fault detection in photovoltaic systems," *Sol. Energy*, vol. 150, pp. 485–499, Jul. 2017.

[22] I. M. Karmacharya and R. Gokaraju, "Fault location in ungrounded photovoltaic system using wavelets and ANN," *IEEE Trans. Power Del.*, vol. 33, no. 2, pp. 549–559, Apr. 2018.

[23] X. Gong, N. Wang, Y. Zhang, S. Yin, M. Wang, and G. Wu, "Fault diagnosis of micro grid inverter based on wavelet transform and probabilistic neural network," in *Proc. 39th Chin. Control Conf. (CCC)*, Jul. 2020, pp. 4078–4082.

[24] D.-M. Tsai, S.-C. Wu, and W.-C. Li, "Defect detection of solar cells in electroluminescence images using Fourier image reconstruction," *Sol. Energy Mater. Sol. Cells*, vol. 99, pp. 250–262, Apr. 2012.

[25] M. W. Hopwood, T. Gunda, H. Seigneur, and J. Walters, "Neural network-based classification of string-level IV curves from physically-induced failures of photovoltaic modules," *IEEE Access*, vol. 8, pp. 161480–161487, 2020.

[26] J. Wu, Z. Yan, and Q. Sun, "Multiple faults detection of three-level NPC inverter based on improved deep learning network," in *Proc. Int. Conf. Appl. Techn. Cyber Secur. Intell.* Springer, 2019, pp. 1575–1583. [Online]. Available: http://link-springer-com-443.webvpn.fjmu.edu.cn/chapter/10.1007%2F978-3-030-25128-4_195#citeas

[27] A. Livera, M. Theristis, G. Makrides, and G. E. Georghiou, "Recent advances in failure diagnosis techniques based on performance data analysis for grid-connected photovoltaic systems," *Renew. Energy*, vol. 133, pp. 126–143, Apr. 2019.

[28] E. F. Alsina, M. Chica, K. Trawiáski, and A. Regattieri, "On the use of machine learning methods to predict component reliability from data-driven industrial case studies," *Int. J. Adv. Manuf. Technol.*, vol. 94, nos. 5–8, pp. 2419–2433, Feb. 2018.

[29] A. P. Talayero, A. Llombart, and J. J. Melero, "Diagnosis of failures in solar plants based on performance monitoring," *Renew. Energy Power Qual. J.*, vol. 18, pp. 33–128, Dec. 2020.

[30] J. M. Freeman, G. T. Klise, A. Walker, and O. Lavrova, "Evaluating energy impacts and costs from PV component failures," in *Proc. IEEE 7th World Conf. Photovolt. Energy Convers.*, Jun. 2018, pp. 1761–1765.

[31] N. Enbar, D. Weng, and G. T. Klise, "Budgeting for solar PV plant operations & maintenance: Practices and pricing," Sandia Nat. Lab. (SNL-NM), Albuquerque, NM, USA, Tech. Rep. SAND-2016-0649R, 2016.

[32] M. C. Peel, B. L. Finlayson, and T. A. McMahon, "Updated world map of the Köppen–Geiger climate classification," *Hydrol. Earth Syst. Sci. Discuss.*, vol. 4, no. 2, pp. 439–473, Dec. 2007.

[33] S. Lokanath, "Central inverter cost of ownership and event analysis," in *Proc. NREL/SNL/BNL PV Rel. Workshops*, Lakewood, CO, USA, Feb. /Mar. 2017, pp. 542–560.

[34] N. Hardeniya, J. Perkins, D. Chopra, N. Joshi, and I. Mathur, *Natural Language Processing: Python and NLTK*. Birmingham, U.K.: Packt, 2016.

[35] J. Lilleberg, Y. Zhu, and Y. Zhang, "Support vector machines and Word2vec for text classification with semantic features," in *Proc. IEEE 14th Int. Conf. Cognit. Informat. Cognit. Comput.*, Jul. 2015, pp. 136–140.

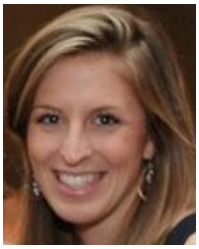
[36] *Application of Machine Learning to Large-Scale PV Plant Faults and Failures*, Electr. Power Res. Inst., Palo Alto, CA, USA, 2020,

[37] F. Pedregosa, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[38] M. E. Roberts, B. M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. K. Gadarian, B. Albertson, and D. G. Rand, "Structural topic models for open-ended survey responses," *Amer. J. Political Sci.*, vol. 58, no. 4, pp. 1064–1082, Oct. 2014.
[39] M. E. Roberts, B. M. Stewart, and E. M. Airoldi, "A model of text for experimentation in the social sciences," *J. Amer. Stat. Assoc.*, vol. 111, no. 515, pp. 988–1003, 2016.
[40] P. DiMaggio, M. Nag, and D. Blei, "Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding," *Poetics*, vol. 41, no. 6, pp. 570–606, Dec. 2013.
[41] J. Ruhl, J. Nay, and J. Gilligan, "Topic modeling the president: Conventional and computational methods," *Geo. Wash. L. Rev.*, vol. 86, p. 1243, Dec. 2018.
[42] J. Silge and D. Robinson, *Text Mining With R: A Tidy Approach*. Newton, MA, USA: O'Reilly Media, 2017.
[43] M. E. Roberts, B. M. Stewart, and D. Tingley, "STM: An R package for structural topic models," *J. Stat. Softw.*, vol. 91, no. 2, pp. 1–40, 2019.
[44] T. Gunda and R. Homan, "Evaluation of component reliability in photovoltaic systems using field failure statistics," Sandia Nat. Lab.(SNL-NM), Albuquerque, NM, USA, Tech. Rep. SAND2020-9231, 2020.
[45] J. M. Bland and D. G. Altman, "Survival probabilities (the Kaplan-Meier method)," *Bmj*, vol. 317, no. 7172, pp. 1572–1580, 1998.
[46] J. M. Fife, M. Scharf, S. G. Hummel, and R. W. Morris, "Field reliability analysis methods for photovoltaic inverters," in *Proc. 35th IEEE Photovolt. Spec. Conf.*, Jun. 2010, p. 2.
[47] J. T. Rich, J. G. Neely, R. C. Paniello, C. C. J. Voelker, B. Nussenbaum, and E. W. Wang, "A practical guide to understanding Kaplan-Meier curves," *Otolaryngology–Head Neck Surgery*, vol. 143, no. 3, pp. 331–336, Sep. 2010.
[48] A. Walker, "PV O&M cost model and cost reduction," in *Proc. Photovolt. Module Rel. Workshop*, 2017, pp. 1–5. [Online]. Available: https://www.nrel.gov/docs/fy17osti/68023.pdf
[49] T. Therneau. (2020). *Survival: R Package for Survival Analysis (Version 3.1-12)*. [Online]. Available: https://cran.r-project.org/package-survival
[50] D. Walter and Y. Ophir, "News frame analysis: An inductive mixed-method computational approach," *Commun. Methods Measures*, vol. 13, no. 4, pp. 248–266, 2019.
[51] J. Flicker, "PV inverter performance and component-level reliability," in *Proc. Photovolt. Module Rel. Workshop*, 2014, pp. 1–5.
[52] G. Ball, B. Brooks, J. Johnson, J. Flicker, A. Rosenthal, J. Wiles, L. Sherwood, M. Albers, and T. Zgonena, "Inverter groundfault detection—Blind spot and mitigation methods," Solar America Board Codes Standards, Orlando, FL, USA, Tech. Rep., 2013. [Online]. Available: http://solarabcs.org/about/publications/reports/blindspot/pdfs/inverter_groundfault-2013.pdf
[53] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 5, pp. 1234–1240, Sep. 2019.
[54] A. Sharma, R. Swaminathan, and H. Yang, "A verb-centric approach for relationship extraction in biomedical text," in *Proc. IEEE 4th Int. Conf. Semantic Comput.*, Sep. 2010, pp. 377–385.
[55] M. Mesbahi. (2018). *Cybersecurity: Peril of Smart Solar and Cybersecurity Measures*. [Online]. Available: https://northamerica.solar-asset.management/whitepaper-cybersecurity
[56] *Data collection practices to facilitate analytics of photovoltaic plant maintenance logs*, Electr. Power Res. Inst., Palo Alto, CA, USA, 2019.
[57] (2019). *Resource Guide: Utility Solar Asset Management and Operations and Maintenance*. [Online]. Available: https://www.solarpowerworldonline.com/2016/02/sepa-om-guide/
[58] A. Koltsidopoulos Papatzimos, T. Dawood, and P. R. Thies, "An integrated data management approach for offshore wind turbine failure root cause analysis," in *Proc. Int. Conf. Offshore Mech. Arctic Eng.*, vol. 57663, 2017, Art. no. V03BT02A012.
[59] H. A. Gohel, H. Upadhyay, L. Lagos, K. Cooper, and A. Sanzetenea, "Predictive maintenance architecture development for nuclear infrastructure using machine learning," *Nucl. Eng. Technol.*, vol. 52, no. 7, pp. 1436–1442, Jul. 2020.
[60] A. Sangwongwanich, Y. Yang, D. Sera, and F. Blaabjerg, "Mission profile-oriented control for reliability and lifetime of photovoltaic inverters," *IEEE Trans. Ind. Appl.*, vol. 56, no. 1, pp. 601–610, Jan. 2020.

THUSHARA GUNDA received the B.S. degree in environmental sciences and the B.A. degree in environmental policy from the University of Virginia and the Ph.D. degree in environmental engineering from Vanderbilt University. She is currently a Senior Member of Technical Staff at Sandia National Laboratories. Her research interests focus on leveraging data science to identify patterns in both time series (e.g., PV performance) and discrete event (e.g., PV O&M) datasets.

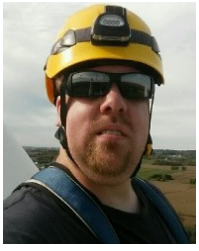
SEAN HACKETT received the Bachelor of Science degree in sustainable technology from Appalachian State University. He is currently an Engineer/Scientist III with the Electric Power Research Institute (EPRI). He is also a key member of the EPRI's solar generation program, which addresses issues across the solar PV asset class, including technology assessment, life cycle and best practices, and fundamental information. He joined EPRI, in 2018, and has worked in the solar industry for more than ten years. His primary experience is in the engineering, procurement and construction of large scale ground mounted and commercial rooftop PV facilities, and managing the project design, procurement, installation, and commissioning activities. He is also a North American Board of Certified Energy Practitioners (NABCEP) PV Installation Professional.



LAURA KRAUS received the B.S. degree in business administration/economics from Saint Louis University, the M.A.T. degree in mathematics from Belmont University, and the M.S. degree in economics from the University of Oregon. She is currently the Manager of performance analytics with Strata Solar, LLC. She leads a team of data analysts and PV system performance engineers in the optimization of PV systems for Strata's 2.5-GW+ operations and maintenance portfolio. Her work focuses on researching and driving best practices in PV system modeling and using data to drive efficiency gains in O&M operations. Her team works on diverse operations activities, including energy forecasting, PV system performance testing, predictive analytics, building machine learning algorithms, identifying performance trends over time, making performance recommendations, preparing cost-benefit analyses, estimating lost production, cost-modeling, and infrared image analysis.

CHRISTOPHER DOWNS received the B.S. degree in chemical engineering with a concentration in sustainable engineering, energy, and environment from North Carolina State University. He is currently a Performance Engineer II with Cypress Creek Renewables. He and his team are responsible for managing production data and automation processes for the performance optimization of PV systems within Cypress Creek Renewable's 3.4-GW fleet. His team's work centers around utilizing production, revenue, and ticketing data to drive plant performance in a positive direction. This is done through a wide array of activities, including, but not limited to, performance recommendations and collaboration with operations engineers, estimating production losses during events, analyzing trends and optimizing models with MET station data, production forecasting, cost-benefit analysis, performance trend analysis over time, and data management of an ever-growing fleet.

RYAN JONES is currently the Director of performance engineering with CD Arevon USA, Inc. He is responsible for leading the development and review of system performance models, energy tests, technical due diligence, and PV system design optimization. In addition to driving root cause analysis, system performance optimization, and reliability through data analytics, he leads activities regarding MET station and SCADA design standardization.



CHRISTOPHER MCNALLEY received the Bachelor of Science degree in materials science and engineering from Michigan Technological University. He is currently a Senior Engineer with Consumers Energy, with engineering responsibilities for wind, solar, and battery assets. His current work focuses on standardization of performance analytics, reliability improvements, and operations for the consumers energy renewables fleet.

MICHAEL BOLEN received the B.S.E.E. degree from Michigan Technological University, in 2005, and the Ph.D. and M.B.A. degrees from Purdue University, in 2010. He was a Senior Advisor to the U.S. Department of Energy's Solar Energy Technologies Office and a Postdoctoral Researcher with the National Renewable Energy Laboratory. He is currently the Principal Project Manager of the Electric Power Research Institute (EPRI). He currently manages the EPRI's Solar Generation Program, which focuses on solar technology, including hardware and software; and lifecycle best practices for large-scale plants, including design, commissioning, operations, maintenance, and end-of-life.

ANDY WALKER received the B.S., M.S., and Ph.D. degrees in mechanical engineering. He is currently a Principal Engineer with the National Renewable Energy Laboratory, where he conducts engineering and economic analysis of energy efficiency and renewable energy projects for Federal agencies and also for commercial and industrial clients. He is currently managing a DOE SunShot Program on PV O&M, producing a Best Practices Guide and PV O&M Cost Model. He holds a patent on the Renewable Energy Optimization (REO) method of planning renewable energy projects across a portfolio of properties based on economic value, which was awarded the Thomas A. Edison Patent Award for innovation and impact. He has taught energy classes at the University of Colorado at Boulder, the Colorado School of Mines, and at the Metropolitan State University of Denver. He is a Fellow of the American Society of Mechanical Engineers and led the Solar Energy Division and is the author of over 28 book chapters, journal articles, and conference papers, including a reference book *Solar Energy: Technologies and Project Delivery for Buildings* (John Wiley). He is a registered Professional Engineer in the State of Colorado.