Fault Detection and Classification in Photovoltaic Arrays using Machine Learning

by

Sunil Srinivasa Manjanbail Rao

A Dissertation Presented in Partial Fulfillment
of the Requirement for the Degree
Doctor of Philosophy

Approved November 2021 by the
Graduate Supervisory Committee:

Andreas Spanias, Co-Chair
Cihan Tepedelenlioglu, Co-Chair
Konstantinos Tsakalis
Devarajan Srinivasan

ARIZONA STATE UNIVERSITY

December 2021

ABSTRACT

Operational efficiency of solar energy farms requires detailed analytics and information on each panel regarding voltage, current, temperature, and irradiance. Monitoring utility-scale solar arrays was shown to minimize the cost of maintenance and help optimize the performance of photovoltaic (PV) arrays under various conditions. This dissertation describes a project that focuses on the development of machine learning and neural network algorithms. It also describes an 18kW solar array testbed for the purpose of PV monitoring and control. The use of the 18kW Sensor Signal and Information Processing (SenSIP) PV testbed which consists of 104 modules fitted with smart monitoring devices (SMDs) is described in detail. Each of the SMDs has embedded, a wireless transceiver, and relays that enable continuous monitoring, fault detection, and real-time connection topology changes. Data is obtained in real time using the SenSIP PV testbed. Machine learning and neural network algorithms for PV fault classification is are studied in depth. More specifically, the development of a series of customized neural networks for detection and classification of solar array faults that include soiling, shading, degradation, short circuits and standard test conditions is considered. The evaluation of fault detection and classification methods using metrics such as accuracy, confusion matrices, and the Risk Priority Number (RPN) is performed. The examination and assessment the classification performance of customized neural networks with dropout regularizers is presented in detail. The development and evaluation of neural network pruning strategies and illustration of the trade-off between fault classification model accuracy and algorithm complexity is studied. This study includes data from the National Renewable Energy Laboratory (NREL) database and also real-time data collected from the SenSIP testbed at MTW under various loading and shading conditions. The overall approach for detection and classification promises to elevate the performance and robustness of PV arrays.

DEDICATION

To my family.

You have the right to perform your duties,

but you're not entitled to the fruits of the duties.

Do not let the fruit be the purpose of your duties,

and therefore you won't be attached to not doing your duties.

Srimad Bhagavad Gita Chapter 2 Verse 47.

ACKNOWLEDGEMENTS

Over the past several years, I have grown both professionally and personally, and I need to thank a number of people for my growth. I owe an outstanding debt of gratitude to my advisor Dr. Andreas Spanias, Dr. Cihan Tepedelenlioglu and Dr. Devarajan Srinivasan, for their constant support and guidance. This Ph.D. would not have been possible without their encouragement and feedback. I am eternally thankful to them for inscribing good research skills in me. I am grateful to Dr. Konstantinos Tsakalis for his valuable time serving on my defense committee and his insightful comments and helpful feedback.

I would like to recognize the invaluable assistance provided to me by the graduate advisors and the School of Electrical, Computer and Energy Engineering at ASU. I would like to thank my friends, peers, and colleagues at the SenSIP center for their kind support, frequent discussions, and memorable days in the lab. I would also like to thank SenSIP industry partners for providing feedback and helping me develop my presentation skills in various I/UCRC meetings. I am grateful for the continuous graduate teaching assistantship support by the ASU ECEE school.

The project was supported by several research grants and agreements including the NSF CPS award 1646542, the NSF MRI award 2019068, the Poundra LLC support, and the SenSIP center and NSF I/UCRC support (award 1540040).

Most importantly, none of this could have happened without my family. Their unconditional love and support have helped me achieve a great deal of success. In particular, I like to thank my parents Dr. Srinivasa and Shakunthala Rao, my guru Dr. Ajey SNR, my fiancé Sushmita Mudholkar and my sisters Dr. Soumya Rao and Supriya Rai for motivating me and sacrificing countless hours of their own time to help me accomplish my goals. I am extremely grateful to them for supporting me during my doctoral studies.

TABLE OF CONTENTS

Page

CHAPTER Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

Chapter 1

INTRODUCTION

The increasing demand for green energy requires expansion and efficiency improvements in renewable sources. Solar arrays on residential roof tops, parking sites, and large commercial structures are being deployed in several countries. In addition, large utility-scale arrays with generation capacity of several megawatts are now connected to the grid. A large number of modules in remote areas makes faults more likely and more challenging to detect and localize. The occurrence of photovoltaic (PV) faults is often unpredictable and requires constant remote monitoring. Even when overcurrent protection devices (OCPD) and ground fault detection interrupters (GFDI) with data transmission capabilities are integrated within the PV array system, recent studies (Alam *et al.* (2015); Zhao *et al.* (2012a); Flicker and Johnson (2013)) have shown that these devices offer diagnosis for a limited set of commonly occurring faults. On-site inspections are also expensive and time consuming. For this reason, there is a need for an automated remote fault detection along with diagnostics and mobile analytics. This requires communications and sensor hardware operating along with online algorithms and software at the panel level.

Our vision for research monitoring and optimizing a large-scale PV array is summarized in Figure 1.1. The various faults occurring with solar arrays can cause issues of power loss or localized panel damage, while others can create safety hazards. Soiling over time and shading (clouds) over an array can cause a significant decrease in power production (Hammond *et al.* (1997)). This can cause an effect known as "hotspotting" (Braun *et al.* (2012a)). When a limited area of an array is under-producing, this section will absorb some of the PV energy from the fully functioning areas and dissi-

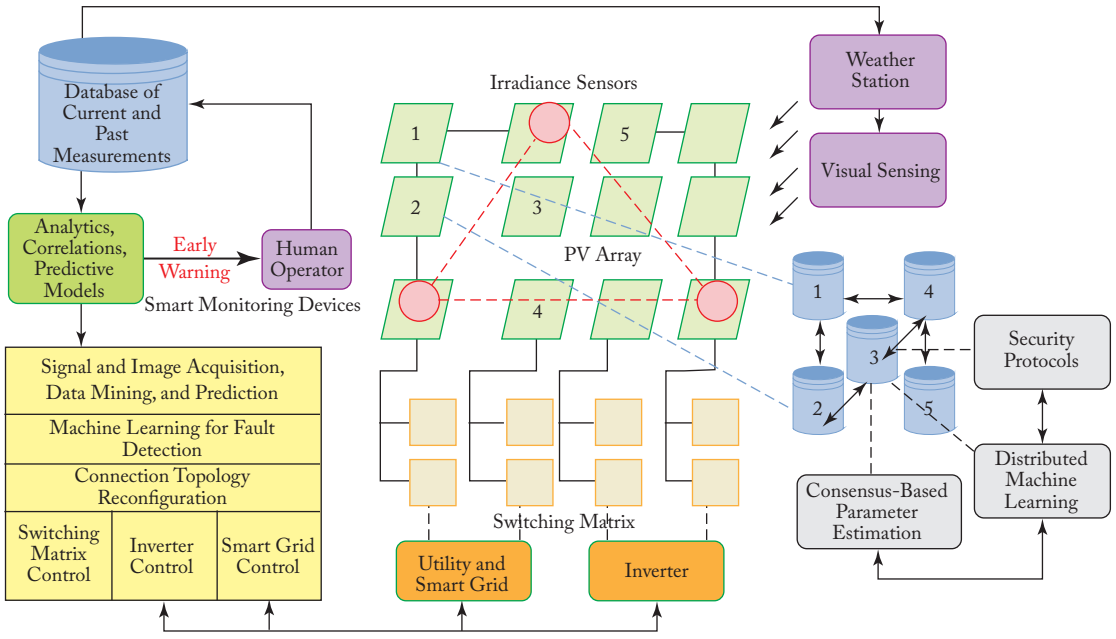

Figure 1.1: Overview of Our Research Vision in Solar Panel Monitoring (Muniraju *et al.* (2017)). Our System Integrates Fault Classification and Diagnosis Modules.

pate it as heat. Due to the parallel and series nature of array segmentation, a small amount of localized degradation in a single panel can have a ripple effect limiting the voltage of all other parallel strings. Besides this, serious safety faults including arc and ground faults are of concern given the high voltages associated with large-scale PV facilities (Alam *et al.* (2015); Wiles (2008)). Given that a suspected problem is recognized, it must then be diagnosed by a technician. This is further complicated by the distinction between a faulty vs. an under-producing system due to environmental conditions or panel age. Trained professional service can be expensive in terms of labor, equipment, compounded with system down-time, and safety. This is not optimal for large-scale arrays where the volume of panel-by-panel metering by technicians increases even further and ultimately is subject to human error.

To support experimental aspects of this research we designed a testing facility (Rao *et al.* (2017)) at the Arizona State University (ASU) research park in Tempe,

Figure 1.2: The SenSIP Solar Monitoring Facility at the ASU Research Park.

Arizona which is shown in Figure 1.2. This research facility consists of 104 modules in a default 8×13 configuration that amounts to approximately 18 kW. Every panel in this solar array is equipped with a smart monitoring device (SMD).

These devices are networked and can provide data to servers, control centers, and ultimately to mobile devices. Each SMD not only provides analytics for each panel but contains relays that can be remotely controlled via wireless access. Relays can bypass or change connectivity configuration, e.g., series to parallel. SMDs, connected to each PV panel, act as intelligent networked sensors providing data that can be used to detect faults, shading, and other problems that cause inefficiencies. Each panel can be monitored individually for voltage, current, and temperature, and all data is transmitted via a wireless channel to a central hub. Additionally, each SMD can reconfigure connections with its nearest neighbors. In fact, the SMDs can accommodate various connection topologies. Data collected from the SMDs and reconfiguration testing will be used to design and evaluate automated fault detection, diagnosis, and mitigation algorithms.

This dissertation describes machine learning (ML) and signal processing tech-

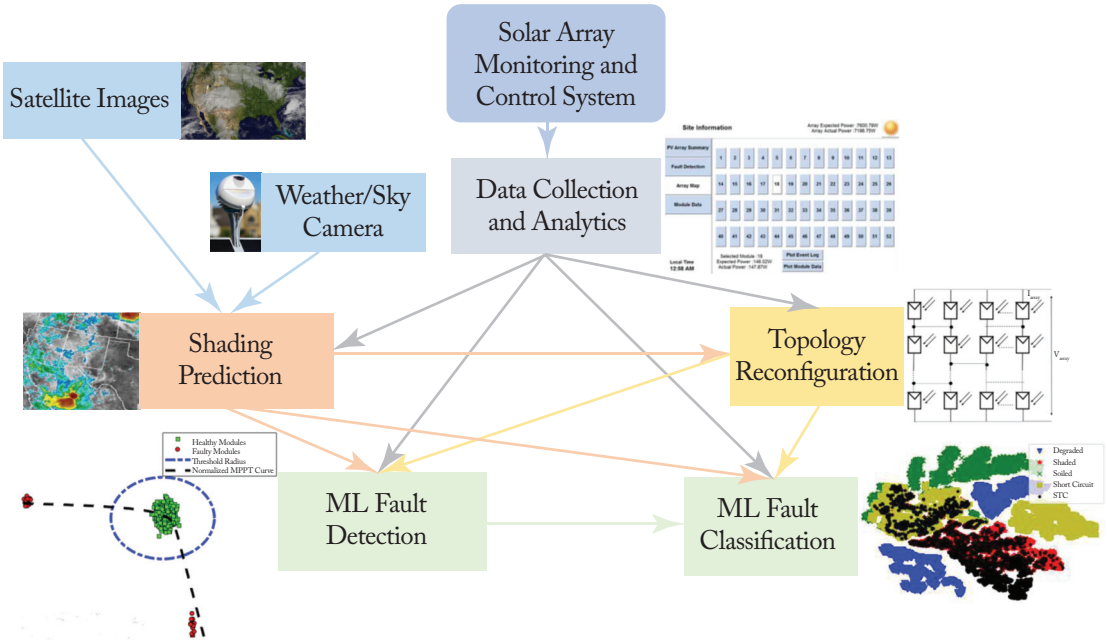

Figure 1.3: Systems and Algorithms Needed for a Holistic Solar Array Monitoring And Control System. The Direction of the Arrow Indicates the Information Flow. For Instance, Topology Reconfiguration Requires the PV Current-voltage (I-V) Measurement Data and Shading Predictions in Order to Switch the Connection Topology. The Information Regarding the New Topology Is Then Passed on to the Fault Detection/Classification Stage. While This System Describes a Holistic Approach for Solar Arrays, the Focus of This Thesis Is Centred Around ML for Fault Detection and Classification in the Solar Array. Rao *et al.* (2020a)

niques that have been shown to improve power generation and robustness in large utility-scale facilities. These methods make possible automated system monitoring, fault detection, and predictive modeling. PV power generation is largely dependent on the irradiance over the modules and cloud cover serves as the major hindrance to the constant power output. A faulty panel could simply be bypassed from the system automatically, improving PV electrical production and eliminating system downtime. Figure 1.3 shows the design of a system by combining these individual components.

We first build an ML algorithm operating on I-V measurements to detect PV panel faults and then an ML classifier is integrated to classify the type of faults detected. The potential to detect and localize PV faults remotely provides opportunities for bypassing faulty modules and retaining power, without disrupting the inverters.

Thus, with the help of ML algorithms (Shanthamallu *et al.* (2017); Rao *et al.* (2019)) for fault detection and classification, our system can simultaneously reconfigure the topology and bypass faulty modules in order to achieve the maximum power generation output, even under non-ideal conditions.

In our previous work (Braun *et al.* (2012a)), we discussed smart PV array monitoring techniques by developing signal processing methods for fault detection, array reconfiguration, and monitoring. In contrast, this dissertation presents advanced intelligent techniques that combine PV module data and shading predictions, for optimal topology reconfiguration and ML-based fault classification/diagnosis. The ML methods presented later in this dissertation are shown to significantly improve performance of the traditional techniques previously described by some of our co-authors in 2012 (Braun *et al.* (2012a)). Moreover, the introduction of ML and deep learning techniques shows potential for additional gains. In addition, our literature review that covers several cyber-physical systems (CPS) (Rao *et al.* (2019); Braun *et al.* (2012b,a, 2016); Katoch *et al.* (2018a,b); Rao *et al.* (2016)), provides a comprehensive bibliography of recent advances in fault detection and diagnosis in PV arrays.

1.1 Motivation

Reliability is a critical factor for a PV system. Issues such as ground faults, arc faults, open circuits, short circuits, soiling, and partial shading can all reduce efficiency and need to be addressed. Some of these faults are undetected for a prolonged

length of time in real-world situations. This leads to reduced and inefficient functioning of the PV array and a significantly lower power output. Unnoticed faults in PV can be dangerous and potentially life threatening. A real-world example would be the Bakersfield fire which was caused due to an undetected ground fault (Mellit *et al.* (2018)). Although ground faults can now be detected with the use of inverters, faults such as soiling and short circuits between modules often go undetected (Braun *et al.* (2012a)). Human operators are currently required to manually perform fault detection and identification. Studies have (Braun *et al.* (2012b); Maish *et al.* (1997); King (1997)) showed that the current method for mean time to repair (MTTR) is at approximately 19 days. There is a significant need to reduce MTTR to reduce power losses from the PV array. We first develop a system to collect data in real-time. We then use machine learning methods to address the MTTR for PV arrays. Fault identification and localization problems pose several challenges and research opportunities. A system must first accurately classify the PV array condition and then react to unseen data to correctly classify the condition of operation of the PV array. Considering these challenges, we explore the use of machine learning techniques (Rao *et al.* (2016)). Semi-supervised learning can be used to label many realistic faults from few measured examples.

1.2 Problem Statement

The I-V data in a PV array can be measured at the panel-level inexpensively. I-V measurements have high correlation. This data can be used to build correlation models. Such models are useful in predicting possible ground faults, arc faults, soiling, shading, etc. (Rao *et al.* (2019)). The I-V curve is modeled using the single diode model as a function of temperature, irradiance, open circuit voltage (V_{OC}) and short circuit current (I_{SC}). Each panel has a peak operating point known as the maximum

power point (MPP). Fault detection using I-V data can be accomplished by measuring MPPs and observing the variation of the measured MPP from the actual MPP.

We review the standard test conditions and some of the faults in PV arrays namely, shading, degraded modules, soiling, and short circuits. We consider the approach of fault detection and classification by monitoring signals such as maximum power point tracking (MPPT) parameters. Standard Test Condition (STC) values correspond to the measurements yielding maximum power under the irradiance values of a particular instance.

1.2.1 Shading

A module is shaded if the irradiance measured is considerably lower than STC, usually caused by overcast conditions, cloud cover, a tree or building obstruction. As a result, the power produced by the PV array is significantly reduced. The irradiance levels measured also are significantly lower compared to STC values.

1.2.2 Degradation

Degraded modules are a result of modules aging or regular wear and tear of the PV modules. Consequently, the degraded modules affect the entire string of the array as it includes both good and degraded modules owing to the lower values of either open-circuit voltage V_{OC} and short circuit current I_{SC}.

1.2.3 Soiling

Since PV modules are exposed to the environment, modules get soiled due to dust, snow, bird droppings and other particulate matter accumulating on the PV module. While the irradiance measured remains the same as STC, the power produced drops significantly.

1.2.4 Short Circuit

The final fault type considered in this dissertation is the short circuit. This not only causes significant power loss but can also create potential fire hazards and cause severe damage to the modules.

To improve the efficiency of PV arrays and prevent safety hazards, we need to identify and localize these faults automatically.

1.3 Statement of Contributions

We consider the problem of detection and classification of faults occurring in utility-scale PV array systems. To address this, we first develop a solar array as described below in Section 1.3.1. We then address fault detection and classification using Machine Learning techniques as described in Section 1.3.2.

1.3.1 Development of Solar Array

As part of detection and identification of faults in PV arrays, we needed to design and develop a load which will collect data in real time. These load banks are used to verify fault detection algorithms developed using simulation models. The load has multiple resistors and can switch between different resistor values according to varying maximum power points (MPP) through the day. We have two load banks which can switch between two different array configurations (12 series 1 parallel and 4 series 3 parallel array configurations). Each load bank has a total of seven resistors which can obtain data in real time. Resistors are programmed such that they adapt to changing MPP values through the day. The load is also programmable such that it can switch between array configurations. We also collect irradiance data in real time at a sampling interval of 1s. We use the obtained real-time data to train our

Machine Learning algorithms. We describe the data collection process in more detail in Appendix A.

1.3.2 Machine Learning

We develop and train customized Machine Learning models for fault detection and classification in solar arrays. First, we develop and train an autoencoder for fault detection. More specifically, we use our custom features to train a 3-layer autoencoder to detect faults. We use the reconstruction error from the autoencoder to create an error histogram, which is used to identify faults. Next, we train a NN for PV fault classification using dropout and concrete dropout regularizers. We compare NNs against the standard machine learning (ML) classification algorithms described in reference (Goodfellow *et al.* (2016)), such as SVM, K-nearest neighbor (KNN), and random forest classifier (RFC). Additionally, we associate the performance of the classification algorithms to the hardness of data separation in PV arrays. We perform dimensionality reduction using the state-of-the-art Distributed Stochastic Neighbor Embedding (t-SNE) algorithm (Maaten and Hinton (2008)) and visualize clusters of faults which are inseparable. Our results show that the 2× pruned networks perform better than standard ML classifiers and concrete dropout has the best performance among all methods examined.

1.3.3 Pruned NN and Lottery Ticket Hypothesis

Pruned NN on embedded hardware greatly improve computational performance and reduce memory requirements with a slight reduction in the model's accuracy(Frankle and Carbin (2019). We integrate the lottery hypothesis optimisation methods to develop a NN architecture such that a dense neural network contains a subnetwork that is initialized such that when trained in isolation it can match the test accuracy of

the original network after training for at most the same number of iterations. Using Monte Carlo simulations, we demonstrate that the test accuracy of a network pruned by 62% (a significant reduction of weights) reduces only by 4% as compared to a fully connected neural network.

1.3.4 Operation and Management of PV Arrays

To provide a practical perspective, we studied the nature of these faults from an operations and management perspective. Faults in PV arrays can be classified into a list of three safety categories. In addition to safety, we also assigned a Risk Priority Number (RPN) to each type of solar fault.

This RPN is calculated as: RPN $= S \times O \times D$, where S denotes the severity (or a numerical subjective estimate of the effect of a failure), O denotes a numerical subjective estimate of likelihood of failure and D the numerical subjective estimate of detection. Failure modes with high RPN are more critical compared to the ones with lower RPN. Each S, O and D estimate is assigned a value between 1 and 10 Dhillon (1999).

1.3.5 Real time assessments and comparisons

We collect and obtain data in real-time. We obtain about 2000 data points for the four classes mentioned above in Section 1.2. Data was collected at a sampling interval of $10s$ for each of the four classes. We develop and train our custom neural network algorithms for fault detection and classification using these data points. Our results compare favorably against existing methods as well as results obtained using PVWatts' time-series dataset. We describe the results in detail in Section 4.6.

1.4 Organization of the Thesis

The organization of the rest of the thesis is given below.

In Chapter 2, we discuss the background work. We provide a literature survey of the area. In addition, to provide a practical perspective, we studied the nature of the faults described Section 1.2 from an operations and management perspective. Faults in PV arrays can be classified into a list of three safety categories. In addition to safety, we also assigned a Risk Priority Number (RPN) to each type of solar fault. Furthermore, the description of synthetic dataset generated by MATLAB Simulink model and the time series data from NREL's PVWatts dataset is provided, which are used to develop ML classification and topology reconfiguration algorithms in the subsequent Chapters. In order to study faults from an operations and management perspective, we classify faults using their Risk Priority Number (RPN).

In Chapter 3, we discuss the construction and development of the solar array test bed at the ASU Research Park. We describe the array which is fitted with Smart Monitoring Devices and collect data in real-time. These measurements include voltage, current, temperature and irradiance. We also describe the construction of a real-time load which can vary the maximum power point through the day. This load is programmed to switch among various configurations depending on the time of the day to allow for maximum power output. We detail the steps needed to obtain data in real-time from the solar array.

Chapter 4 provides a comprehensive study of various ML and signal processing techniques for fault detection and classification in solar arrays. The background on different classes of faults and their diagnostics is discussed. These faults are studied with the help of current-voltage and power-voltage curve characteristics. Next, we discuss and evaluate various ML techniques including k-means, k-nearest neigh-

bors, support vector machines, artificial neural networks, in terms of classification performance, as well as computational complexity.

Appendix A provides a detailed description on the safety parameters and best practices considered for the development and construction of the solar array. We also provide detailed instructions on the connections required to obtain real-time measurements.

Chapter 2

BACKGROUND AND LITERATURE SURVEY

The efficiency of solar energy farms requires detailed analytics and information on each panel regarding voltage, current, temperature, and irradiance. Monitoring utility-scale solar arrays was shown to minimize cost of maintenance and help optimize the performance of the array under various conditions. Faults in utility-scale solar arrays (Köntges *et al.* (2014, 2017); Kuitche *et al.* (2011); Mellit *et al.* (2018)) often lead to increased maintenance costs and reduced efficiency. Since photovoltaic (PV) arrays are generally installed in remote locations, maintenance and annual repairs due to faults incur large costs and delays. To automatically detect faults, PV arrays can be equipped with smart electronics that provide data for analytics. Smart Monitoring Devices (SMDs) (Takehara and Takada (2013)) that have remote monitoring and control capability have been proposed (Braun *et al.* (2012a)) to provide data from each panel and enable detection and localization of faults and shading. The presence of such SMDs renders the solar array system as a cyber-physical IoT networked system (Spanias (2017)) that can be monitored and controlled in real-time with algorithms and software.

2.1 Literature Survey

Traditional methods such as the Support Vector Machines (SVM) (Mellit *et al.* (2018)), decision tree based approach (Zhao *et al.* (2012b)), and the Minimum Covariance Determinant (MCD) distance metric (Braun *et al.* (2012b)) were proposed to identify fault conditions in PV arrays. Real-time fault detection in PV systems was studied in (Ali *et al.* (2017)), wherein a threshold based approach was developed

to identify faulty modules. Another statistical method in (Platon *et al.* (2015)) proposed a 3-sigma statistical rule for detecting faults in PV modules. Methods to detect partial shading in PV systems have been addressed in (Hariharan *et al.* (2016)). An unsupervised monitoring procedure for detecting anomalies in photovoltaic systems using a one-class SVM was shown in (Harrou *et al.* (2019)) and a semi-supervised graph approach for fault detection and classification was proposed in (Zhao *et al.* (2014)). Although the above methods provide encouraging results, they are based on aggregated data and generally cannot localize and distinguish between electrical faults and shading in PV systems. The ability to classify faults accurately and automatically with various PV array connection topologies is still a challenging problem (Braun *et al.* (2016)).

While neural networks (NNs) have been used in the past for fault detection and classification tasks (Rao *et al.* (2019); Mellit *et al.* (2018); An and Cho (2015); Zhao *et al.* (2014, 2012b)), the set of hyper-parameters to be chosen and the type of architecture is a challenge. As shown in Figure 2.1, the array can be used to collect data in real time. Data collected from the array is used for fault detection and classification studies. Switches with remote access also allow for dynamic topology reconfiguration. In this study, we use an autoencoder machine learning framework (Goodfellow *et al.* (2016)) to perform fault detection. An autoencoder is used to learn efficient representations (also called encodings) of the data through unsupervised dimensionality reduction. A decoder can then reconstruct the original input from the learned encoding. This unsupervised machine learning approach can be used to identify faults. We then implement fully connected NNs and dropout NNs (Srivastava *et al.* (2014)) trained specifically for fault classification in PV arrays. In our results section, we discuss performance based on accuracy and computational complexity in terms of weighted accuracy for various architectures. To reduce computation and redundancy

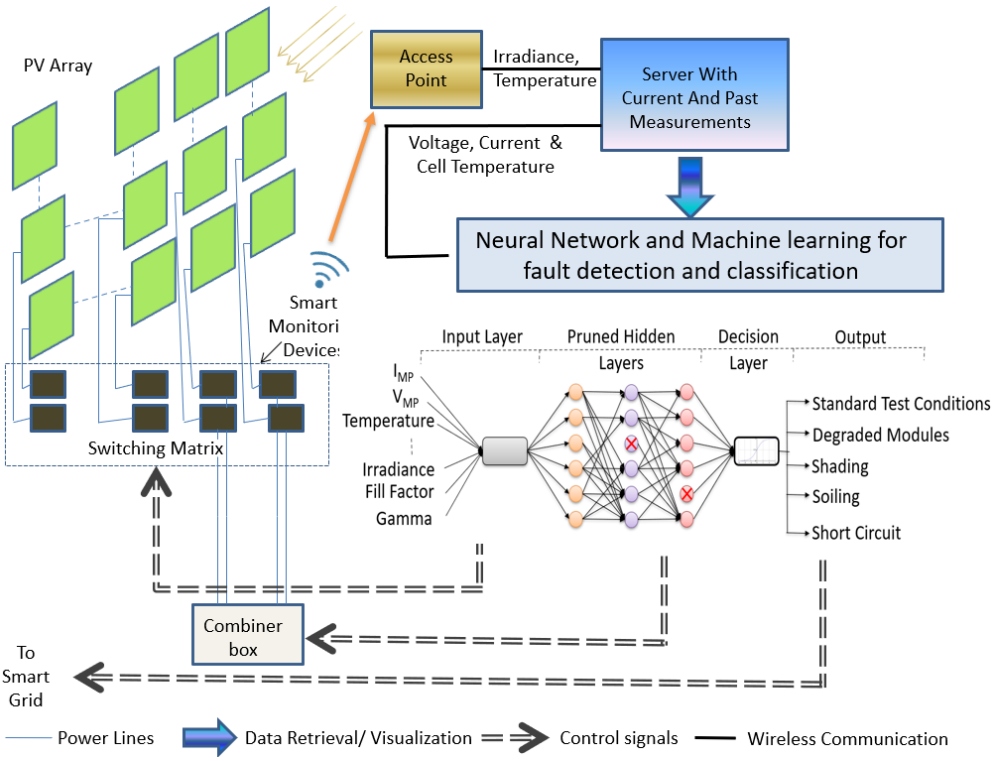

Figure 2.1: Smart Solar Array Monitoring System with Fault Detection and Classification Systems. The Autoencoder Is Used for Fault Detection While the Pruned Neural Network Is Used for Fault Classification.

and to customize the NN, we also perform network pruning using the *lottery ticket hypothesis* optimization process (Frankle and Carbin (2019)) to design sparse NN architectures. We achieve a 2× reduction in the size of the NN. Along with custom hardware, which enables monitoring voltage, current, temperature, and irradiance at the module level (Muniraju *et al.* (2017)), a custom NN with reduced parameters and high accuracy will be beneficial for the development of compact and specialized hardware for fault classification in PV arrays.

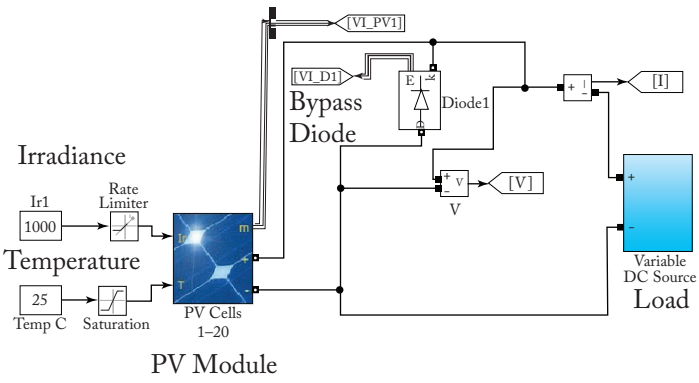

Figure 2.2: Simulink Model Used to Create the Fault Classification Synthetic Dataset. The Model Can Be Used to Generate Simulated Data for Various Shading and Fault Conditions Of a Single PV Module. An Array Can Be Created by Integrating Several Such Modules.

2.2 The Simulink Model for PV Fault Simulations

Simulated data to generate Maximum Power Points (MPPs) was obtained using MATLAB's Simulink model. The model is shown in Figure 2.2. The Simulink model will be used for data generation for various fault(King *et al.* (2004)) and shading conditions in the subsequent sections. The Simulink model uses the Sandia Flicker and Johnson (2013) PV module performance model. Through MATLAB, the user enters parameters for the Sandia model such as open circuit voltage (V_{OC}), short-circuit current (I_{SC}), temperature, and irradiance. The output of the module includes maximum voltage (V_{MP}) and maximum current (I_{MP}). The Simulink model can be used to generate synthetic data from a single PV module.

2.3 The PVWatts Dataset

In this section, we briefly discuss the data used in our experiments. We use the National Renewable Energy Laboratory's (NREL) PVWatts Calculator (Dobos (2014))

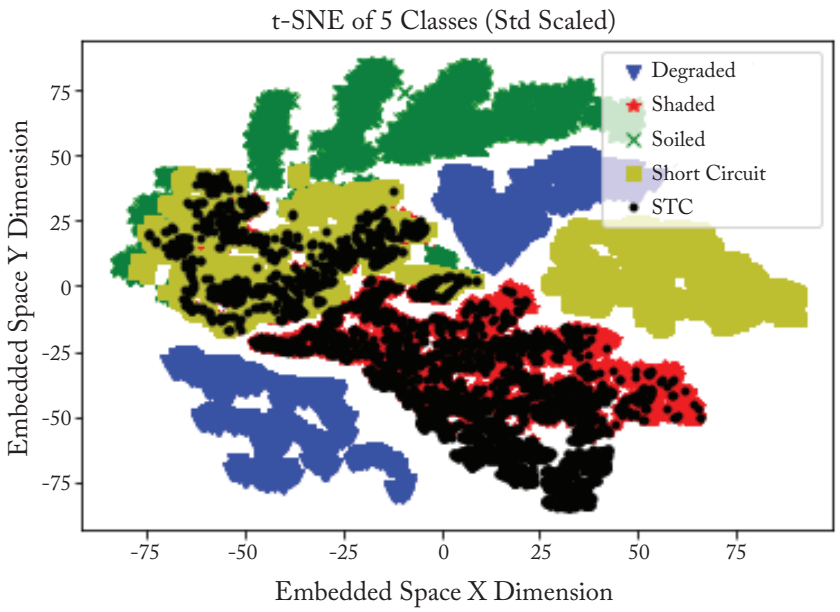

Figure 2.3: A t-distributed Stochastic Neighbor Embedding (t-SNE)(Maaten and Hinton (2008)) Plot Shows the Overlapping Data Points Between the Five Classes. We Project the Nine Dimensional Input Feature Matrix onto to a 2D Space and Visualize the Data Clusters.

which estimates the cost and amount of energy produced by grid-connected photovoltaic energy systems worldwide. The dataset available from PVWatts includes 4 faults used in this study, as well as the standard test conditions of PV arrays. We study these faults as these can not only be replicated using our solar array test bed described later in Section 3.2 but are also associated with high RPNs mentioned in Table 2.2. Faults are classified in terms of the following categories: shaded modules, soiled modules, short-circuited modules, and degraded modules. The data was obtained for a period of one year (January to December 2006) at a sampling duration of one hour. Data points include irradiance, temperature, and maximum power (P_{mp}) measurements along with a time stamp, amounting to 4000 hours of data.

Each data point was hand-labeled to one of the 4 PV array faults or as STC

(normal operation). Data points were labeled as STC if the measured irradiance was $1000W/m^2$ or has an ambient temperature of approximately $25°C$. A data point was considered shaded if the irradiance was lower than STC by 25% or more. If the measured irradiance was as per STC but the power measured was low, then the module was classified as soiled. Alternatively, if the irradiance and temperature were as per STC but the measured maximum current (I_{mp}) was low, then that data point was labeled as a short circuit or a line to line fault. Finally, if the measured open circuit voltage (V_{oc}) and or, short circuit current (I_{sc}) were lower than the rating of the PV module by 25% or more, the data point was classified as a degraded module (Platon *et al.* (2015)).

We consider a set of 9 custom input features, which includes maximum voltage (V_{mp}), maximum current (I_{mp}), measured irradiance, temperature, fill factor (FF), V_{oc}, I_{sc}, P_{mp} and Gamma (γ) - the ratio of power over irradiance. These features are derived from the IV-curves of the NREL's PVWatts Calculator (Dobos (2014)) dataset. In order to understand the data, we perform t-SNE to visually show that the data has overlapping faults as shown in Figure 2.3. This method projects the input 9-dimensional feature matrix into two dimensions by minimizing the Kullback-Leibler divergence of the data distributions between the higher and the mapped lower dimensional data (Maaten and Hinton (2008)).

In order to understand the data, we perform t-SNE to visually show that the data has overlapping faults, as shown in Figure 2.3. This method projects the input nine-dimensional feature matrix onto lower dimensions (2D) by minimizing the Kullback–Leibler divergence of the data distributions between the higher and the mapped lower dimensional data(Maaten and Hinton (2008)).

2.4 Operation and Management of PV Arrays

To provide a practical perspective, we studied the nature of these faults from an operations and management perspective. Faults in PV arrays can be classified into a list of three safety categories, as shown in Table 2.1 (Köntges *et al.* (2014)). In addition to safety, we also assigned a Risk Priority Number (RPN) to each type of solar fault.

This RPN is calculated as: RPN $= S \times O \times D$, where S denotes the severity (or a numerical subjective estimate of the effect of a failure), O denotes a numerical subjective estimate of likelihood of failure and D the numerical subjective estimate of detection. Failure modes with high RPN are more critical compared to the ones with lower RPN. Each S, O and D estimate is assigned a value between 1 and 10 (Dhillon (1999)).

Safety Category	Description
A	Failure has no effect on safety.
B (f,e,m)	Failure may cause fire (f), electrical shock (e) or physical danger (m) if failure repeats and/or second failure occurs.
C (f,e,m)	Failure causes direct safety problem.

Table 2.1: Broad Safety Categories in PV Arrays. Faults in Category C Have a Higher RPN as Shown in Table 2.2

.

Faults in this dissertation including shading, degradation and soiling can be con-

Fault Type	S	O	D	RPN
Standard Test Conditions (STC)	1	1	1	1
Soiling (Sepanski and et.al (2018))	8	3	6	144
Shading (Sepanski and et.al (2018))	1	6	5	30
Degradation (Chattopadhyay *et al.* (2014); Sepanski and et.al (2018))	2	10	8	160
Short Circuit (Shrestha *et al.* (2014); Rajput *et al.* (2019))	8	5	6	240

Table 2.2: RPN of All Faults Considered in This Dissertation. Higher RPN Could Indicate a Safety Category of Type B or Type C as Shown in Table 2.1

.

sidered as type A faults while short circuits are considered a type C (f,e,m) fault. The corresponding risk priority numbers are shown in Table 2.2.

We study faults with RPN as mentioned in Table 2.2. Since faults with high RPN possess a greater safety threat as shown in Table 2.1, detection and classification of these faults is critical.

Chapter 3

SOLAR ARRAY RESEARCH TESTBED

The efficiency of solar energy farms requires detailed analytics and information on each panel regarding voltage, current, temperature, and irradiance. Monitoring utility-scale solar arrays was shown to minimize cost of maintenance and help optimize the performance of the array under various conditions. We describe the design of an 18 kW experimental facility that consists of 104 modules fitted with smart monitoring devices. Each of these devices embeds sensors, wireless transceivers, and relays that enable continuous monitoring, fault detection, and real-time connection topology changes. The facility enables networked data exchanges via the use of wireless data sharing with servers, fusion and control centers, and mobile devices.

3.1 The SenSIP 18 kW Solar Array Testbed

Our vision for research monitoring and optimizing a large-scale PV array is summarized in Figure 3.1. To support experimental aspects of this research we designed a solar monitoring testbed (Rao *et al.* (2016)) at the ASU research park in Tempe, Arizona which is shown in Figure 1.2. This research facility consists of 104 modules in an 8×13 configuration that amounts to approximately 18 kW. Every panel in this solar array is equipped with an SMD. These devices are networked and can provide data to servers, control centers, and ultimately to mobile devices. Each SMD not only provides analytics for each panel but contains relays (actuators) that can be remotely controlled and via wireless access. Relays can bypass or change connectivity configuration, e.g., series to parallel. SMDs connected to each PV panel act as intelligent networked sensors (Spanias (2017)) providing data that can be used

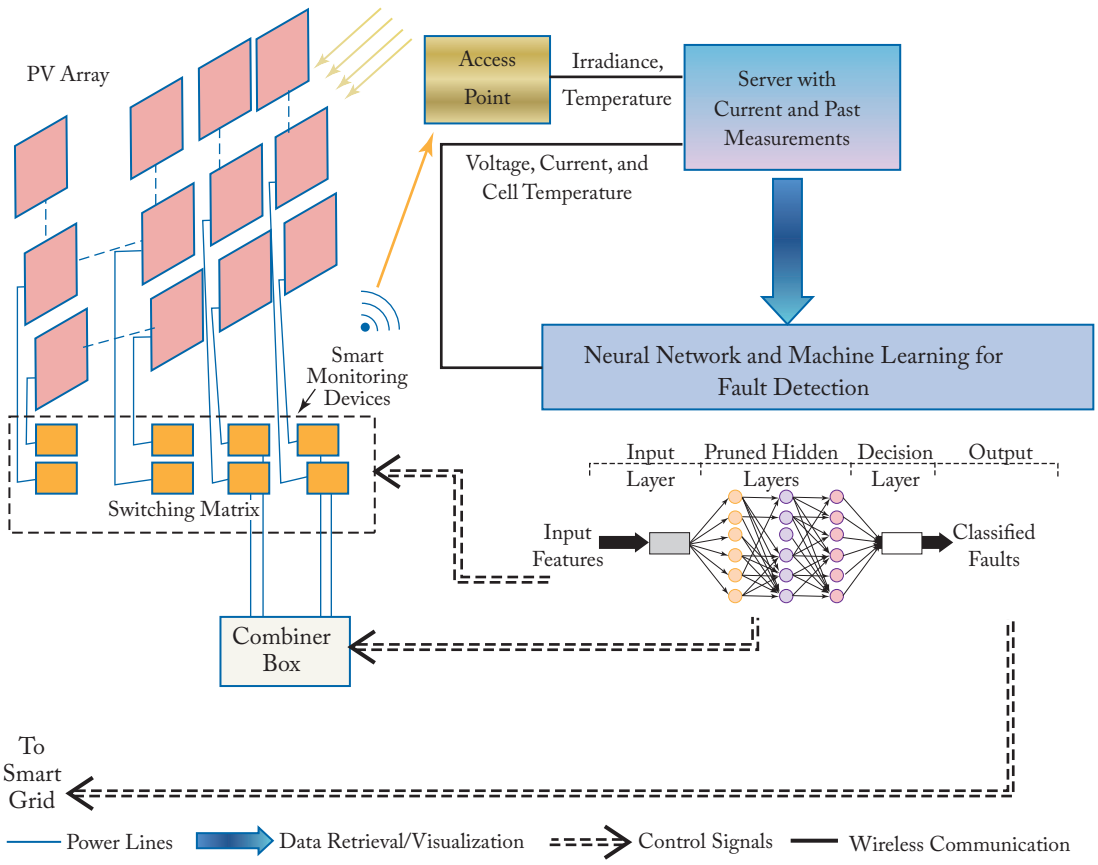

Figure 3.1: Smart Solar Array Monitoring System with Topology Reconfiguration, and Fault Detection and Diagnosis Systems.

to detect faults, shading, and other problems that cause inefficiencies. Each panel can be monitored individually to acquire voltage, current, and temperature, and all data is wirelessly transmitted to a central hub with minimal power loss. Additionally, each smart hardware device can reconfigure connections with its nearest neighbors. Data collected from the SMDs and reconfiguration testing will be used to design and evaluate automated fault detection, diagnosis, and mitigation algorithms.

Figure 3.2: Smart Monitoring Device (SMD) Attached to a Solar Panel(Takehara and Takada (2013)).

3.2 Design of the Solar Array Testbed

The research testbed is shown in Figure 1.2. This facility is built outdoors on the ground level of the ASU MTW building for ease of access by researchers. Each SMD and panel can be accessed from under the raised frame. The structure stands over 4 m tall at the tallest point, but is otherwise freely accessible. A weather monitoring station nearby records environmental conditions for fusion with collected PV data. This structure consists of 104 PV modules, each with an SMD, installed atop an elevated steel frame. Each SMD (Figures 3.2 and 3.3) can measure current, voltage, irradiance, and temperature of the associated panel. This data communicates to a server through a wireless network.

The facility is intended to enable experimental research with results obtained for various loading and shading conditions that will validate and extend various theoretical results reported in(Braun *et al.* (2012a); Peshin *et al.* (2015); Braun *et al.* (2016, 2012b)). Other studies that are based on a similar framework is reported in(Hammond *et al.* (1997); Kolodenny *et al.* (2008); Dirks *et al.* (2006)). The SMD as shown

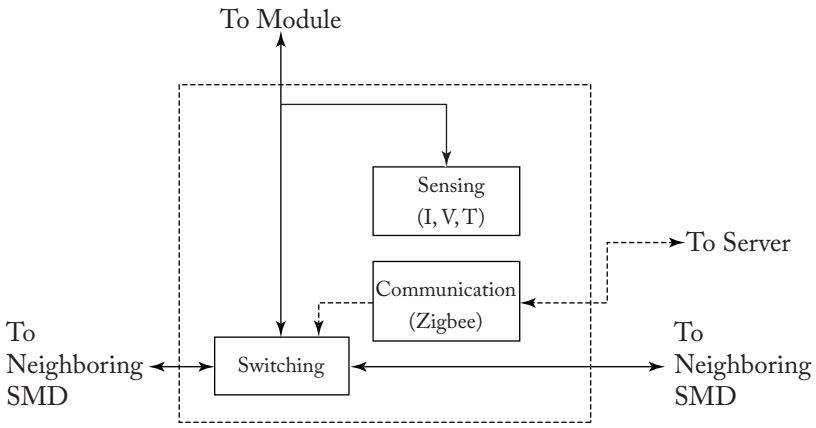

Figure 3.3: A Block Diagram of the Internal Connections of an SMD(Takehara and Takada (2012)).

in Figure 3.2, has six connectors. Two of the connectors are for the positive and negative leads for the associated panel and two leads each are assigned to the two neighboring SMDs. These interconnections allow for dynamic reconfiguration of the series and parallel strings.

Each SMD includes sensors and actuators (relays). Relays are used to change the topology configuration of the modules within the array. Three modes are available in the SMDs, i.e., series, parallel, and bypass. A faulty panel can easily be removed from the system to prevent mismatch losses by using the bypass mode. In some cases, the default topology may be suboptimal(Braun *et al.* (2016)). In these cases, the series and parallel modes are used to define an alternate topology. Neighboring modules are connected first in parallel and then in series, a configuration known as the total cross tied (TCT) topology.

Figure 3.4 shows a schematic of the communication between the SMDs and the server. Each SMD communicates wirelessly to an access point located at one of the PV modules. This access point in turn communicates with a central gateway which is connected to a server through USB.

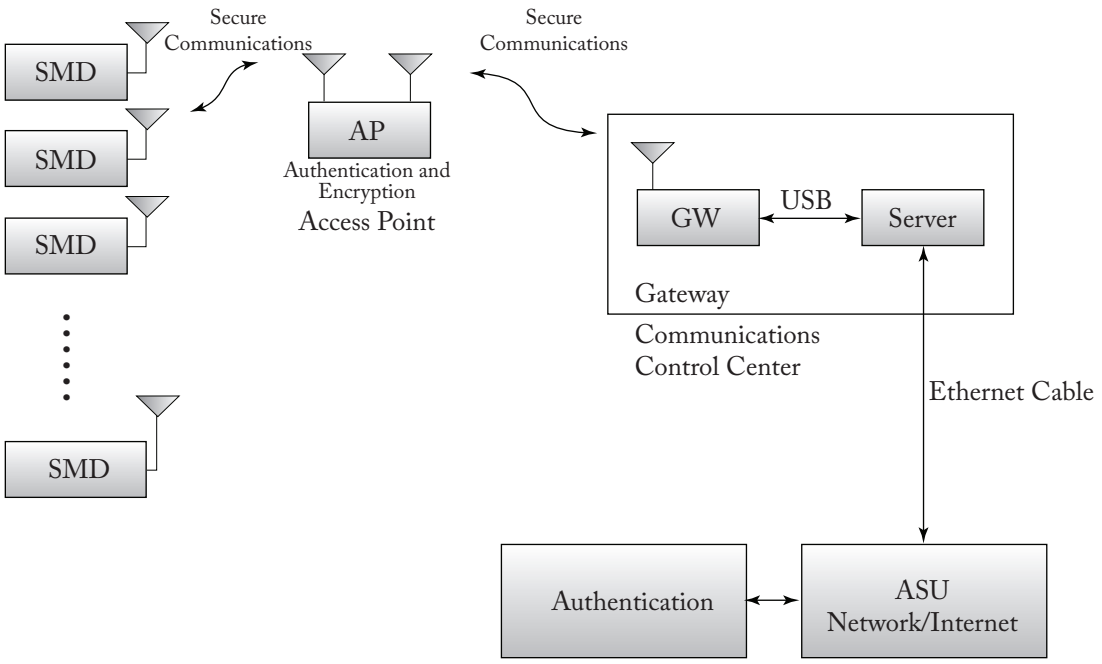

Figure 3.4: Block Diagram Depicts Communication Between SMDs and Server.

Each of the SMDs within the array is equipped with ZigBee wireless communication hardware. To minimize power consumption by the SMDs, the ZigBee transceivers do not transmit continuously. Instead, they periodically transmit voltage, current, and temperature measurements. A ZigBee hub device connected to the server receives all the reported data and transmits control signals to the networked SMDs. The newly built PV array facility is used to gather data for testing, training, optimization, and evaluation of algorithms. Common shading and fault conditions is safely generated in order to build a comprehensive dataset for designing and evaluating monitoring techniques.

The algorithmic and image/data analysis unit are equipped with various state of the art algorithms for imaging, data mining, and prediction that identify and track various important time-varying events and patterns. The algorithms operate on PV array measurements and on parametric models to detect and remedy faults using

SMD panel switching (Figure 3.1) or bypassing if necessary(Rao *et al.* (2016)).

Continuing work on machine learning for fault detection and classification, we investigate real world settings for detection and identification of faults in solar arrays. First, we built a load for a real-time scenario. The load is capable of MPP tracking and collect data through the day to validate the results discussed. We study the convergence of pruned neural networks and address the issues of overfitting in machine learning models.

3.3 Design and implementation of a Load for data collection

Data obtained include current and voltage readings from the PV array in real time. In addition, we also obtain irradaince values in real time at a sampling interval of $1s$. From the obtained current and voltage readings, we characterise the IV curve of the array and obtain the MPP. These obtained data points are used as inputs in various Machine Learning algorithms to detect and identify faults in PV Arrays. Data obtained helped in identify various loading and shading conditions along with faults as they lie along distinct regions in the two-dimensional space of the IV curve.

Figure 3.5 and 3.6 show the switching and control box developed to perform MPP tracking. Figure 3.7 shows the load bank which is controlled through the switches and the control box.

Data obtained includes current and voltage readings from the PV array in real time. From the obtained current and voltage readings, we can characterise the IV curve of the array and obtain the MPP. These obtained data points are used as inputs in various Machine Learning algorithms to detect and identify faults in PV Arrays. Data obtained will thus help identify various loading and shading conditions along with faults as they lie along distinct regions in the two-dimensional space of the IV curve. We discuss these results in Chapter 4. Figure 1.2 shows the PV array

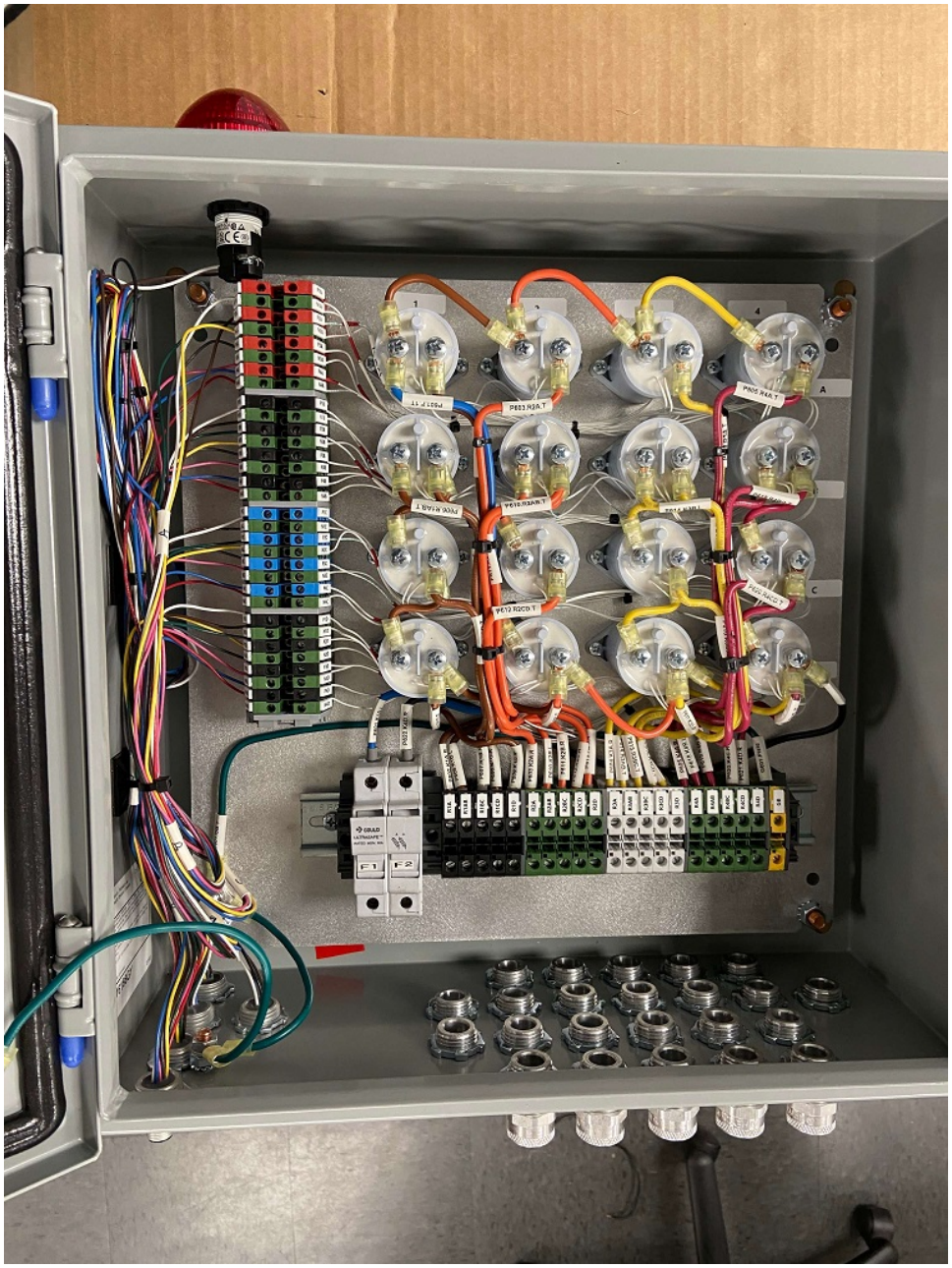

Figure 3.5: The Load Is Controlled Through High Powered DC Switches. These Switches Have a Rating of $600w$ Each. They Are Dynamically Controlled Through the Control Box Using the PLC.

currently set up at ASU Research Park. The load is placed under the array and can perform real time switching between array configurations to optimize power output.

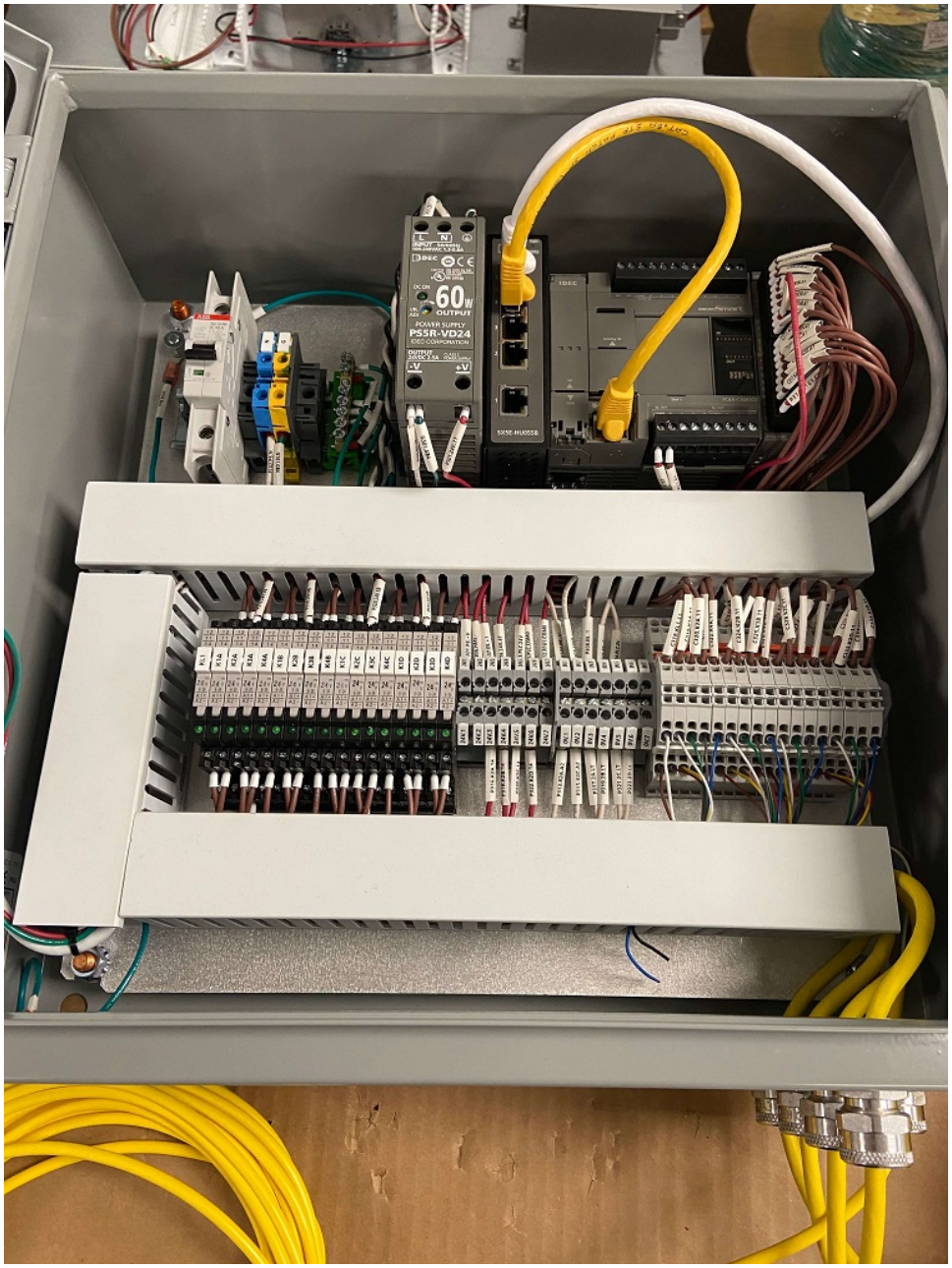

Figure 3.6: The MPP Tracking Is Controlled Automatically Through the Control Box Shown Here. The PLC Controls the Switches and Will Either Turn the Resistors on or off Depending on the Time of the Day.

These loads are programmable which can switch between different resistor values to suit the MPP for an array configuration and can also switch between load banks to

Figure 3.7: A Design of a 7 Resistor Load Bank. Four of the Resistors Are Always on and the Remaining Set of Resistors Turn on or off to Implement Various Conditions for Testing.

suit another array configuration.

Chapter 4

FAULT DETECTION AND CLASSIFICATION USING MACHINE LEARNING

Detecting faults in PV is important for the overall efficiency and reliability of a solar power plant. Ground faults, series and parallel arc faults, high resistance connections, soiling, and partial shadowing need to be detected. The I-V data in a PV array can be measured at the panel-level. This data is useful in predicting possible ground faults or arc faults. The I-V characteristic is a function of temperature, incoming solar irradiance (direct and diffused), angle of incidence, and the spectrum of sunlight. The panel has an optimal operating point for maximum power. Fault detection using I-V data can be accomplished by identifying outliers in the I-V feature space. I-V measurements are typically highly correlated. Moreover, the dynamics of the I-V measurements lend themselves to predictive models. Current practice is to identify faults via a human operator examining data collected at the inverter. One study identified a Mean Time to Repair (MTTR) of 19 days(Braun *et al.* (2012a)) for a centrally monitored system of residential installations. With the addition of more and higher quality data from SMDs, MTTR could be significantly reduced. Several challenges and research opportunities are evident in the fault diagnosis and localization problems. First, of course, a system must accurately classify the PV array's condition. It should be able to react to the "unknowns"—faults the system designers did not anticipate. Considering these challenges, several ML approaches can be examined. Simulated fault datawere obtained using the Sandia PV module performance model and a MATLAB circuit simulation package(Fan *et al.* (2020a,b)).

4.1 PV Fault Detection using Autoencoders

We propose the use of an autoencoder for fault detection. An autoencoder is an unsupervised learning algorithm designed to identify faults based on reconstruction errors. An autoencoder consists of an encoder and a decoder. A simple schematic of an encoder can be seen in Figure 4.1. The encoder maps the input to a lower dimensional embedded space also called latent space and the decoder maps the latent space to the original input space. The difference between the original input and the reconstructed output can be used to identify anomalies in the data (An and Cho (2015)) and hence detect the presence of faults.

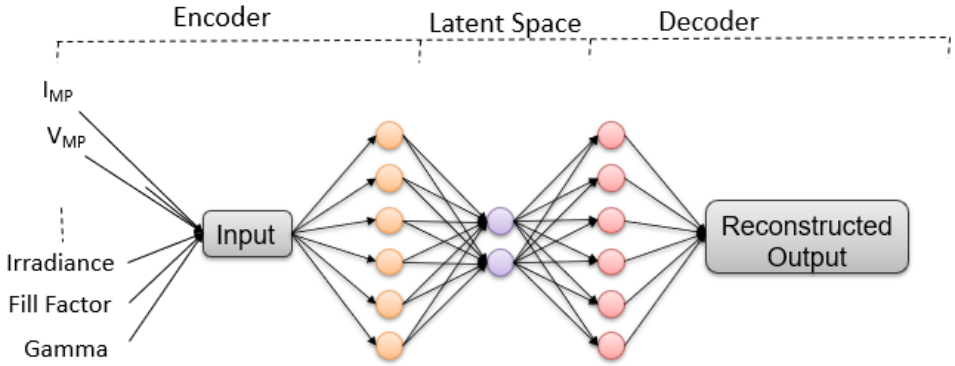

Figure 4.1: A Figure Illustrating an Autoencoder Used for Fault Detection. The Original Input Is Mapped to a Lower Dimension (Also Called Latent Space). The Reconstructed Output Maps the Latent Space Back to the Original Input Space. We Detect Faults Based on Reconstruction Errors. Higher Reconstruction Error Indicates the Presence of a Fault.

We train a three layered autoencoder for fault detection. The nine dimensional input feature matrix is given as an input to the autoencoder. The autoencoder is trained on STC irradiance data while the fault data is treated as anomalous and is

used to test the algorithm. The latent space consists of two neurons and the decoder maps the latent space to the original input dimensions. As seen in the error histogram in Figure 4.2 and Figure 4.3, we detect faults based on reconstruction errors. We observe that while STC irradiance data have low reconstruction errors, fault data have higher reconstruction errors. Using this method, we can identify anomalous data from observed measurements and hence detect the presence of faults.

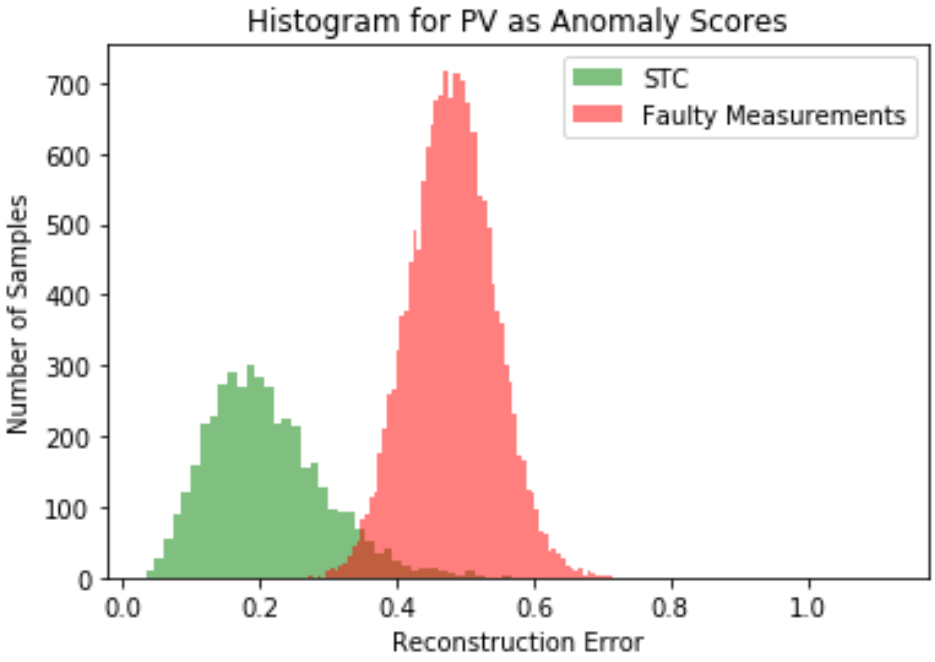

Figure 4.2: PV Fault Detection Using an Autoencoder on NREL Data. An Autoencoder Is Used for Fault Detection. Samples from the Same Class Have Lower Reconstruction Error While Samples from Fault Classes Have Higher Reconstruction Error.

4.2 Faults in PV Arrays

In this section, we review the standard test conditions and the commonly occurring faults namely, shading, degraded modules, soiling, and short circuits. STC values

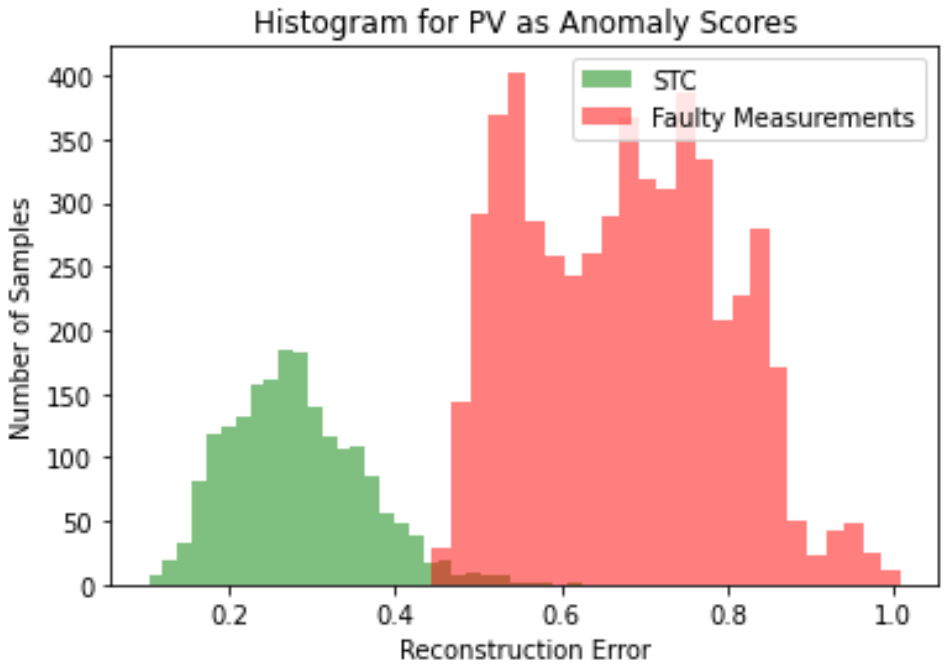

Figure 4.3: PV Fault Detection Using an Autoencoder on Real Data. An Autoencoder Is Used for Fault Detection. Samples from the Same Class Have Lower Reconstruction Error While Samples from Fault Classes Have Higher Reconstruction Error.

correspond to the measurements yielding maximum power under the temperature and irradiance values of a particular day.

4.2.1 Standard Test Conditions

Standard Test Conditions (STCs) are the industry standard for the conditions under which a solar panel are tested. By using a fixed set of conditions, all solar modules can be more accurately compared and rated against each other. The temperature of the cell is taken to be 25°C and the irradiance is 1000 W/m^2. STC values correspond to the measurements yielding maximum power under the temperature and irradiance values of a particular day. Data points are labeled as STC if the irradiance, temperature, and power were the highest possible values for that particular day. Figure 1.2

Figure 4.4: Example of a Soiled Solar Panel.

shows a solar array under the STC.

4.2.2 Soiling

Since PV modules are exposed to the environment, modules get soiled due to dust, snow, and bird droppings accumulating on the PV module as shown in Figure 4.4. While the irradiance measured remains the same as STC, the power produced drops significantly. The solution to this problem involves manually cleaning the modules regularly. If the measured irradiance was as per STC but the power measured was low, then the module was soiled. Soiling is caused by dry deposition affects the power output of PV modules, especially under dry and arid conditions that favor natural atmospheric aerosols (wind-blown dust)(Cordero *et al.* (2018)).

4.2.3 Degraded Modules

Degraded modules are a result of modules aging or regular wear and tear of the PV modules, as shown in Figure 4.5. Consequently, such modules produce lower power values owing to the lower values of open-circuit voltage V_{OC} and short-circuit current I_{SC}. If the measured open circuit voltage (V_{OC}) and or short-circuit current

Figure 4.5: Example of a Degraded Solar Panel.

(I_{SC}) were lower than the rating of the PV module by 25% or more, the data point was classified as a degraded module. Solar modules degrade by approximately 1% per year; however, if the measured current is less than 20% of the expected value after adjusting for sunlight conditions then the module maybe failing(Mellit *et al.* (2018)).

4.2.4 Arc Fault

An Arc-circuit fault is mainly due to bad wiring in a PV string or between PV strings. This not only causes significant power loss but also creates potential fire hazards and severe damage to the modules, as shown in Figure 4.6. To improve power production, the efficiency of the solar array, and prevent safety hazards, identifying and localizing these faults automatically is critical. If the irradiance and temperature were as per STC but the measured maximum current (I_{MP}) was low, then that data point was labeled as a short circuit or a line to line fault. Short circuit in the wiring is a bad or loose connection, incorrect wiring, or an internal problem with the solar module. It is possible that the connection point is sufficient enough for full voltage reading, but limited current(Mellit *et al.* (2018)).

Figure 4.6: Example of Damage from an Arc Fault.

4.2.5 Shading

Shading is a serious concern in PV arrays. A module is shaded if the irradiance measured is considerably lower than STC, usually caused by overcast conditions, cloud cover, and building obstruction. As a result, the power produced by the PV array is significantly reduced. A data point was considered as shaded if the irradiance measured was lower than STC by 25% or more. Figure 4.8 shows a module under various shading conditions. Because the PV module output current is completely dependent on the amount of sunlight and varies linearly with the sunlight conditions available, shading considerations are extremely important when designing and siting the location of an array installation. The cells of a solar module are wired in series and the maximum output current is dependent on the weakest cell, as the current is the same through each cell. The module maximum output current is dependent on the maximum current available from the weakest cell. Therefore, if a single cell in a PV module is shaded, the output current from the entire module goes to zero. If any part of cell is shaded, the output current from the PV module is reduced by the proportional amount that the cell is shaded(Patel and Agarwal (2008); Nguyen and Lehman (2006); Quaschning and Hanitsch (1996)).

Figure 4.7: Example of a Partially Shaded Solar Panel Array.

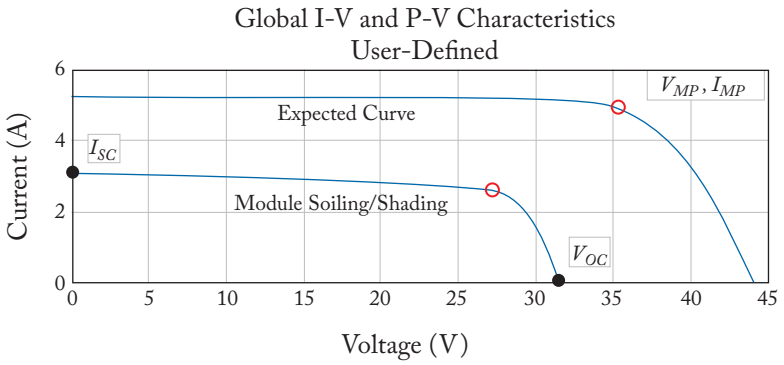

Figure 4.8: I-V Curves of the PV Module under Shading Conditions.

4.3 Key Contributions in Machine Learning for PV Applications

The use of ML in fault diagnosis can be formulated as a multiple hypothesis testing problem. ML is useful for the detection and the identification of the type of the fault. For example, if one of the arrays receives less sunlight due to shading, ML could help identify the error in the shading conditions. It was previously shown that fault detection can be performed using statistical outlier detection techniques(Braun *et al.*

(2012b)). However, performing diagnosis and localization of a fault is a much deeper problem. It requires data on array behavior under each fault condition. Moreover, PV arrays come in all shapes and sizes and may behave very differently from one another under similar fault conditions. A comprehensive PV fault dataset does not currently exist. Since array operators are rarely involved in academic research and may wish to keep the performance of their systems proprietary. Gathering data from fault conditions is difficult to obtain unless continuous monitoring is enabled. Finally, the overwhelming majority of arrays are fitted with I-V sensors only at the inverter, allowing minor faults which do not cause a large drop in output to persist undetected. Studies that attempt to quantify the likelihood and severity of different conditions were reported in(Maish *et al.* (1997)). On the other hand, extensive work has been done to characterize the behavior of normally operating modules and arrays(King (1997)).

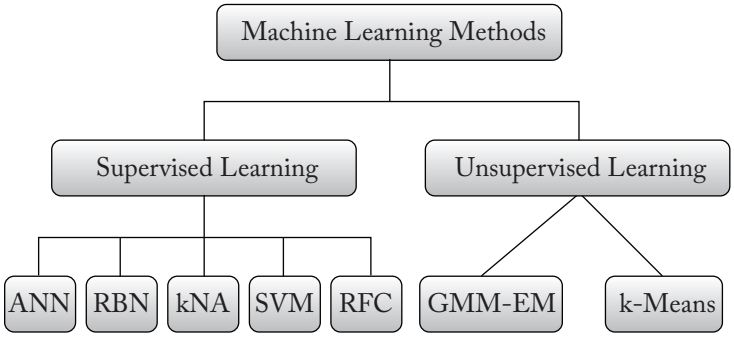

Figure 4.9: Algorithms Considered in This Dissertation for Fault Classification Are Artificial Neural Networks (ANN), K-nearest Neighbor Algorithm (KNA), Support Vector Machine (SVM), Random Forest Classifier (RFC), Radial Basis Network (RBN), Gaussian Mixture Model—expectation Maximization Algorithm (GMM-EM), And the k-means Algorithm(Bishop (2006)). These Algorithms Are Used to Identify Shading and Fault Conditions in PV Arrays.

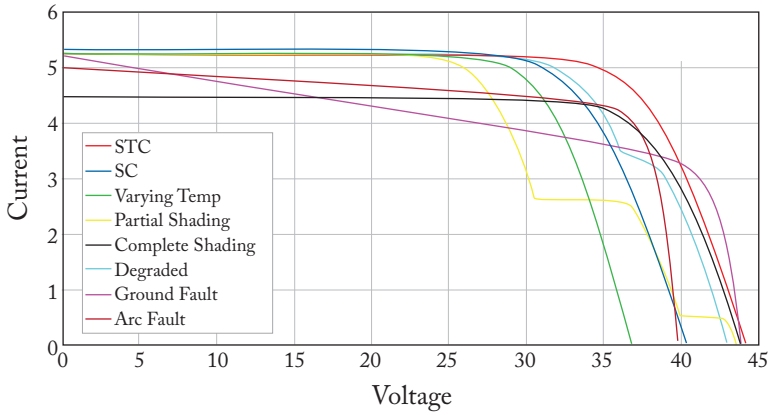

Figure 4.10: I-V Curves of the PV Module under Various Fault and Shading Conditions.

A classification algorithm for fault detection must have the following properties. First it must accurately classify the PV array's condition. It must be adaptable to different array configurations without extensive data collection for each individual array. It must be able to recognize each fault class from a very small number of

training examples. It should take advantage of our prior knowledge of the electrical behavior of PV arrays (e.g., equal current within a string), rather than having to learn these relationships through the training data. It should be capable of reacting to the "unknown unknowns," i.e., faults the system designers did not anticipate. In light of these requirements, several ML approaches are worth examining. Semi-supervised learning could allow the generation of many realistic faults from a few measured examples(Zhao *et al.* (2014)). This would mitigate the problem of lopsided data, where very few examples of faults are available. We study a number of such algorithms in this dissertation, as shown in Figure 4.9. These algorithms can help identify multiple PV conditions using their I-V curves. Figure 4.10 shows the I-V curve of the PV module under various loading, fault and shading conditions.

ML algorithms are widely classified as supervised, semi-supervised, and unsupervised algorithms. In supervised learning, "true" or "correct" labels of the input dataset are available. The algorithm is "trained" using the labeled input dataset (training data) which means ground truth samples are available for training. In the training process, the algorithm makes appropriate predictions on the input data and improves its estimates using the ground truth and reiterating until the algorithm reaches a desired level of accuracy. In almost all the ML algorithms, we optimize a cost function or an objective function. The cost function is typically a measure of the error between the actual output and the algorithm estimates. By minimizing the cost function, we train our model to produce estimates that are close to the correct values (ground truth).

In the case of unsupervised algorithms, there are no explicit labels associated with the training dataset. The objective is to draw inferences from the input data and then model the hidden or the underlying structure and the distribution in the data in order to learn more about the data. Clustering is the most common example

of an unsupervised algorithm. Semi-supervised learning is an approach to ML that combines a small amount of labeled data with a large amount of unlabeled data during training. Semi-supervised learning falls between unsupervised learning (with no labeled training data) and supervised learning (with only labeled training data).

4.3.1 The k-Means Algorithm

The k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. The k-means algorithm is used to partition a given set of observations into a predefined amount of k clusters. The algorithm as described by(James (1967)) starts with a random set of k center-points (μ). During each update step, all observations x are assigned to their nearest center-point (see Equation 4.1). In the standard algorithm, only one assignment to one center is possible. If multiple centers have the same distance to the observation, a random one would be chosen:

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - \mu_i^{(t)} \right\|^2 \leq \left\| x_p - \mu_j^{(t)} \right\|^2 \ \forall j, 1 \leq j \leq k \right\}. \tag{4.1}$$

Afterward, the center-points are repositioned by calculating the mean of the assigned observations to the respective center-points:

$$\mu_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j. \tag{4.2}$$

The update process reoccurs until all observations remain at the assigned center-points and therefore the center-points would not be updated anymore. Figure 4.11 shows the use of the k-means algorithm to identify ground faults and arc faults from MPPT datapoints. The k-means algorithm can accurately detect and identify faults by forming clusters on the I-V curve.

While generating MPPs, we consider a variance of ± 5 V for V_{MP} and a variance of ± 1 A for I_{MP} to account for variability in real-time scenarios(Mekki *et al.* (2016)).

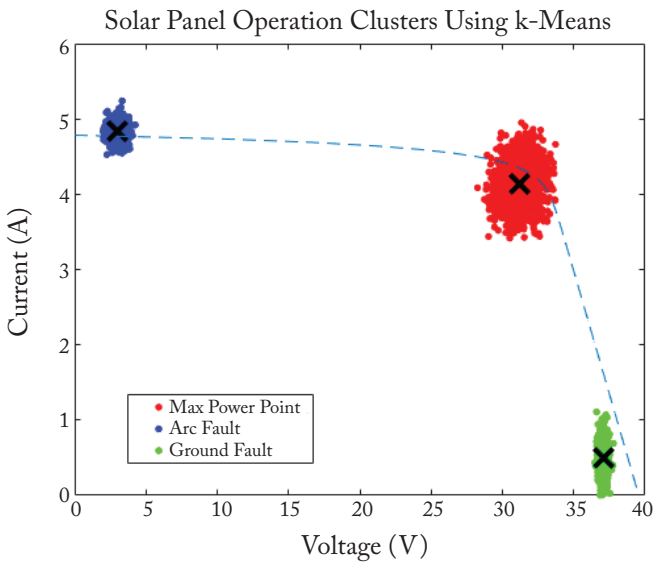

Figure 4.11: Fault Classification Using the k-means Algorithm. Using the k-means Algorithm, We Identify Three Clusters In the I-V Curve.

To simulate a varying temperature panel, the simulated panel was assigned a higher temperature value. The data was obtained and trained with the k-means algorithm. The results obtained are shown in Figure 4.12. Each set of data points represent one condition associated with the PV array. Using k-means with voltage, current, and temperature as our three axes, we successfully identify ground faults (Gnd), arc faults (Arc), standard test conditions with irradiance at 1000 W/m^2, and a module temperature of 25°C (STC), shaded conditions (shading), and varying temperature conditions.

However, certain other conditions such as soiling and short circuits are not identified using this method due to the lack of labels in the dataset. Soiling and short-circuit condition have MPPs which lie in similar areas in the 2D I-V curve space. The k-means algorithm in this setting also does not identify partial shading vs. complete shading of modules. Therefore, there is a need for the use of neural network algorithms to detect and identify faults in PV arrays.

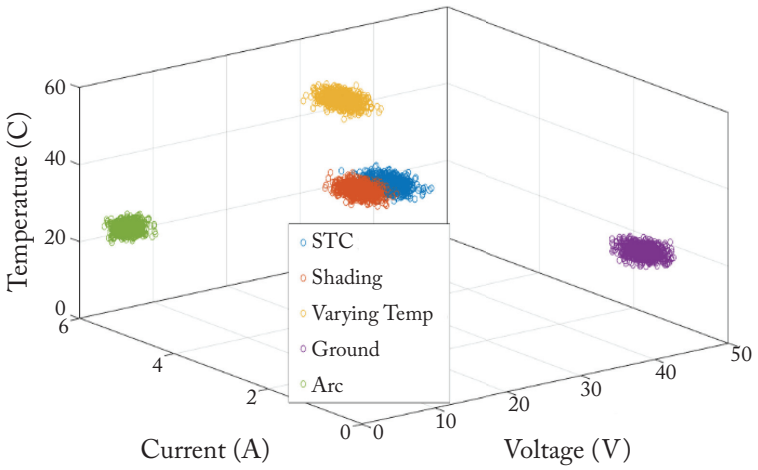

Figure 4.12: Clustering Using the k-means Algorithm. The Synthetic Data Was Obtained Using the Simulink Model Described in Section 2.2.

4.3.2 The Kernel SVM

Kernel SVM is a soft margin classifier robust to outliers. Computing the soft margin classifier is equivalent to minimizing the loss function,

$$\mathcal{L}_{svm} = \frac{1}{n}\left[\sum_{i=1}^{N} \max\big(0, 1 - y_i(w \cdot \phi(x_i) - b)\big)\right] + \lambda||w||^2, \tag{4.3}$$

where λ is a hyper-parameter which regularizes the weights and $\phi(\cdot)$ is the kernel function. Loss function in Equation (4.3) can be reduced to a quadratic programming problem and solved by a convex solver. Common choices of kernel functions $\phi(\cdot)$ are polynomial kernel, Gaussian radial basis kernel, and hyperbolic tangent kernel. Success of SVM depends on the right choice of kernel, which is hard to select for a given data set(Cortes and Vapnik (1995)).

4.3.3 The k-Nearest Neighbor Algorithm

The k-nearest neighbor algorithm (kNA) is a simple nonparametric classifier, where classification is based on local membership scores. In training phase, simi-

larity measure for each data point with its closest k neighboring data points is stored. To classify a test sample, similarity measure between the test sample and all the data points are calculated, and the class label assigned is the label corresponding to the majority of k-closest samples based on the similarity score. Similarity score is generally computed using Euclidean, Manhattan, Minkowski, and Hamming distance. The main drawback of k-nearest neighbor (KNN) method is the large computation time during test phase(Altman (1992)).

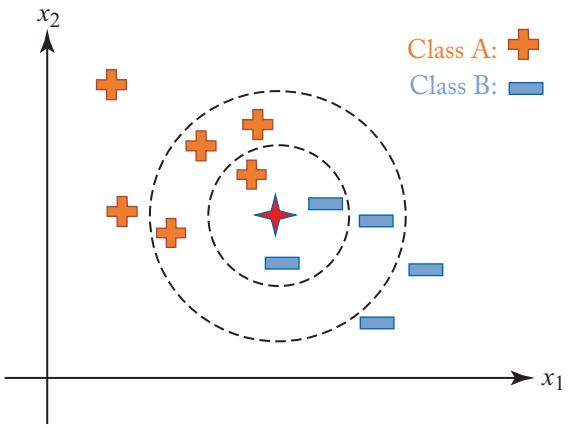

Figure 4.13: A Simple kNA Model for Different Values of k. For $k = 3$, the Test Point (Star) Is Classified as Belonging to Class B and for $k = 6$; The Point Is Classified as Belonging to Class A.

4.3.4 Random Forest Classifiers

The Random forest classifier (RFC) is a classification algorithm based on an ensemble of decision trees. A decision tree is constructed by set of input features randomly sampled batch of data from the dataset. To classify a test sample, each decision tree provides a vote for a particular class, and the label assigned is the class which has the majority of the votes. RFC involves two hyper-parameters: number of decision tress and the depth of the decision tree. RFCs are capable of modeling

complex data sets and are robust to outliers(Ho (1995)).

4.3.5 *Radial Basis Function Networks*

The Radial Basis Networks (RBNs) are nonlinear classifiers that use radial basis functions as the activation functions of the hidden layers. The RBN is a supervised learning algorithm, where each point in the dataset is passed through the network and labeled with its true classification. This network classifies by measuring the similarity of the input vectors to the labeled examples gathered from the training set Pedersen *et al.* (2019). A simplified representation is shown in 4.14.

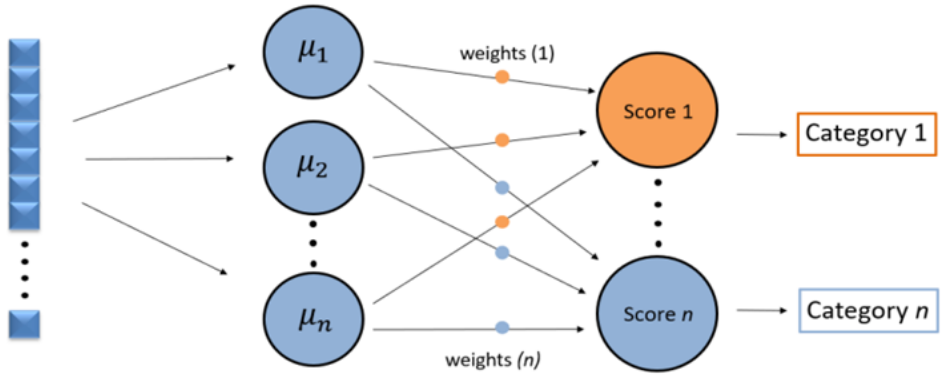

Figure 4.14: Simplified Representation of the RBFN Architecture. In Our Case, the Radial Basis Function Is Used as the Activation Function..

4.4 Key Contributions in Neural Networks

Various signal processing and statistical methods have been developed for detection and identification of faults in utility scale PV arrays. However, there is a need for a comprehensive algorithm which captures a wide variety of faults. While several methods have been proposed in the past for fault detection, neural networks aim to detect and identify the type of fault occurring in PV arrays. Figure 4.10 shows the I-V curve for the multi-class classification problem. While traditional signal process-

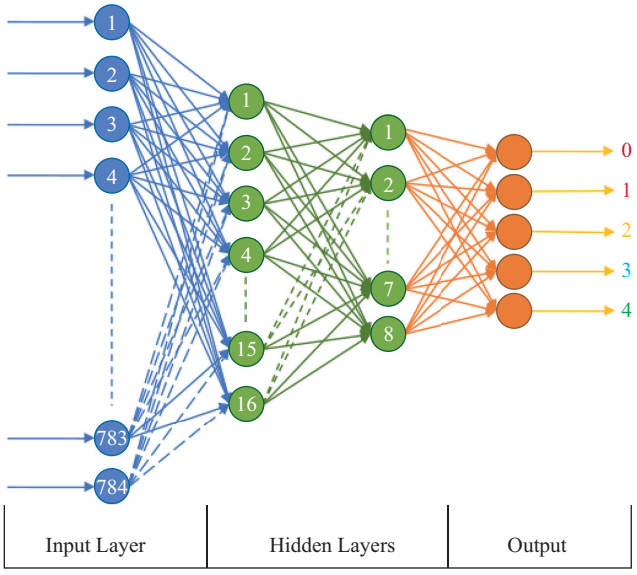

Figure 4.15: An Example of a Neural Network. This Neural Network Has One Input Layer, Two Hidden Layers, and One Output Layer.

ing algorithms use the statistical properties of a single I-V curve of a given module, most methods do not cover multiple cases. Using neural networks allows not only detection but identification of the fault type with a high accuracy(Rao *et al.* (2019, 2020b)). Previous studies that used neural nets have been used to make binary decisions on fault detection, i.e., detect faults but not classify the type of fault(Mekki *et al.* (2016); Chine *et al.* (2016); Chen *et al.* (2017); Hariharan *et al.* (2016)).

4.4.1 The Feature Matrix

Studies show that nine inputs namely V_{OC}, V_{MP}, I_{SC}, I_{MP}, temperature of module (Temp), irradiance of module (Irr), Fill Factor (FF) (a ratio of the product of the short circuit current (I_{SC}), and open circuit voltage (V_{OC}) over product of V_{MP} and I_{MP}), gamma (γ)—the ratio of power over irradiance, and power, to classify eight different faults. The eight faults classified are ground fault (Gnd), arc fault (Arc), complete module shading (Fully Shaded), partial module shading (Partial Shading), varying

temperatures of module (Varying Temp), soiling (Degraded), short circuits (SC), and standard test conditions with irradiance at 1000 W/m^2 and a module temperature of 25°C (STC). An example of a row vector and their corresponding class is shown in Table 4.1.

Data Labeling:

The data points were labelled as belonging to one of the five classes (i.e., standard test conditions (STC), shaded, soiled, short circuit and degraded) based on the input feature vector. The data points were labeled as:

1. *STC*: If the measured irradiance was 1000 W/m^2 or has an ambient temperature of approximately 25°C.

2. *Shaded*: If the irradiance was lower than STC by 25% (i.e., lower than 750 W/m^2) or more.

3. *Soiled*: If the measured irradiance was as per STC (i.e., 1000 W/m^2 or 25°C) but the power output was less than 25% of power output under STC conditions.

4. *Short circuit fault*: If the irradiance and the temperature were as per STC, but the measured maximum current I_{mp} was less than 25% of measured maximum current I_{mp} at STC.

5. *Degraded module*: If the measured open circuit voltage V_{oc} or, short circuit current I_{sc} were lower than the rating of the PV module by 25% or more.

Table 4.1: An Example of the Row Vector and Their Corresponding Class. Each Such Row Vector Is Classified into One of the Five Classes.

V_{MP}	I_{MP}	Temp	Irr	FF	γ	P_{MP}	V_{OC}	I_{SC}	Class
36.33	1.36	25°C	281.11	4.66	0.17	49.77	44.33	5.242	STC
36.33	1.02	25°C	281.11	6.62	0.13	37.33	44.48	5.56	Soiled
36.33	1.23	25°C	281.11	5.05	0.15	44.79	44.26	5.11	SC
36.33	1.02	25°C	210.83	6.62	0.17	37.33	44.48	5.56	Shaded
36.33	1.09	25°C	281.11	2.59	0.14	39.81	28.80	3.58	Degraded

4.5 Real Time Experiments

For normal operations:

Standard Test Conditions (STCs) are the industry standard for the conditions under which a solar panel are tested. By using a fixed set of conditions, all solar modules can be more accurately compared and rated against each other. STC values correspond to the measurements yielding maximum power under the temperature and irradiance values of a particular day. Data points are labeled as STC if the irradiance, temperature, and power were the highest possible values for that particular day. For normal operations, we propose to:

1. Keep the array on, vary the resistors depending on time of the day.
2. Record measurements- V_{MP}, I_{MP}, temperature. These measurements are recorded by the Smart Monitoring Device (SMD) in the array developed at ASU Research Park.
3. Record irradiance measurements using the TES132 meter. The meter gives values at a sampling rate of 1 second. These readings are beneficial in identifying soiling versus shading modules.

Shading:

Shading is a serious concern in PV arrays. A module is shaded if the irradiance measured is considerably lower than STC, usually caused by overcast conditions, cloud cover, and building obstruction. As a result, the power produced by the PV array is significantly reduced. To run shading experiments, we propose to:

1. Run a piece of obstruction over a few modules. Scenarios include partial shading and complete shading. 25% of the modules are covered during this process. An irradiance value drop of 25% or more shows significant power loss under shading conditions.

2. Measurements will include V_{MP}, I_{MP} and temperature. Our experiments on the PVWatts dataset have shown that neural networks are effective in classifying shaded modules with high accuracy. Figure 4.16 shows a PV panel shaded at the ASU Research Park.

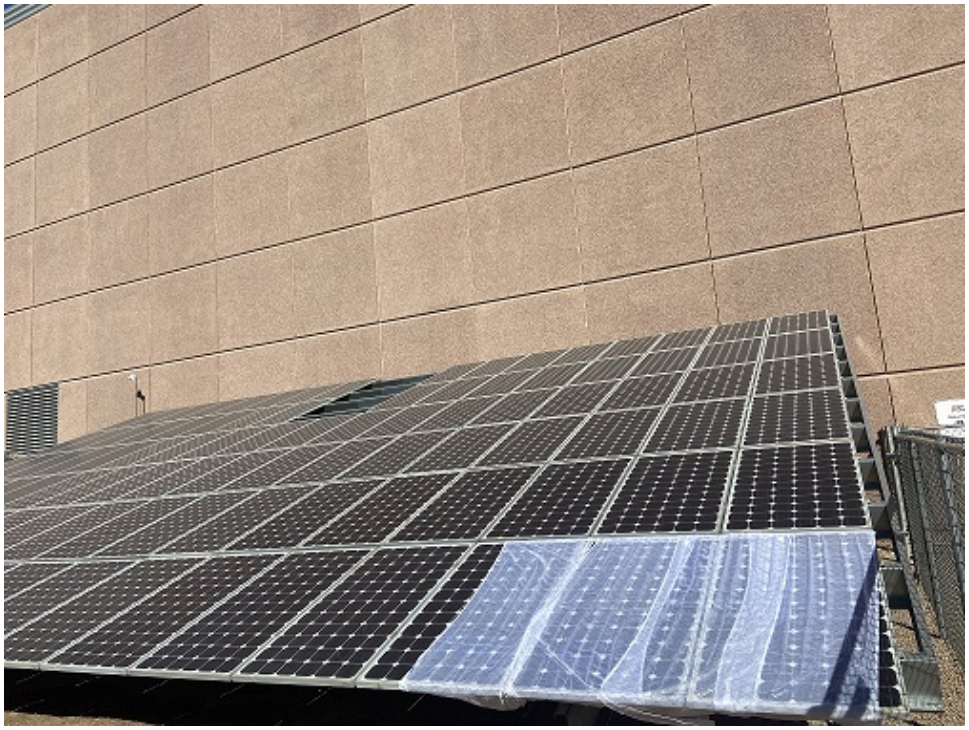

Figure 4.16: An Example of a Simulated Shaded Module at ASU Research Park. This Corresponds to 25% Shading.

However, significantly covering the module could turn off the array and not record any values. Therefore, we propose to cover not more than 50% of the module at a time.

Soiling:

While the irradiance measured remains the same as STC, the power produced drops significantly. The solution to this problem involves manually cleaning the modules regularly. If the measured irradiance was as per STC but the power measured was low, then the module was soiled. Soiling is caused by dry deposition affects the power output of PV modules, especially under dry and arid conditions that favor natural atmospheric aerosols (wind-blown dust). For soiling experiments, we propose to:

1. We use the uncleaned modules for these experiments. These modules have dust particles on them which are a result of frequent dust storms.

2. Record the same measurements as provided by the SMD which include V_{MP}, I_{MP} and irradiance. An illustration of soiling versus STC modules is shown in Figure 4.17.

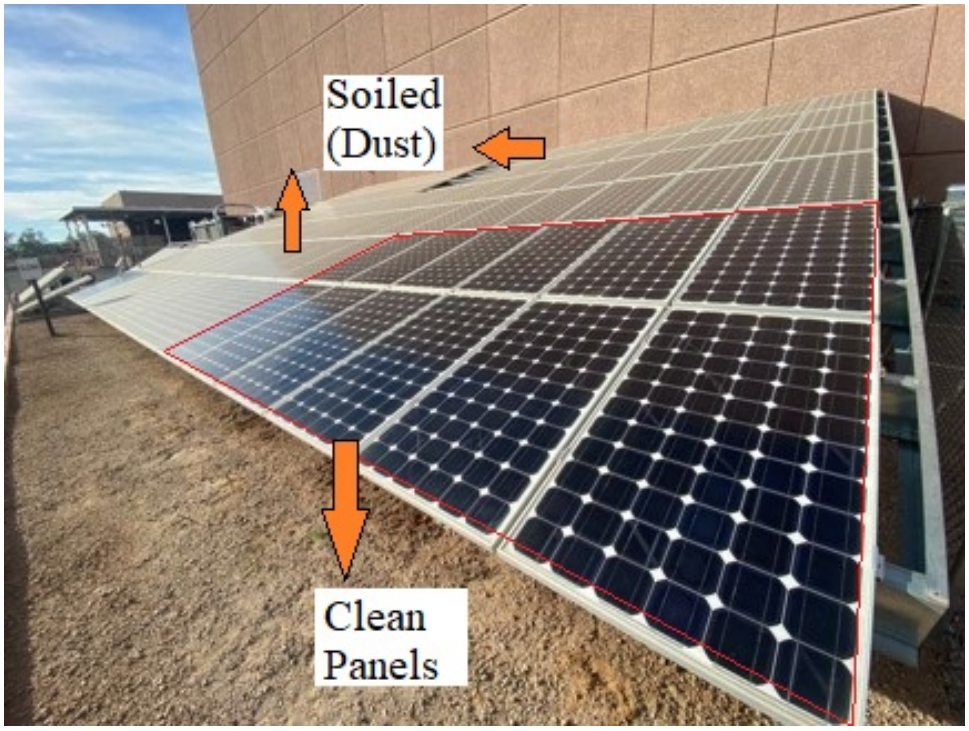

Figure 4.17: An Example of a Soiled Module at ASU Research Park Versus a STC Module.

Degraded Modules:

Degraded modules are a result of modules aging or regular wear and tear of the PV modules. Consequently, such modules produce lower power values owing to the lower values of open-circuit voltage V_{OC} and short-circuit current I_{SC}. For degraded modules, we propose to:

1. We use the clean modules. However, we measure the open circuit voltage V_{OC}. Some of the modules were measured with low V_{OC}. These modules are typically old and are inefficient. However, identifying such modules is critical as they reduce the power output significantly.

2. Continue to record the same measurements such as V_{MP}, I_{MP} and irradiance.

All of these experiments need to have high safety protocols installed. We describe in the Appendix the best practices prescribed by NREL.

4.5.1 The Confusion Matrix

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm. It allows easy identification of confusion between classes, e.g., one class is commonly mislabeled as the other. Most performance measures are computed from the confusion matrix. A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix. The confusion matrix shows the ways in which your classification model is confused when it makes predictions. It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made. An example figure is shown in Figure 4.18.

		Actual Class		
		Positive	Negative	
Predicted Class	Positive	65	38	103
	Negative	74	123	197
		139	161	

Figure 4.18: An Example of a Confusion Matrix. This Shows a Simple Binary Classification Problem of The Predicted Class Vs. The Actual Class.

4.5.2 Feedforward Neural Networks

Using the features mentioned, we apply them as inputs to a multilayer feedforward neural network, popularly called as the multilayer perceptron (MLP). We use a five layered neural network (NN) with backpropagation to optimize the weights used in each layer. Each layer uses six neurons. Information flows through the neural networks in two ways: (i) in forward propagation the MLP model predicts the output for the given data; and (ii) in backpropagation the model adjusts its parameters considering the error in the prediction. The activation function used in each neuron allows the MLP to learn a complex function mapping. The MLP architecture used for Fault Classification is shown in Figure 4.19.

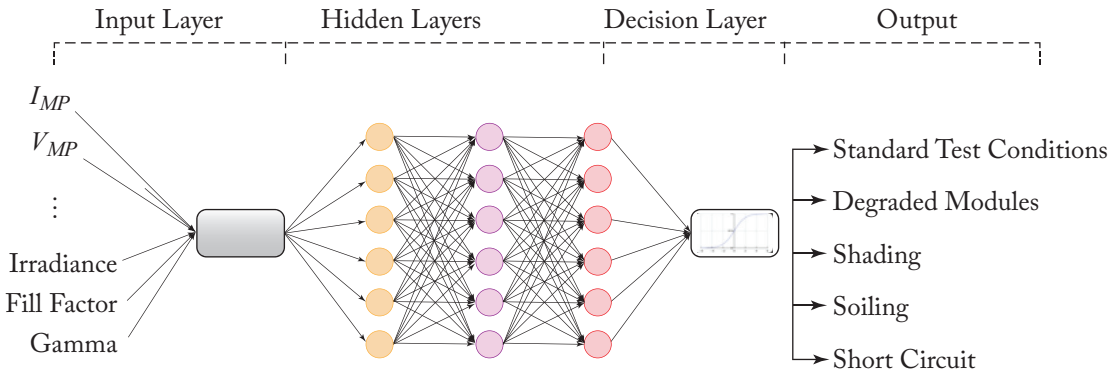

Figure 4.19: Neural Network Architecture Used for Fault Detection and Classification. This NN with Six Neurons in Every Hidden Layer Was Used for Fault Classification on Synthetic Data.

Let $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$ represent the d-dimensional PVWatts data and $\mathcal{Y} = \{\mathbf{y}_i\}_{i=1}^N$ represents one-hot encoded labels for c classes. We consider a NN with L layers. We denote the lth layers weight matrix as $\mathbf{W}_l$ and bias vector as $\mathbf{b}_l$. We use hyperbolic tangent function as the activation function $a(\cdot)$ for the hidden layers and SoftMax function $\sigma(\cdot)$ for the output layer. The output of the lth layer for input $\mathbf{x}_i$ is denoted by $\mathbf{z}_i^{(l)}$. Our goal is to learn a classifier $\mathcal{F}$, such that $\mathcal{F}(x_i, \{\mathbf{W}_k\}_{k=1}^L, \{\mathbf{b}_k\}_{k=1}^L) = y_i$. The update equations of the feedforward NN is given by

$$\mathbf{z}_i^{(1)} = a(\mathbf{W}_1 \mathbf{x}_i + \mathbf{b}_1) \tag{4.4}$$

$$\mathbf{z}_i^{(l)} = a(\mathbf{W}_l \mathbf{z}_i^{(l-1)} + \mathbf{b}_l) \tag{4.5}$$

$$\hat{\mathbf{y}}_i = \sigma(\mathbf{z}_i^{(L)}). \tag{4.6}$$

Weights of each neuron are trained using a scaled gradient back propagation algorithm. Each layer is assigned a tanh (hyperbolic tangent) activation function. From our experiments, we see that the tanh decision boundary gives the best accuracy. The output layer uses the SoftMax activation function to categorize the type of fault in the PV array.

We simulate each fault type vs. shading vs. standard conditions so as to have the same number of data points and avoid bias in the training of the NN. For the training of the NN, we use 70% of labeled data for training, 15% of data for validation, and the remaining 15% data as a test dataset, allowing the algorithm to classify the "unknown" testing data points. The results of the algorithm are shown in the form of a confusion matrix in Figure 4.20.

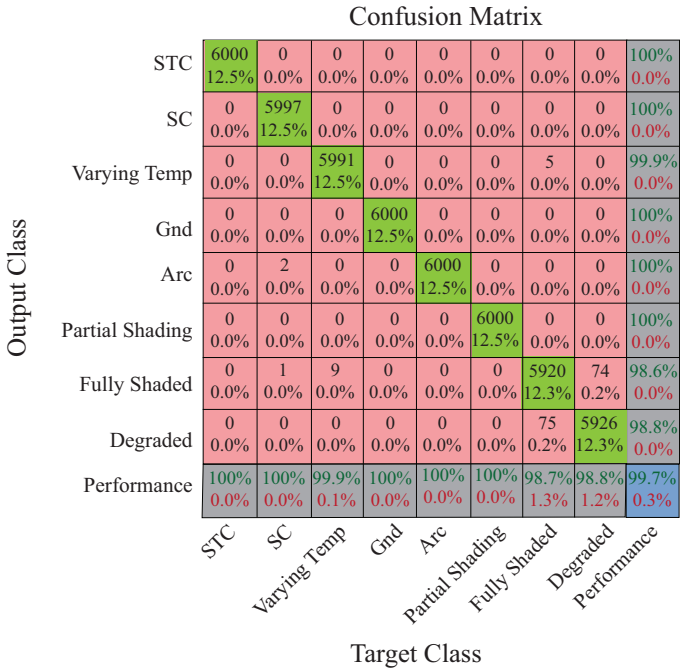

Figure 4.20: Confusion Matrix for Fault Identification. The Results Shown Are on Simulated Data Using the Simulink Model Shown In Figure 2.2. The Simulated Data Is Produced in a Noiseless And Ideal Environment.

However, these results were obtained by Simulink under an ideal noiseless environment and there is a need for a more noisy and realistic scenario. Therefore, in the subsequent sections, we use various NN architectures using the dataset described in Section 2.3.

4.5.3 Pruned Neural Networks

Pruned NN on embedded hardware greatly improve computational performance and reduce memory requirements with a slight reduction in the model's accuracy(Frankle and Carbin (2019). Also called The Lottery Ticket Hypothesis, it is a randomly-initialized, dense neural network contains a subnetwork that is initialized such that when trained in isolation it can match the test accuracy of the original network after training for at most the same number of iterations. Consider a fully connected NN with N neurons in each layer initialized by weight matrices $\mathcal{W}^0 = \{\mathbf{W}_i^0\}_{i=1}^L$. After training this network for t epochs, the resulting weights of the network are $\mathcal{W}^t$. Next, compute a mask $\mathcal{M}$(Frankle and Carbin (2019)) by pruning $p\%$ of the of weights closer to zero by taking the absolute value. Reinitialize the network with $\mathcal{W}^0$ masked by $\mathcal{M}$. The network training and network pruning process is iterated until $2.5\times$ compression is achieved, after which the networks performance degrades due to underfitting of the data(Frankle and Carbin (2019)).

We employ iterative pruning with resetting. The steps are described below:

1. Randomly initialize a neural network f(x;m$\odot$ θ) where $\theta = \theta_0$ and $m = 1^{|\theta|}$ is a mask.

2. Train the network for j iterations, reaching parameters $m \odot \theta_j$.

3. Prune s% of the parameters, creating an updated mask m' where $P_{m'} = (P_m - s)\%$.

4. Reset the weights of the remaining portion of the network to their values in θ_0.

5. Let $m = m'$ and repeat steps 2 through 4 until a sufficiently pruned network has been obtained.

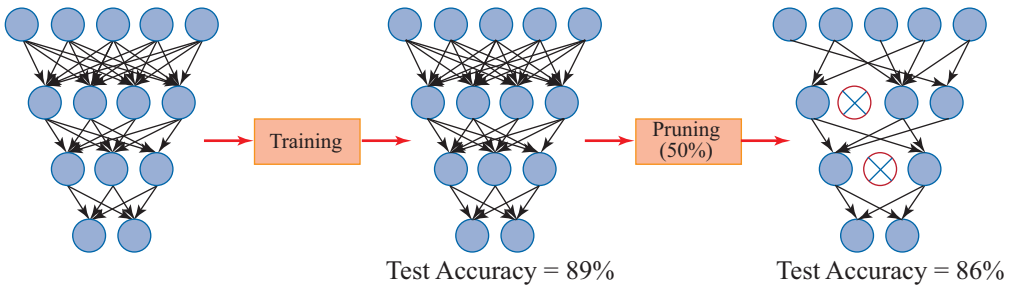

Figure 4.21: A Figure Illustrating the Use of Neural Networks Pruned By 50% for Solar Array Fault Classification.

4.5.4 Dropout Neural Network

In dropout NN, for the lth layer, we select a dropout ratio $p \in (0, 1)$ and sample a vector of Bernoulli random variables $\beta^{(l)}$ with a probability p of being 1 and $1 - p$ of being 0. In both forward pass and back-propagation update, we mask the weights of neurons by computing element-wise product of $\mathbf{z}^{(l)}$ and $\beta^{(l)}$. Masking these weights during the update regularizes the network and avoids over-fitting. Dropout is implemented as, let $\beta_i^{(l)} \sim (p)$ then Equations (4.4), (4.5), and (4.6) are updated as follows:

$$\hat{\mathbf{z}}_i^{(l)} = \beta_i^{(l)} * \mathbf{z}_i^{(l)} \tag{4.7}$$

$$\mathbf{z}_i^{(l+1)} = a(\mathbf{W}_l \hat{\mathbf{z}}_i^{(l)} + \mathbf{b}_l) \tag{4.8}$$

$$\hat{\mathbf{y}}_i = \sigma(\mathbf{z}_i^{(L)}), \tag{4.9}$$

where $*$ denotes element-wise product(Srivastava *et al.* (2014).

4.5.5 Concrete Dropout Neural Networks

Since p is a hyper-parameter, the problem of selecting p for a given dataset is crucial and performing a brute force search on a continuous variable p is computationally expensive. To address this issue, concrete dropout was introduced in(Gal

et al. (2017)), in which the dropout ratio p is optimally selected for each layer by auto-tuning p, i.e., by updating p by gradients with respect to dropout probabilities. Since gradients cannot be computed for the Bernoulli distribution, concrete dropout replaces the Bernoulli distribution during training by a Gumbel–Softmax distribution, so that reparameterization trick can be used to compute gradients with respect to dropout probabilities(Gal *et al.* (2017)).

4.6 Fault Detection and Computational Complexity

We developed a set of nine-dimensional unique custom input feature matrix for the NN. These nine input features are known to provide high accuracy for fault classification on simulated data(Rao *et al.* (2019)). The dataset contains a total of 22,000 samples. We feed the $22{,}000 \times 9$ feature matrix to the NN. We considered a 3-layer neural network with 50 neurons in each layer, as in(Mellit *et al.* (2018)), with tanh as our activation function for each layer. This architecture was fixed for all the NN simulations to avoid any bias which may occur during training and testing. We consider multiple uniform dropout architectures with dropout probabilities $p \in (0.1, 0.2, 0.3, 0.4, 0.5)$, where p is the probability of neurons dropping out in each layer, i.e, in each layer, neurons are dropped randomly based on p.

Along with dropout neural networks, we performed fault classification using the traditional ML classifiers, as reported in Table 4.2 and Table 4.3, and compared the results against those previously reported with fully connected NN (baseline)(Mellit *et al.* (2018); Rao *et al.* (2019)). We run a Monte Carlo simulation on all the architectures mentioned to obtain estimates for training and testing. The training (70%) and testing (30%) dataset were sampled randomly in each run of the Monte Carlo simulation. Among all the dropout architectures we see an improvement of 0.5% when using a concrete dropout architecture in comparison to the fully connected NN.

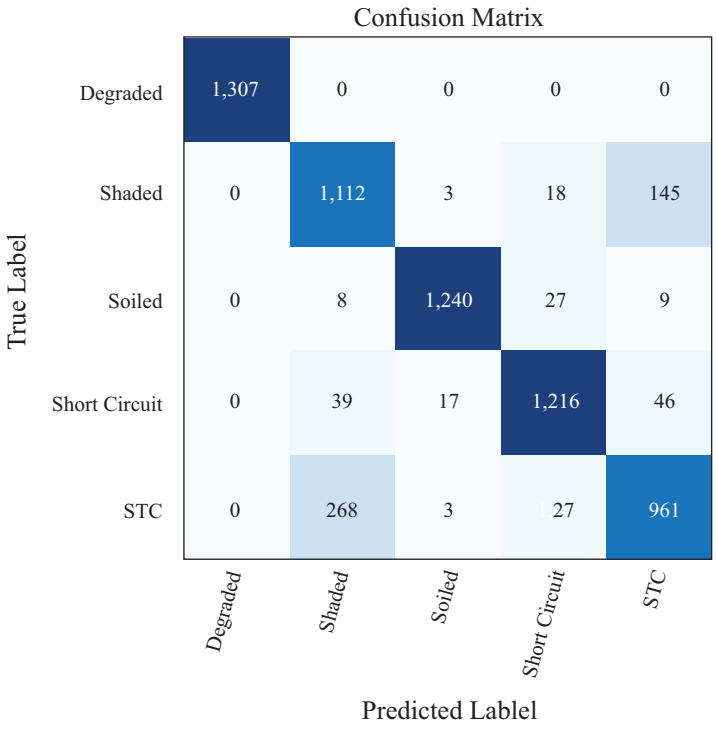

Figure 4.22: Confusion Matrix Obtained with Concrete Dropout. The Dataset Used to Obtain These Results Is Described In 2.3.

We also obtained data in real time from the PV array described in 3. We developed a set of eight-dimensional unique custom input feature matrix for the NN. These eight input features are known to provide high accuracy for fault classification on simulated data(Rao *et al.* (2019)). The dataset contains a total of approximately 8000 samples. We feed the $8{,}000 \times 8$ feature matrix to the NN. The results are shown in Figure 4.23.

We also compared NNs performance with standard machine learning algorithms such as RFC, SVM and KNNs, and the results are reported in Table 4.2 and Table 4.3. For the ML algorithms, we empirically searched over a range of parameters and chose the best configuration. The RFC classifier was trained with 300 estimators with a depth of 50, SVM was trained with radial basis kernel and KNN with 30 nearest neighbors. We observe that techniques such as the RFC overfits the training data,

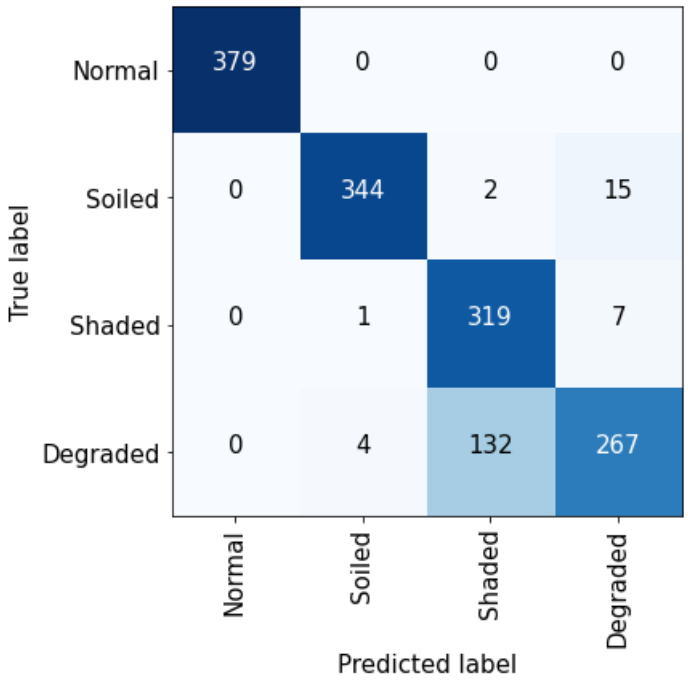

Figure 4.23: Confusion Matrix Obtained with Concrete Dropout. The Dataset Used to Obtain These Results Is Described In Appendix A.

while other classifiers such as SVM and KNN perform poorly compared to NNs. In order to evaluate the model's ability to classify the data points belonging to the group with higher risk factors, we compared performance of different models based on RPN weighted accuracy. The RPN weighted accuracy (RWA) is calculated by summing the products of normalized RPN scores with its class-wise accuracy, written as,

$$RWA = \frac{1}{575}(A_1 + 144A_2 + 30A_3 + 160A_4 + 240A_5)$$

where, A_1, A_2, A_3, A_4, A_5 are class-wise accuracy's of standard test conditions, soiling, shading, degraded and short circuit faults, respectively. The coefficients are obtained using Table 2. We observed that the random forest classifier and concrete dropout has better RWA performance over the other models. Note that, even though RFC has higher RWA score than concrete dropout, the overall test accuracy is much lower

than concrete dropout, which suggests that RFC is accurate only in classifying the faults of higher RPN, whereas, concrete dropout is consistent in correctly classifying all faults classes considered in PV array monitoring systems.

For the ML algorithms, we empirically searched over a range of parameters and chose the best configuration. RFC classifier was trained with 300 estimators with a depth of 50, SVM was trained with radial basis kernel, and kNA with 30 nearest neighbors.

We provide a detailed analysis of the hyperparameter design below: Our hyperparameters were designed from the grid search shown below. In order to choose the hyperparameters we used a grid search. Grid-search is used to determine the optimal hyperparameters of a model which results in the most 'accurate' predictions. For RFC, we search for max depth and estimators, KNN search over neighbors, SVM search over soft margin and kernels. The process is listed below:

- Random Forest Classifier:
  - Max depth: {10,25,50,100}
  - Number of Estimators: {5,10,25,50}
- K- Nearest Neighbor Classifier:
  - Number of Neighbors: {5,10,25,50,100,200}
- Support Vector Machine:
  - C (Soft Margin Parameter): {1,10,100,1000}
  - Kernel: {'linear', 'radial basis function'}

Choice of hyperparameter values: (based on the grid search above) we determine the appropriate of the Hyperparameters obtained by 100 Monte Carlo simulations) These are shown below.

- Random Forest Classifier: By grid search, hyper-parameters associated with the best accuracy of 87.35 are obtained with Max depth of 25 with 50 estimators.
- K- Nearest Neighbor Classifier: By grid search, hyper-parameters associated with the best accuracy of 86.18 are obtained with the number of neighbors being 25.
- Support Vector Machine: By grid search, hyper-parameters associated with the best accuracy of 84.23 are obtained with C=1000 with a 'linear' kernel.

We observe that techniques such as the RFC overfits the training data, while other classifiers such as SVM and KNA perform poorly compared to NNs.

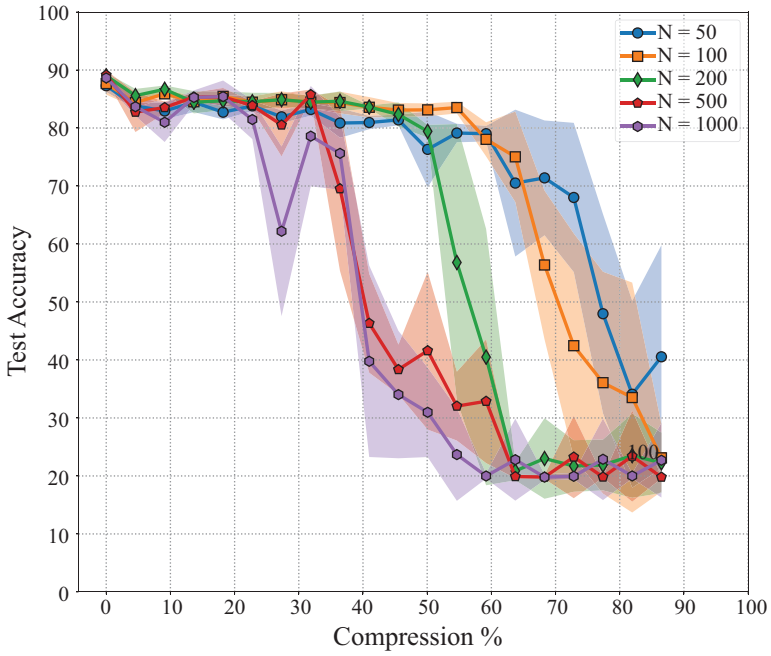

Figure 4.24: Test Accuracy (Mean and Standard Deviation) of Pruned NNs for Different Pruning Compression Percentage for NREL Data. All NNs Have Three Hidden Layers, Each with N Neurons.

Architecture	Train Accuracy(%)	Test Accuracy(%)	Test Accuracy Change	RPN weighted Accuracy
Fully Connected	91.62	89.34	Baseline	85.20
Concrete Dropout	91.45	89.87	+0.5%	85.25
Dropout p=0.1	89.71	89.34	0%	84.53
Dropout p=0.2	89.29	89.13	-0.21%	84.53
Dropout p=0.3	88.92	88.77	-0.57%	84.56
Dropout p=0.4	87.38	88.77	-2.14%	82.39
Dropout p=0.5	85.51	85.42	-3.92%	79.55
RFC	100	86.32	-3.02%	87.57
KNN	87.15	85.76	-3.58%	73.82
SVM	83.51	83.29	-6.05%	79.30

Table 4.2: Comparison of Various Classifiers Used for Fault Classification in PV Arrays. We Note That the Concrete Dropout Architecture Performs Best in Terms of Accuracy Due to an Optimized Hyperparameter Search Within the Architecture.

Architecture	Train Accuracy(%)	Test Accuracy(%)	Test Accuracy Change	RPN weighted Accuracy
Fully Connected	93.4	93.04	Baseline	87.83
Concrete Dropout	92.52	92.23	-0.8%	87.1
Dropout p=0.1	85.1	85.06	-8.17%	85.25
Dropout p=0.2	76.07	76.09	-16.95%	77.07
Dropout p=0.3	71.27	71.24	-21.8%	73.04
Dropout p=0.4	65.15	65.12	-27.92%	68.29
Dropout p=0.5	59.17	59.36	-33.68%	61.32
RFC	73.03	72.51	-20.52%	74.58
KNN	86.77	86.25	-6.78%	87.19
SVM	85.54	85.26	-7.77%	86.77

Table 4.3: Comparison of Various Classifiers Used for Fault Classification in PV Arrays. We Note That the Concrete Dropout Architecture Performs Best is Comparable in Accuracy Due to an Optimized Hyperparameter Search Within the Architecture.

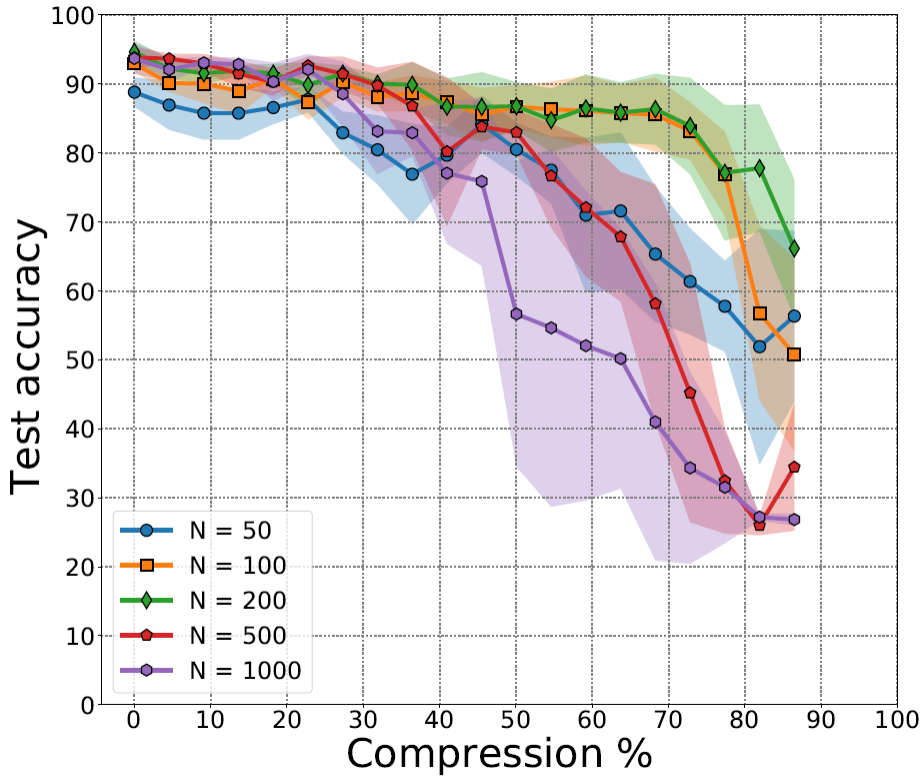

Figure 4.25: Test Accuracy (Mean and Standard Deviation) of Pruned NNs for Different Pruning Compression Percentage for Real Data. All NNs Have Three Hidden Layers, Each with N Neurons.

For the network pruning experiments, we consider NNs with three hidden layers each with $N = \{50, 100, 200, 500, 1000\}$ neurons. All NNs were trained for 150 epochs and at every pruning iteration 10% of the remaining weights were pruned. We find that smaller networks achieve greater compression of about 62% for a drop in accuracy by 4%, as shown in Figure 4.24 and Figure 4.25. The performance of larger networks degrades by up to 40% after pruning the network.

We observe that our pruned neural network algorithms converge faster. This is because there are fewer parameters in our pruned network and hence less misadjustment error. This can be useful for the development of custom hardware for fault classification. We also observe that our pruned neural network algorithms have an accuracy within 2% of the fully connected neural network algorithm for a 40% reduction of the weights of the neural network.

Interestingly, we find that the overlapping points shown in Figure 2.3 correspond the incorrectly classified points in the confusion matrix, shown in Figure 4.22, which is approximately 10% of the data. Hence, accuracy beyond 90% is not achieved by any of these methods Rao *et al.* (2021).

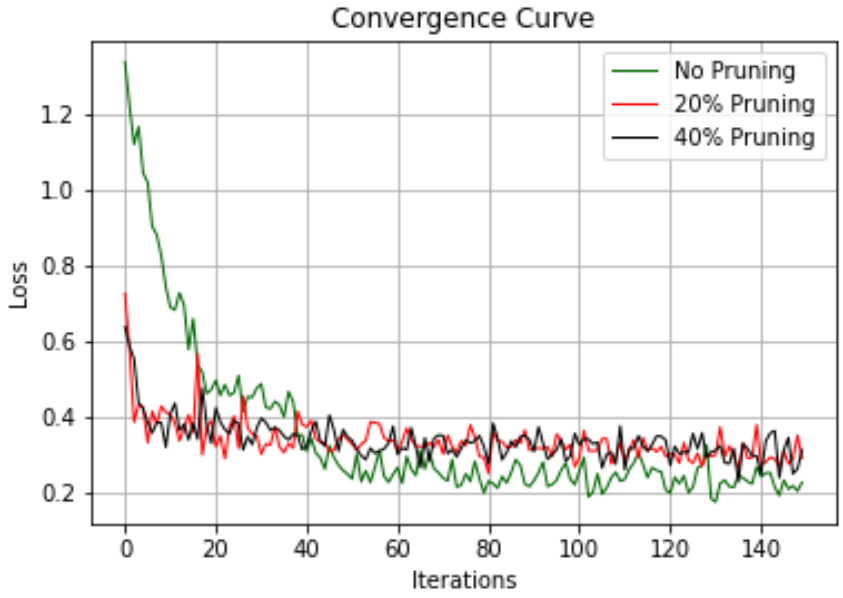

Figure 4.26: The Convergence Plot of the Neural Network with Pruning. We Observe That Pruned Neural Network Algorithms Converge Faster. This Can Be Useful for the Development of Custom Hardware for Fault Classification.

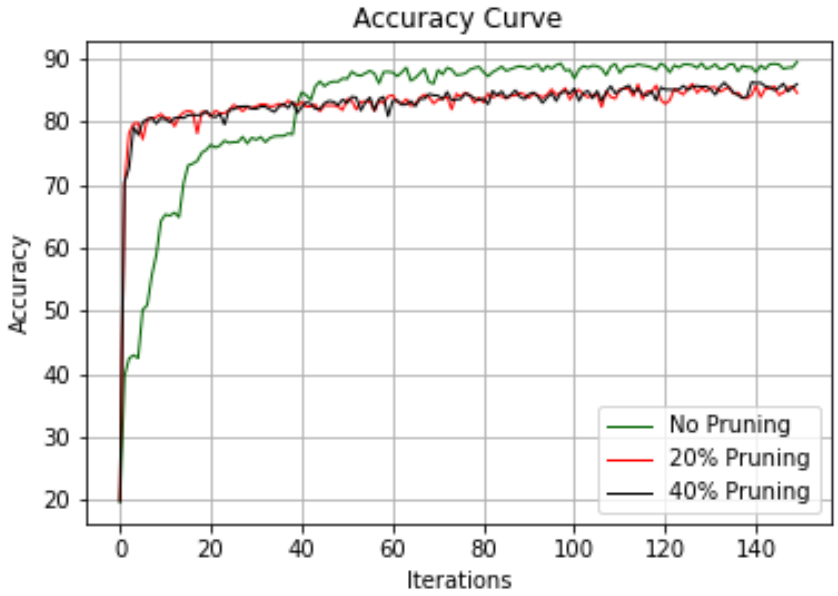

Figure 4.27: The Accuracy Plot of the Neural Network with Pruning. We Observe That Pruned Neural Network Algorithms Have an Accuracy Within 2% of the Fully Connected Neural Network Algorithm for a 40% Reduction of the Weights of the Neural Network.

Chapter 5

CONCLUSIONS

A comprehensive study of fault detection in PV systems including literature review, and machine learning algorithm development was presented in this dissertation. ML and signal processing techniques are utilized to monitor and control PV arrays and develop algorithms to remotely detect and classify the type of fault, thereby enabling fault diagnosis with minimal human involvement. The PV arrays are equipped with SMDs which are capable of switching and controlling panel connections, thereby enabling Maximum Power Point Tracking to optimize and mitigate the effects of module failures and ensure operation of the array at maximum efficiency. In this dissertation, an overview of this system and its design was discussed, followed by ML and signal processing techniques for fault detection and classification.

Chapter 2 described the 18 kW experimental testbed that consists of 104 modules fitted with smart monitoring devices situated at the ASU Research Park. This facility is equipped with SMDs that collect voltage, current, irradiance, and temperature data from individual modules in the array. The data is transmitted wirelessly and is received by a ZigBee hub device connected to a server. This test bed is used to evaluate and validate our algorithms, with real-time data, including ML-based, and graph based techniques for fault detection, and diagnosis.

Chapter 3 addresses the construction of the solar array test bed. We develop a real time load for MPP tracking. We also use this load to collect data for the multiple classes mentioned in this study. We validated the ML methods mentioned in this work using the data obtained from the load using real time measurements.

Chapter 4 addressed the problem of PV array monitoring and control using ad-

vanced NNs and ML algorithms. We describe the formulation of the nine input features used to identify different faults in PV arrays. We collect data in real-time from the ASU SenSIP Solar Array and also use NREL's PVWatts time-series dataset. Results using NNs demonstrated the detection and identification of commonly occurring faults and shading conditions in utility-scale PV arrays. We showed a significant improvement in accuracy of detection and identification of faults compared to traditional and existing methods.

5.1 Summary of Results

In this dissertation, we proposed and characterized efficient neural network architectures for fault detection and classification in utility scale solar arrays using PVWatts time-series dataset as well as real-time data from ASU MTW Research Park. We study the faults and their diagnosis from an operations and management perspective to offer an experimental perspective. We first use an autoencoder to detect faults. We detect faults based on the histogram reconstruction error. We then customize and optimize neural network architectures with concrete dropout mechanisms for fault classification in PV arrays. We examine the fault classification accuracy for each class. We characterize algorithms in terms of performance and complexity and more specifically we compare the proposed concrete dropout method with fixed dropout and fully connected NNs. We also compare our work against standard machine learning algorithms. We observe that concrete dropout outperforms other methods with a classification accuracy of 89.87% and 92.93% as shown in Table 4.2 and Table 4.3 respectively. It also has the fastest run time on the test dataset. In order to reduce complexity, we also explore the use of pruned neural networks. Using Monte Carlo simulations, we demonstrate that the test accuracy of a network pruned by 62% (a significant reduction of weights) reduces only by 4%. The pruned network,

represented by half the number of parameters, will be useful for the development of customized and efficient fault detection hardware and software for PV arrays. In addition, we evaluated faults using their RPN and their corresponding safety category. Some of the faults considered in this dissertation have a high RPN as shown in Table 2.2. We also perform a weighted class average and examine the class wise accuracy of these faults. Since the RPN associated with these faults is high and poses a greater safety threat, the detection and classification of such faults is critical.

5.2 Future Research

5.2.1 Smart Monitoring Device

Currently, we obtain data using Smart Monitoring Devices (SMDs). However, the following concerns could be addressed with respect to SMDs. The SMDs are not designed to be secure. Future research could involve development of custom SMDs which have multiple security protocols. This would ensure that the array continues to operate safely by preventing certain hazardous connections. SMDs should also have the ability to communicate to the server in parallel. In addition, the sampling rate of the SMD could be increased. Currently, the SMD has a sampling rate of $10s$.

5.2.2 Quantized Neural Networks

To improve the throughput and energy efficiency of Deep Neural Networks (DNNs) on customized hardware, lightweight neural networks constrain the weights of DNNs to be a limited combination. In such networks, the multiply-accumulate operation can be replaced with a single shift operation, or two shifts and an add operation (Ding *et al.* (2019)). Future research could look to design a once-for-all network that can be directly deployed under diverse architectural configurations, amortizing the

training cost. The inference is performed by selecting only part of the once-for-all network. It flexibly supports different depths, widths, kernel sizes, and resolutions without retraining. In continuation, research could involve decoupling the model training stage and the neural architecture search stage. In the model training stage, focus should be on improving the accuracy of all sub-networks that are derived by selecting different parts of the once-for-all network. In the model specialization stage, a sample subset of sub-networks to train an accuracy predictor and latency predictors could be used. This is estimated to reduce the total cost of design (Cai *et al.* (2019)).

Taking these assumptions into consideration, a light neural network algorithm could be developed, which in the future can be implemented on the solar module for fast and efficient fault detection and classification.

BIBLIOGRAPHY

Alam, M. K., F. Khan, J. Johnson and A. J. Flicker, "Comprehensive review of catastrophic faults in pv arrays: Types, detection, and mitigation techniques", IEEE Journal of Photovoltaics **5**, 3, 982–997 (2015).

Ali, M. H., A. Rabhi, A. El Hajjaji and G. M. Tina, "Real time fault detection in photovoltaic systems", Energy Procedia **111**, 914–923 (March 2017).

Altman, N. S., *An introduction to kernel and nearest-neighbor nonparametric regression* (The American Statistician, 1992).

An, J. and S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability", Special Lecture on IE **2**, 1, 1–18 (2015).

Bishop, C., *Pattern Recognition and Machine Learning* (Springer-Verlag, New York, 2006).

Braun, H., S. Buddha, V. Krishnan, C. Tepedelenlioglu, A. Spanias, M. Banavar and D. Srinivasan, "Topology reconfiguration for optimization of photovoltaic array output", Sustainable Energy, Grids and Networks (SEGAN) **Vol 6**, pp. 58–69 (Elsevier, 2016).

Braun, H., S. T. Buddha, V. Krishnan, A. Spanias, C. Tepedelenlioglu, T. Takehara, S. Takada, T. Yeider and M. Banavar, "Signal processing for solar array monitoring, fault detection, and optimization", in Synthesis Lectures on Power Electronics, edited by J. Hudgins (Morgan & Claypool Publishers, 3(1, 2012a).

Braun, H., S. T. Buddha, V. Krishnan, A. Spanias, C. Tepedelenlioglu, T. Yeider and T. Takehara, "Signal processing for fault detection in photovoltaic arrays", in 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1681–1684 (IEEE, 2012b).

Cai, H., C. Gan, T. Wang, Z. Zhang and S. Han, "Once-for-all: Train one network and specialize it for efficient deployment", arXiv preprint arXiv:1908.09791 (2019).

Chattopadhyay, S., R. Dubey, V. Kuthanazhi, J. J. John, C. S. Solanki, A. Kottantharayil, B. M. Arora, K. Narasimhan, V. Kuber, J. Vasi *et al.*, "Visual degradation in field-aged crystalline silicon pv modules in india and correlation with electrical degradation", IEEE Journal of photovoltaics **4**, 6, 1470–1476 (2014).

Chen, Z., L. Wu, S. Cheng, P. Lin, Y. Wu and W. Lin, "Intelligent fault diagnosis of photovoltaic arrays based on optimized kernel extreme learning machine and IV characteristics", Applied energy **204**, 912–931 (2017).

Chine, W., A. Mellit, V. Lughi, A. Malek, G. Sulligoi and A. M. Pavan, *novel fault diagnosis technique for photovoltaic systems based on artificial neural networks* (Renewable Energy, 2016).

Cordero, R., A. Damiani, D. Laroze, S. Macdonell, J. Jorquera, E. Sepúlveda, S. Feron, P. Llanillo, F. Labbe, J. Carrasco *et al.*, "Effects of soiling on photovoltaic (pv) modules in the atacama desert", Scientific reports **8**, 1, 1–14 (2018).

Cortes, C. and V. Vapnik, *Support-vector networks* (Machine Learning, 1995).

Dhillon, B. S., *Engineering Maintainability: How to Design for Reliability and Easy Maintenance* (Gulf Professional Publishing, 1999).

Ding, R., Z. Liu, T.-W. Chin, D. Marculescu and R. D. Blanton, "Flightnns: Lightweight quantized deep neural networks for fast and accurate inference", in Proceedings of the 56th Annual Design Automation Conference 2019, pp. 1–6 (2019).

Dirks, E., A. Gole and T. Molinski, *Performance evaluation of a building integrated photovoltaic array using an internet based monitoring system* (IEEE Power Engineering Society General Meeting, 2006).

Dobos, A. P., "Pvwatts version 5 manual", Tech. rep., National Renewable Energy Lab.(NREL), Golden, CO (United States) (2014).

Fan, J., S. Rao, G. Muniraju, C. Tepedelenlioglu and A. Spanias, *Fault classification in photovoltaic arrays using graph signal processing* (ICPS, Tampere, 2020a).

Fan, J., S. Rao, G. Muniraju, C. Tepedelenlioglu and A. Spanias, "Fault classification in photovoltaic arrays using graph signal processing", US patent application number **63** (2020b).

Flicker, J. and J. Johnson, *Analysis of fuses for blind spot ground fault detection in photovoltaic power systems* (Sandia National Laboratories Report, 2013).

Frankle, J. and M. Carbin, *The lottery ticket hypothesis: Finding sparse, trainable neural networks* (ICLR, 2019).

Gal, Y., J. Hron and A. Kendall, *Concrete dropout* (Advances in Neural Information Processing Systems, 2017).

Goodfellow, I., Y. Bengio and A. Courville, *Deep learning* (MIT Press, November 2016).

Hammond, R., D. Srinivasan, A. Harris, K. Whitfield and J. Wohlgemuth, "Effects of soiling on pv module and radiometer performance", Photovoltaic Specialists Conference pp. 1121–1124 (1997).

Hariharan, R., M. Chakkarapani, G. S. Ilango and C. Nagamani, "A method to detect photovoltaic array faults and partial shading in PV systems", IEEE Journal of Photovoltaics **6** (Sept 2016).

Harrou, F., A. Dairi, B. Taghezouit and Y. Sun, "An unsupervised monitoring procedure for detecting anomalies in photovoltaic systems using a one-class support vector machine", Solar Energy **179**, 48–58 (2019).

Ho, T. K., "Random decision forests", in Proc. of the 3rd International Conference on Document Analysis and Recognition, pp. 278–282 (August 14, 1, 1995).

James, M., "Some methods for classification and analysis of multivariate observations", in Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297 (1(14), 1967).

Katoch, S., G. Muniraju, S. Rao, A. Spanias, P. Turaga, C. Tepedelenlioglu, M. Banavar and D. Srinivasan, "Shading prediction, fault detection, and consensus estimation for solar array control", IEEE Industrial Cyber-Physical Systems (ICPS) pp. 217–222 (2018a).

Katoch, S., P. Turaga, A. Spanias and C. Tepedelenlioglu, "Fast non-linear methods for dynamic texture prediction", in 25th IEEE International Conference on Image Processing (ICIP), pp. 2107–2111 (Athens, 2018b).

King, D., *Photovoltaic module and array performance characterization methods for all system operating conditions* (AIP Conference Proceedings, 1997).

King, D. L., W. E. Boyson and J. A. Kratochvill, *Photovoltaic array performance model* (Sandia Report, 2004).

Kolodenny, W., M. Prorok, T. Zdanowicz, N. Pearsall and R. Gottschalg, "Applying modern informatics technologies to monitoring photovoltaic (pv) modules and systems", Photovoltaic Specialists Conference, PVSC'08, 33rd IEEE pp. 1–5 (2008).

Köntges, M., S. Kurtz, C. Packard, U. Jahn, K. A. Berger, K. Kato, T. Friesen, H. Liu, M. Van Iseghem, J. Wohlgemuth *et al.*, "Review of failures of photovoltaic modules", IEA International Energy Agency (2014).

Köntges, M., G. Oreski, U. Jahn, M. Herz, P. Hacke, K.-A. Karl-Anders Weiss, G. Razongles, P. Marco, P. David, T. Tanahashi *et al.*, "Assessment of photovoltaic module failures in the field", IEA International Energy Agency (2017).

Kuitche, J. M., G. Tamizh-Mani and R. Pan, "Failure modes effects and criticality analysis (fmeca) approach to the crystalline silicon photovoltaic module reliability assessment", in "Reliability of Photovoltaic Cells, Modules, Components, and Systems IV", vol. 8112 (International Society for Optics and Photonics, 2011).

Maaten, L. V. D. and G. Hinton, "Visualizing data using t-sne", JMLR **9**, 2579–2605 (2008).

Maish, A. B., C. Atcitty, S. Hester, D. Greenberg, D. Osborn, D. Collier and M. Brine, *Photovoltaic system reliability* (Photovoltaic Specialists Conference, 1997).

Mekki, H., A. Mellit and H. Salhi, "Artificial neural network-based modelling and fault detection of partial shaded photovoltaic modules", Simulation Modeling Practice and Theory **67**, 1–3 (2016).

Mellit, A., G. M. Tina and S. A. Kalogirou, "Fault detection and diagnosis methods for photovoltaic systems: A review", Renewable and Sustainable Energy Reviews **91**, 1–17 (2018).

Muniraju, G., S. Rao, S. Katoch, A. Spanias, C. Tepedelenlioglu, P. Turaga, M. K. Banavar and D. Srinivasan, "A cyber-physical photovoltaic array monitoring and control system", IJMSTR **5**, 3, 33–56 (May 2017).

Nguyen, D. and B. Lehman, "Modeling and simulation of solar PV arrays under changing illumination conditions", Computers in Power Electronics, COMPEL'06, IEEE Workshops on pp. 295–299 (2006).

Patel, H. and V. Agarwal, "MATLAB-based modeling to study the effects of partial shading on pv array characteristics", Energy Conversion, IEEE Transactions on **23**, 1, 302–310 (2008).

Pedersen, E., S. Rao, S. Katoch, K. Jaskie, A. Spanias, C. Tepedelenlioglu and E. Kyriakides, "PV array fault detection using radial basis networks", in 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA), pp. 1–4 (IEEE, 2019).

Peshin, S., D. Ramirez, J. Lee, H. Braun, C. Tepedelenlioglu, A. Spanias, M. Banavar and A. p. p. a. m. s. D. Srinivasan, *34th IASTED International Conference on Modelling, Identification and Control* (Innsbruck, 2015).

Platon, R., J. Martel, N. Woodruff and T. Y. Chau, "Online fault detection in pv systems", IEEE Transactions on Sustainable Energy **6**, 4, 1200–1207 (2015).

Quaschning, V. and R. Hanitsch, "Numerical simulation of current-voltage characteristics of photovoltaic systems with shaded solar cells", Solar Energy **56**, 6 (1996).

Rajput, P., M. Malvoni, N. M. Kumar, O. Sastry and G. Tiwari, "Risk priority number for understanding the severity of photovoltaic failure modes and their impacts on performance degradation", Case Studies in Thermal Engineering **16**, 100563 (2019).

Rao, S., H. Braun, J. Lee, D. Ramirez, D. Srinivasan, J. Frye, S. Koizumi, Y. Morimoto, C. Tepedelenlioglu, E. Kyriakides and A. Spanias, "An 18 kw solar array research facility for fault detection experiments", in 2016 18th Mediterranean Electrotechnical Conference (MELECON), IEEE (Limassol, Cyprus, 2016, 2016).

Rao, S., S. Katoch, V. Narayanaswamy, G. Muniraju, C. Tepedelenlioglu, A. Spanias, P. Turaga, R. Ayyanar and D. Srinivasan, "Machine learning for solar array monitoring, optimization, and control", Synthesis Lectures on Power Electronics **7**, 1, 1–91 (2020a).

Rao, S., S. Katoch, P. Turaga, A. Spanias, C. Tepedelenlioglu, R. Ayyanar, H. Braun, J. Lee, U. Shanthamallu, M. Banavar and A. D. Srinivasan, "cyber-physical sys-tem approach for photovoltaic array monitoring and control", in Proc. IEEE In-ternational Conference on Information, Intelligence, Systems and Applications,(Larnaca, 2017).

Rao, S., G. Muniraju, C. Tepedelenlioglu, D. Srinivasan, G. Tamizhmani and A. Spanias, "Dropout and pruned neural networks for fault classification in photovoltaic arrays", IEEE Access **9**, 120034–120042 (2021).

Rao, S., A. Spanias and C. Tepedelenlioglu, "Machine learning and neural nets for solar array fault detection, patent pre-disclosure", provisional US patent application number **62** (2020b).

Rao, S., A. Spanias and C. Tepedelenlioglu, "Solar array fault detection using neural networks", in 2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS), (Taiwan, May, 2019).

Sepanski, A. and et.al, "Assessing fire risks in photovoltaic systems and developing safety concepts for risk minimization", Report by U.S. Department of Energy, Solar Energy Technologies Office (June, 2018).

Shanthamallu, U., A. Spanias, C. Tepedelenlioglu and A. M. Stanley, "brief survey of machine learning methods and their sensor and IoT applications", in Proc. 8th International Conference on Information, Intelligence, Systems and Applications (IEEE IISA), (Larnaca, 2017).

Shrestha, S. M., J. K. Mallineni, K. R. Yedidi, B. Knisely, S. Tatapudi, J. Kuitche and G. TamizhMani, "Determination of dominant failure modes using FMECA on the field deployed c-Si modules under hot-dry desert climate", IEEE Journal of Photovoltaics **5**, 1, 174–182 (2014).

Spanias, A. S., "Solar energy management as an Internet of Things (IoT) application", in 8th International Conference on Information, Intelligence, Systems & Applications (IISA), (IEEE, Larnaca, Cyprus, 2017).

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting", JMLR **15**, 1, 1929–1958 (2014).

Takehara, T. and S. Takada, "Network topology for monitoring and controlling a solar panel array", US **8264195** (2012).

Takehara, T. and S. Takada, "Photovoltaic panel monitoring apparatus", US Patent 8,410,950 (2013).

Walker, H., "Best practices for operation and maintenance of photovoltaic and energy storage systems", Tech. rep., National Renewable Energy Lab.(NREL), Golden, CO (United States) (2018).

Wiles, J., "Ground-fault protection for pv systems", IAEI pp. 2–7 (2008).

Zhao, Y., R. Ball, J. Mosesian, J.-F. de Palma and B. Lehman, "Graph-based semi-supervised learning for fault detection and classification in solar photovoltaic arrays", IEEE Transactions on Power Electronics **30**, 5, 2848–2858 (2014).

Zhao, Y., J.-F. D. Palma, J. Mosesian, R. Lyons and B. Lehman, "Line—line fault analysis and protection challenges in solar photovoltaic arrays", IEEE Transactions on Industrial Electronics **60**, 9, 3784–3795 (2012a).

Zhao, Y., L. Yang, B. Lehman, J.-F. de Palma, J. Mosesian and R. Lyons, "Decision tree-based fault detection and classification in solar photovoltaic arrays", in 2012 Twenty-Seventh Annual IEEE Applied Power Electronics Conference and Exposition (APEC), (February, 2012b).

APPENDIX A

SOLAR ARRAY FACILITY OPERATION

A.1 Solar Array Description

The increasing demand for green energy requires expansion of renewable sources. Solar arrays on residential roof tops, parking sites, and large commercial structures are being deployed in several countries. In addition, large utility-scale arrays with generation capacity of several megawatts are now connected to the grid. The large number of modules in remote areas makes faults more likely and more difficult and expensive to detect and localize. For this reason, there is a need for automated remote fault detection along diagnostics and mobile analytics. This requires localization techniques, communications and sensor hardware operating along with online algorithms and software at the panel level.

The solar array testbed can perform load switching and data collection in real time. In order to establish safe and correct connections, we first propose to establish the safety protocols recommended by NREL. To support experimental aspects of this research we designed a testing facility at the ASU research park in Tempe, Arizona which is shown in Figure 1.2. This solar array research facility consists of 104 modules in an 6×18 configuration that amounts to approximately 18 kW. Every panel in this solar array is equipped with a smart monitoring device (SMD). These devices are networked and can provide data to servers and control centers. Each SMD not only provides analytics for each panel but contains relays that can be remotely controlled and via wireless access. Relays can bypass or change connectivity configuration, e.g., series to parallel. SMDs, connected to each PV panel, act as intelligent networked sensors providing data that can be used to detect faults, shading, and other problems that cause inefficiencies. Each panel can be monitored individually for voltage, current, and temperature, and all data is wirelessly transmitted to a central hub with minimal power loss. Additionally, each smart hardware device can reconfigure connections with its nearest neighbors. Data collected from the SMDs and reconfiguration testing are used to design and evaluate automated fault detection, diagnosis, and mitigation algorithms. We discuss this in detail below.

To automatically detect faults, this solar array is equipped with smart electronics that provide data for analytics. Smart monitoring devices (SMDs) Takehara and Takada (2013) (FigureA.1.(b)) that have remote monitoring and control capability have been proposed Braun *et al.* (2012a) to provide data from each panel and enable detection and localization of faults and shading. The presence of such SMDs renders the solar array system as a cyber-physical system Spanias (2017) that can be monitored and controlled in real-time with algorithms and software. Figure A.1 shows a cyber-physical 18 kW PV testbed described in Rao *et al.* (2016).

Each SMD includes relays to alternate the topology configuration of the modules within the array. Three modes are available: series, parallel, and bypass. A faulty panel can easily be removed from the system to prevent mismatch losses by using the bypass mode. Figure A.2 shows a schematic of the communication between the SMDs and the server. Each SMD communicates wirelessly to an access point located at one of the PV modules. This access point in turn communicates with a central gateway which is connected to a server through USB.

Each of the SMDs within the array is equipped with ZigBee wireless communication hardware. To minimize power consumption by the SMDs, the ZigBee transceivers do not transmit continuously. Instead they periodically report voltage, current, and

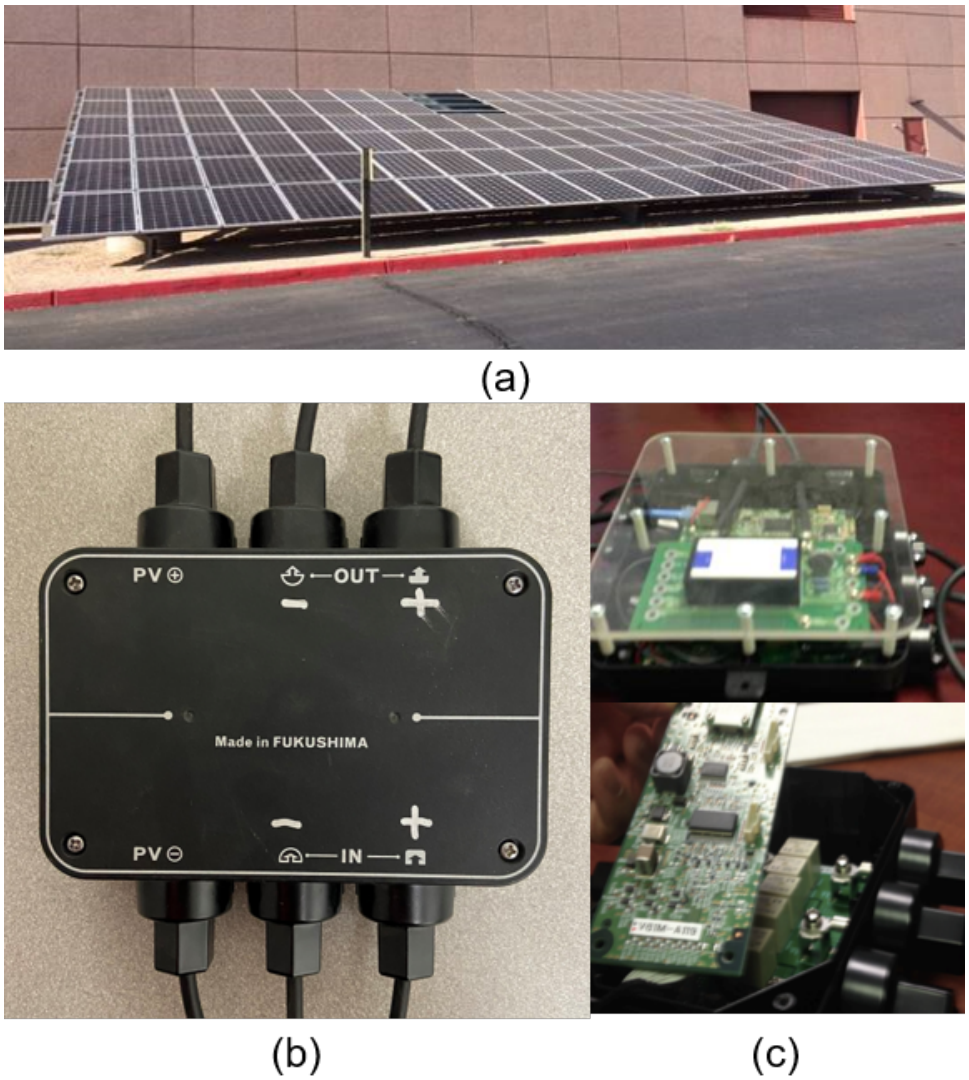

Figure A.1: Smart solar array testbed monitoring system with SMDs at the ASU Research Park. (a) Solar array at the ASU Research Park consisting of 104 modules. (b) SMD which is fitted on to each individual panel. (c) SMD radio and relay switches which allow for real time switching and remote monitoring and control.

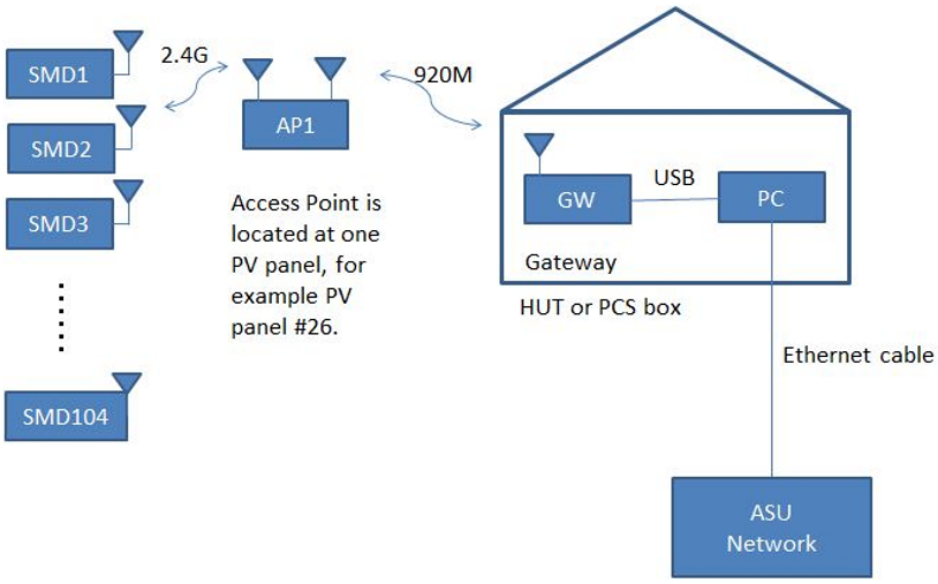

Figure A.2: Block diagram depicts communication between SMDs and server Rao *et al.* (2016).

temperature measurements. A ZigBee hub device connected to the server receives all the reported data and transmits control signals to the networked SMDs.

Figure A.3 shows the computer which was designed for communication with the SMDs and receiving data using the SMDs. An 18kW experimental facility that consists of 104 modules (6×13) fitted with smart monitoring devices has been built at ASU Research Park. The facility is equipped with SMDs that collect voltage, current, irradiance, and temperature data from individual modules in the array. This data is transmitted wirelessly and received by a ZigBee hub device connected to a server. The facility is used to evaluate and validate several different algorithms including novel machine learning based techniques for fault detection and diagnosis.

The solar array testbed can perform load switching and data collection in real time. In order to establish safe and correct connections, we first propose to establish the safety protocols recommended by NREL. We discuss this in detail below.

A.2 Manual Electrical Testing

Manual electrical testing such as open-circuit voltage, operating current, or field I-V curve tracing is used as a method to detect faults in the DC system that the monitoring system is not able to detect. The accuracy of testing equipment is limited by the combined accuracy of irradiance, temperature, and electrical sensors, and in the case of I-V tracing, it is limited to around 5% for standard field units. This testing reveals only defects that are currently causing significant module failure. However, these signatures can be important for understanding underlying module-quality issues, in addition to allowing early detection of possible fire risks. Manual testing is performed over several days or weeks to test a large array. Because this testing must be performed inside the isolated combiner while the system is operational with

Figure A.3: The computer developed for controlling the 18kW solar array. This computer is connected to the transceiver which communicates to the SMDs and receives data.

suitable PPE required for testing. Manual inspection and testing requires that inverter wiring enclosures, re-combiner boxes, combiner boxes, and eventually module junction boxes be opened to access the circuits. These safety recommendations are advised by NREL (Walker (2018)).

A.2.1 Visual Inspection

1. All PV modules are permanently installed (confirm modules are in good condition).
2. All combiner boxes permanently installed.
3. All disconnects and switch gear permanently installed.
4. Wiring is completed (no loose connections or damaged wires).
5. No potential for wire damage (e.g., deburred metal and proper sheathing to protect wires).
6. Utility power connected.

7. Internet connection operational (if applicable).
8. Physical installation is per design drawing and manufacturer's specification, and it meets.
9. System is compliant with applicable building and electrical codes.
10. Protective fencing and enclosures are installed.
11. Verify grounding of metallic surfaces that might become energized.
12. Wire and conduit sizes installed per plan.
13. Fuses and breakers are sized and installed properly.
14. Document as-built conditions.
15. All equipment is labeled as required.

A.2.2 Performance Testing

1. Measure and record open-circuit voltage (V_{OC}) and polarity of each string. (Verifies all strings have the same number of modules.)
2. Measure and record short-circuit current (I_{SC}) of each string.
3. Measure and record maximum power point current (I_{MP}) for each string. (Current measurements for each string should be within a 0.1A range of each other, assuming consistent weather conditions and all string having same tilt and azimuth angle. If a string is outside the range, check for shading or a ground fault.)
4. Confirm the system output under actual conditions meet minimum expected output. Actual performance should be within about 5% of expected, calculated performance. This procedure includes system nameplate rating (kW), solar irradiance measurement (W/m^2) and module cell temperature (C). Procedure is best conducted during consistent weather conditions, where no array shading is present, and solar irradiance is not less than 400 W/m^2.

After having established the safety protocols mentioned, we can begin to obtain data from the array.

A.3 Solar Array Operation Steps

We have established connections to three subarrays each consisting of 12 modules. These subarrays can be connected to the load to establish connections. The layout for each of the subarray is shown in Figure A.4. We describe the steps to connect the array to the load below.

We built a load for a real-time scenario. This is shown in figure A.5. The load is capable of MPP tracking and collect data through the day to validate the results discussed. Data obtained include current and voltage readings from the PV array in

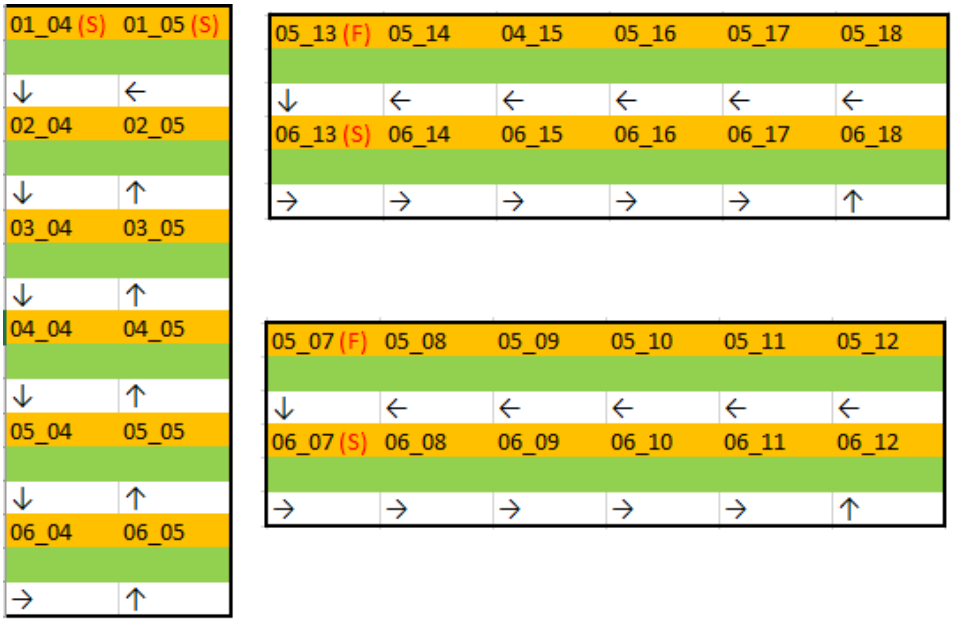

Figure A.4: Three subarrays can be connected to the load. The configurations of the three subarrays are shown here. Each subarray has 12 modules.

real time. In addition, we also obtain irradiance values in real time at a sampling interval of 1s. From the obtained current and voltage readings, we characterise the IV curve of the array and obtain the MPP. These obtained data points are used as inputs in various Machine Learning algorithms to detect and identify faults in PV Arrays. Data obtained helped in identify various loading and shading conditions along with faults as they lie along distinct regions in the two-dimensional space of the IV curve. The load is controlled using the switching box as shown in figure A.6.

1. The load is configured for two configurations, namely 3 series 4 parallel and 12 series and 1 parallel.

2. Put the SMDs in the corresponding mode of operation (series, parallel or bypass). To maintain safe connections, the final SMD should always be in series mode. This will ensure that the PV+ terminal is always connected to IN+ terminal of the SMD and the PV- terminal is always connected to the OUT+ terminal of the SMD.

3. To operate the SMD in series, parallel, bypass series and bypass parallel modes, we need to used the commands 0602, 0603, 0600 and 0601 respectively. This can be performed through custom software or MATLAB.

4. The message ID should be set to 11 to allow for duplex communication. Using the custom software or Matlab, we can issue these commands to each SMD using its MAC address.

Figure A.5: A load bank consisting of multiple resistors to collect data in real time. There are 7 resistors. 4 of these resistors are always on and the remaining 3 are turned on depending on the time of the day.

5. We first need to verify the open circuit voltage depending on the configuration chosen. For the $3S4P$ combination V_{OC} should be close to $120V$ and for the $12S1P$, V_{OC} should be approximately $480V$.

6. After having established V_{OC}, the load can be turned on using the monitor of the control box shown in Figure A.6. All the resistors must be turned on initially. This corresponds to $4R$ on the monitor of the control box.

7. Verify the current flowing through the PV wire using a clamp ammeter. This current should not exceed the value of the fuse inserted in the switching box shown in Figure 3.5.

8. After having verified that the current is safe, we can turn off the resistors one at a time, depending on the time of the day. To do this, we need to use the $3R$, $2R$, R buttons on the touch screen of the control box.

9. To obtain data in real time from the SMD, we need to use the command 03. This will provide us with measurements for each individual panel. Measurements include voltage and current.

10. Use the irradiance meter to record irradaince measurements. The meter has a sampling rate of 1 second per reading. Figure A.7 shows the irradiance meter used. Figure A.8 shows the software which helps obtain real-time irradiance

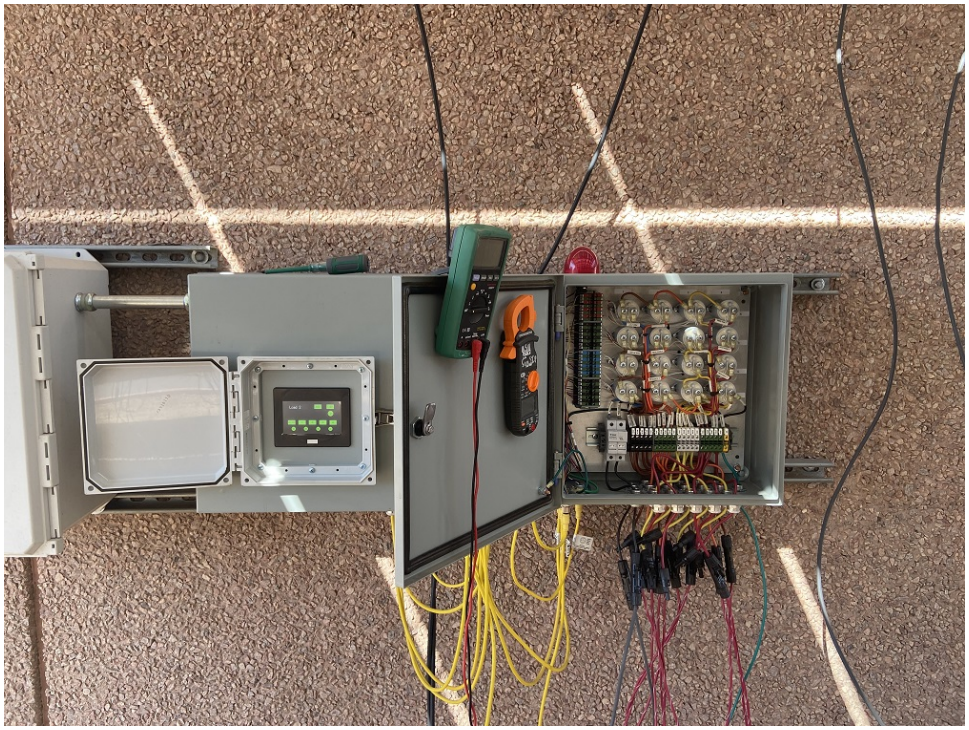

Figure A.6: The control box connects to the load shown in A.5. The control box has relay switches and a PLC to allow for varying loads during the day.

measurements. The values obtained from the meter can be saved locally into a server.

11. Use the recorded measurements to build the feature matrix shown in Table 4.1.
12. The feature matrix is used to train ML algorithms.

Procedure to establish connections

1. Mount the resistors on the cage built.
2. Establish safe connections between terminal box and the resistors. Keep the switches off and the SMDs in bypass mode to prevent arcing.
3. Establish safe connections between modules and the switching box. Turn off the DC contactors and again keep the SMDs in bypass modes to prevent arcing and short circuits.
4. Once connections are established, turn on all resistors for maximum safety to prevent safety hazards while keeping the DC contactors off.

Figure A.7: The irradiance meter used to obtain irradiance measurements in real time. The irradiance meter has a sampling rate of $1s$ and it is placed on top of the PV module.

Turning on the array

1. Keeping all resistors on, set the SMDs to parallel or series mode depending on the configuration needed.

2. Verify the open circuit voltage of the array after connecting all the modules in series or parallel. The open circuit voltage should be a sum of all voltages if the modules are series and should be divided among all modules if they are in parallel.

3. Turn on the DC contactors and establish a closed circuit. This can be established using the monitor on the control box. A programmable PLC controls the DC relays which enable for safe operations of the load.

4. Vary the resistors depending on the time of the day. This is to ensure that we perform maximum power point tracking. By varying the load through the day, we ensure the array is operating under design conditions.

A.4 Experiments

For normal operations

Standard Test Conditions (STCs) are the industry standard for the conditions under which a solar panel are tested. By using a fixed set of conditions, all solar modules can

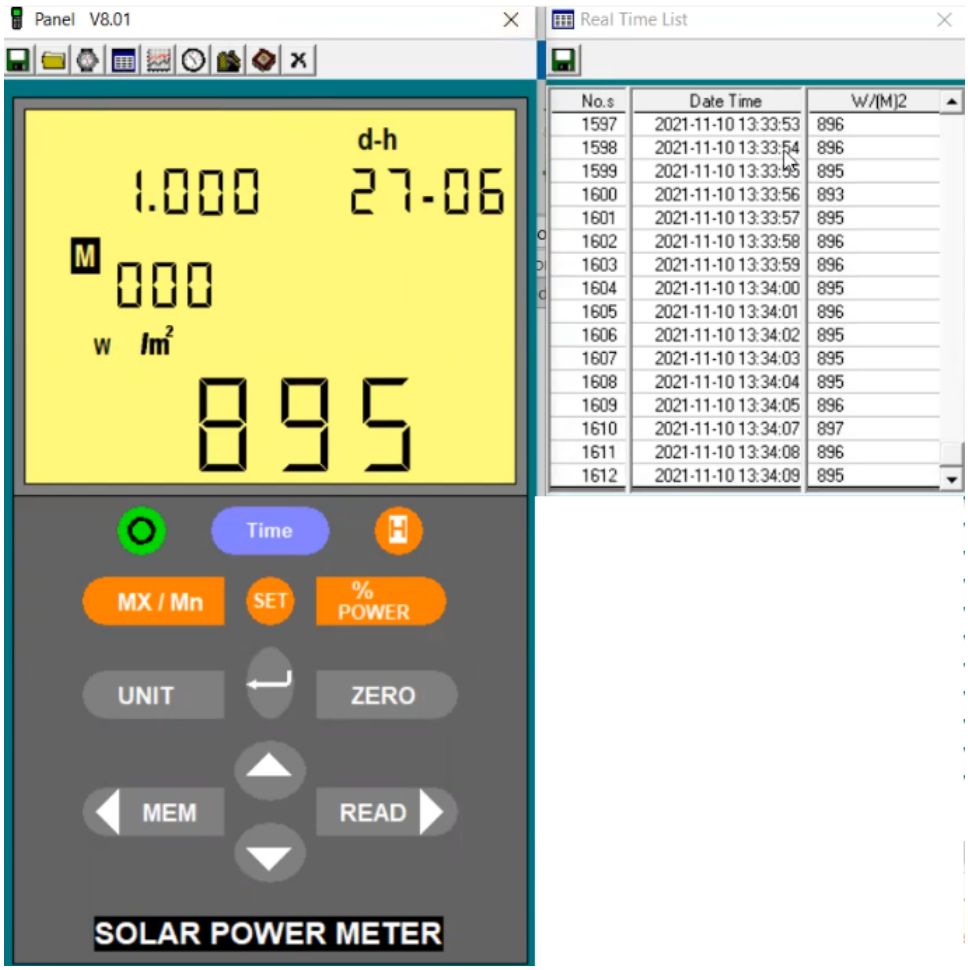

Figure A.8: The irradiance meter can be connected to the computer using a USB port. The meter readings can be saved locally into a server.

be more accurately compared and rated against each other. STC values correspond to the measurements yielding maximum power under the temperature and irradiance values of a particular day. Data points are labeled as STC if the irradiance, temperature, and power were the highest possible values for that particular day. For normal operations, we propose to:

1. Keep the array on, vary the resistors depending on time of the day.
2. Record measurements- V_{MP}, I_{MP}, temperature. These measurements are recorded by the Smart Monitoring Device (SMD) in the array developed at ASU Research Park.
3. Record irradiance measurements using the TES132 meter. The meter gives values at a sampling rate of 1 second. These readings are beneficial in identifying soiling versus shading modules.

Shading

Shading is a serious concern in PV arrays. A module is shaded if the irradiance measured is considerably lower than STC, usually caused by overcast conditions, cloud cover, and building obstruction. As a result, the power produced by the PV array is significantly reduced. To run shading experiments, we propose to:

1. Run a piece of obstruction over a few modules. Scenarios include partial shading and complete shading. 25% of the modules are covered during this process. An irradiance value drop of 25% or more shows significant power loss under shading conditions.

2. Measurements will include V_{MP}, I_{MP} and temperature. Our experiments on the PVWatts dataset have shown that neural networks are effective in classifying shaded modules with high accuracy. Figure A.9 shows a PV panel shaded at the ASU Research Park.

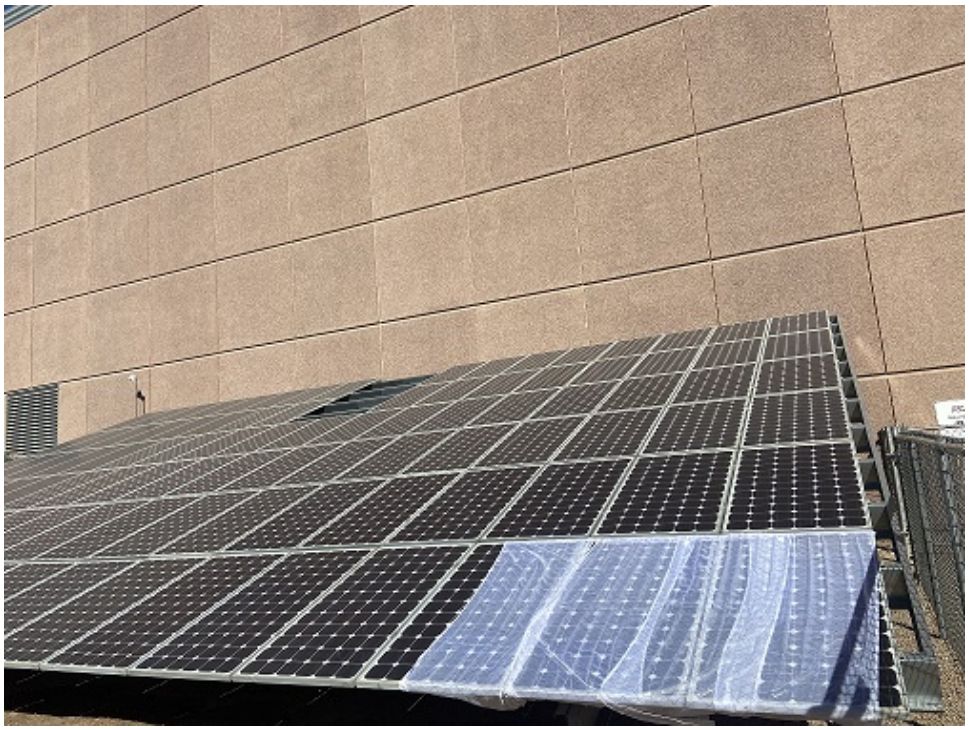

Figure A.9: An Example of a Simulated Shaded Module at ASU Research Park. This Corresponds to 25% Shading.

However, significantly (100%) covering the module could turn off the array and not record any values. Therefore, we propose to cover not more than 50% of the module at a time.

Soiling

While the irradiance measured remains the same as STC, the power produced drops significantly. The solution to this problem involves manually cleaning the modules regularly. If the measured irradiance was as per STC but the power measured was low, then the module was soiled. Soiling is caused by dry deposition affects the power output of PV modules, especially under dry and arid conditions that favor natural atmospheric aerosols (wind-blown dust). For soiling experiments, we propose to:

1. We use the uncleaned modules for these experiments. These modules have dust particles on them which are a result of frequent dust storms.

2. Record the same measurements as provided by the SMD which include V_{MP}, I_{MP} and irradiance. An illustration of soiling versus STC modules is shown in Figure A.10.

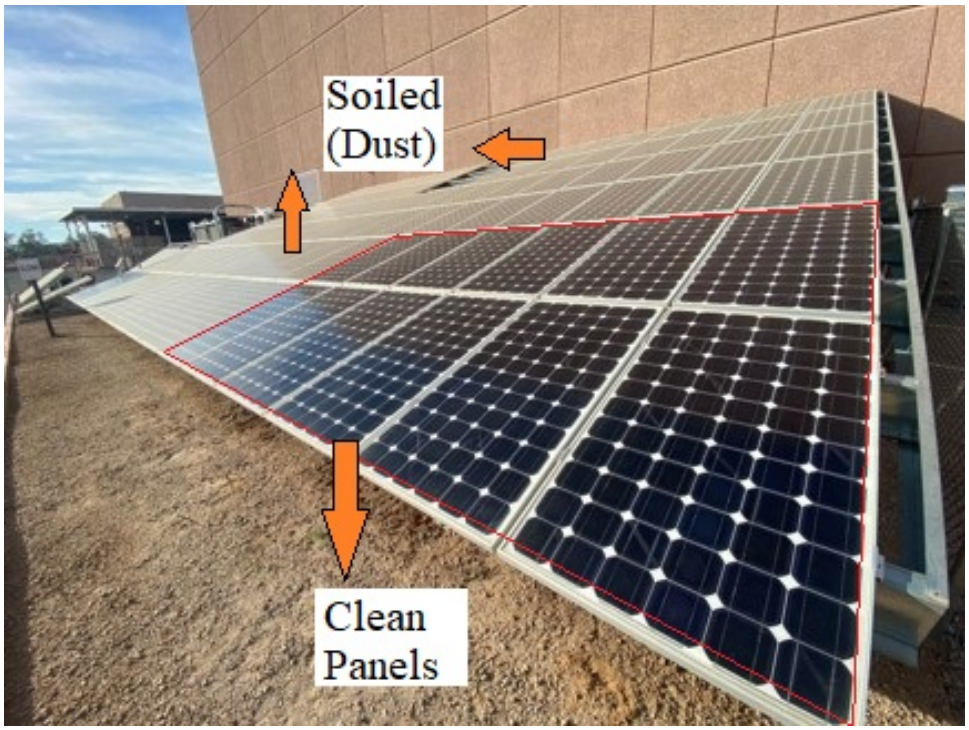

Figure A.10: An Example of a Soiled Module at ASU Research Park Versus a STC Module.

Degraded Modules

Degraded modules are a result of modules aging or regular wear and tear of the PV modules. Consequently, such modules produce lower power values owing to the lower values of open-circuit voltage V_{OC} and short-circuit current I_{SC}. For degraded modules, we propose to:

1. We use the clean modules. However, we measure the open circuit voltage V_{OC}. Some of the modules were measured with low V_{OC}. These modules are typically old and are inefficient. However, identifying such modules is critical as they reduce the power output significantly.

2. Continue to record the same measurements such as V_{MP}, I_{MP} and irradiance.

All of these experiments need to have high safety protocols installed. We describe in the Appendix the best practices prescribed by NREL.

A.4.1 Safety Considerations

We use PVWatts data to develop the design parameters of the load. However, we face multiple fire and safety hazards. In order to avoid fire hazards, we have considered a safety factor of 1.5×. The load has been designed to hold more current than the solar array is equipped to supply. For shading and soiling experiments, the power values measured are expected to be lower than STC. Because of the inherent inaccuracies of data monitoring alone, it is necessary to implement a secondary check of the DC array to detect string- and module-level faults through periodic inspection and testing. The two main methodologies used for these inspections are manual electrical testing and aerial thermal-imaging inspections. We will focus on manual electrical testing. However, we describe all the safety considerations in detail in the Appendix.

BIOGRAPHICAL SKETCH

Sunil Rao received a B.E. degree in electronics and communications engineering from Visvesvaraya Technological University, India, in 2013, and an M.S. degree in electrical engineering from Arizona State University, Tempe, AZ, USA, in 2018. He is currently pursuing a Ph.D. degree at the School of Electrical, Computer, and Energy Engineering, Arizona State University. He is a member of the SenSIP center. He also taught the lab section of the DSP course for four years. Most recently, in the Summer of 2020, he was an intern at Bosch. His research interests include solar array fault classification using machine learning, signal processing, and deep learning.