Utilizing AMI Interval Data and Machine Learning Algorithms to Identify Distribution System Topology and DER Connectivity

by

Elizabeth M. Cook

B.S. in Electrical Engineering, University of Pittsburgh, 2004

M.S. in Electrical Engineering, Kansas State University, 2011

Submitted to the Graduate Faculty of

the Swanson School of Engineering in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2021

UNIVERSITY OF PITTSBURGH

SWANSON SCHOOL OF ENGINEERING

This dissertation was presented

by

Elizabeth M. Cook

It was defended on

October 6th, 2020

and approved by

Brandon Grainger, Ph.D., Assistant Professor, Dept. of Electrical and Computer Engineering

Yang Weng, Ph.D., Assistant Professor, Arizona State University, Dept. of Electrical and Computer Engineering

Katrina Kelly-Pitou, Ph.D., Urban Systems Strategy, Smith Group

Murat Akcakaya, Ph.D., Assistant Professor, Dept. of Electrical and Computer Engineering

Ahmed Hassan Dallal, Ph.D., Assistant Professor, Dept. of Electrical and Computer Engineering

Masoud Barati, Ph.D., Assistant Professor, Dept. of Electrical and Computer Engineering

Utilizing AMI Interval Data and Machine Learning Algorithms to Identify Distribution System Topology and DER Connectivity

Elizabeth M. Cook, PhD

University of Pittsburgh, 2021

The ongoing deployment of Distributed Energy Resources (DERs), while bringing benefits, introduces significant challenges to the electric utility industry, especially in the distribution grid. These challenges call for closer monitoring through state estimation, where real-time topology recovery is the basis for accurate modeling. With the dramatic increase of the residential photovoltaic (PV) systems (i.e., DER), utilities need to know the locations of these new assets to manage the unconventional two-way power flow for sustainable management of distribution grids. Previous methods to maintain the system connectivity are either based on outdated maps or an ideal assumption of an isolated sub-network for topology recovery, e.g., within one transformer. This requires field engineers to identify the association, which is costly and may contain errors. As it has been shown that, historical records are not always up-to-date.

To solve these problems, a density-based clustering method is proposed that leverage both voltage domain data from the Advanced Measurement Infrastructure (AMI) and the geographical space information. The goal of such a method is to efficiently segment data sets from a large utility customer pool, after which other topology reconstruction methods can carry over. Specifically, it is shown how to use the voltage data and GIS information to refine the connectivity within one transformer. To give a guarantee, a theoretic bound for the proposed clustering method is shown, providing the ability to explain the performance of the machine learning method. Numerical results on both IEEE test systems and utility networks show the outstanding performance of the new method. An implementation is also demonstrated in the field.

In this dissertation, we consider the rich potential of large utility datasets, in which physical laws are inherently embedded, to identify system information and utilization by using machine learning algorithms. In order to provide situational awareness and tackle

practical issues such as limited measurements and un-scalability, we start with proposing a customized data-driven approach to provide an accurate model for distribution grid control and planning.

Table of Contents

List of Tables

List of Figures

Preface

I would like to take a moment to express my gratitude to each and everyone that has been a part of my journey towards achieving this milestone. First and foremost, I want to thank my husband, my rock, my love, my best friend, Jason, for his devoted support and ability to stay the course and allow me to pursue my dream and hold down the fort while I did so. You are amazing. Thank you.

I also want to thank my six beautiful children, Alexandra, Jameson, George, Maggie, Aiden, and Finley. The six of you have only known me as a full-time working-mom while also attending school for literally your whole lives thus far. Thank you for being your bright and brilliant selves giving your mom the clarity and ability to focus as you all are such beautiful, steady, and loving individuals that makes being your Mom easy.

I thank my parents, Richard and Victoria Graham who instilled in me a growth mindset and provided the confidence and wisdom to know I can achieve anything I put my mind to and always saw me which instilled a strong sense of self. My Dad was the man who has always been my wisdom teacher and continues to guide me through this journey we call life. And my Mom, the woman who always made the small moments magical and continues to show me how to capture those moments each day. Thank you.

After working as an electrical engineer for over a decade while simultaneously starting a family with young children, I began on a quest to discover what it was about my work and my family life that made me quite literally want to skip across the parking lot to go to work. I realized it came down to two things. 1) A career and company that allowed me to feel seen and heard. And 2) centering my actions from a place of service in both my professional and personal life. I want to acknowledge and thank my first boss, John Paserba, and second boss, Donald Shoup, who introduced me to the role of a power system consultant. You inspired me never to stop asking questions and seeking the answer as well as believing in my ability to contribute and consult in the power industry as well as to continue my education and earn my doctorate. Whoever thought my journey could be so exciting, it all started when I was taught to analyze every power infrastructure as you ride down the highway and

memorize the equipment to understand what it did and how it fits into serving the larger good. I also want to acknowledge two amazing teammates and friends that believed in me and supported me throughout my professional and academic career, Lucas Collette and the late Robert Hellested.

I would like to thank Dr. Katrina Kelly-Pitou for her continued support and spirit to keep me positive and provide the inspirational thought and gusto in believing the work that we are doing will have an impact and we can be the change we want to see in the world.

And from the bottom of my heart I would like to say a big thank you for all the support, energy, guidance, and patience, given to me by Dr. Yang Weng, Bilal Saleem, and Shuman Luo. It was serendipity that our paths met on this journey and I am grateful each day for it. It truly has been a joy to work and learn from you.

Finally, a special thank you to Dr. Brandon Grainger; without your help and wise guidance, this achievement would not have been possible. You never did give up on me, and you will always have a forever friend as we continue to work together on the largest machine on earth, the power grid.

Elizabeth Cook

1.0 Introduction

The power grid has been built and designed over the last 120 years standardizing on the idea of one-way power flow from the transmission system to the distribution system [34] as shown in Fig. 1.

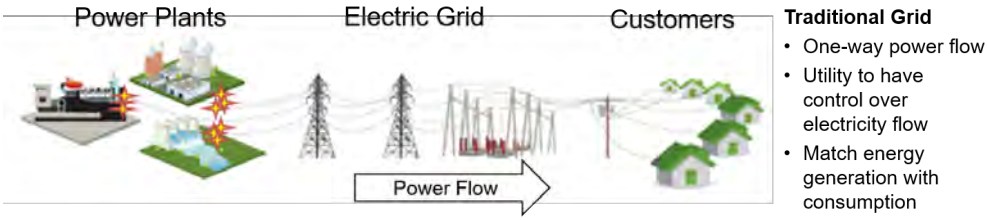


Figure 1: Illustration of the Current Power Grid.

Recently, there is a rapidly expanding usage of distributed energy resources (DERs) in the distribution grids, which is proliferating emerging technologies that ultimately facilitate renewable energy (i.e., photovoltaic and storage devices). However, with the continuous growth of DER penetration, the one-way power flow is starting to change direction from the distribution system to the transmission system, leading to a two-way power flow [21] as shown in Fig. 2.

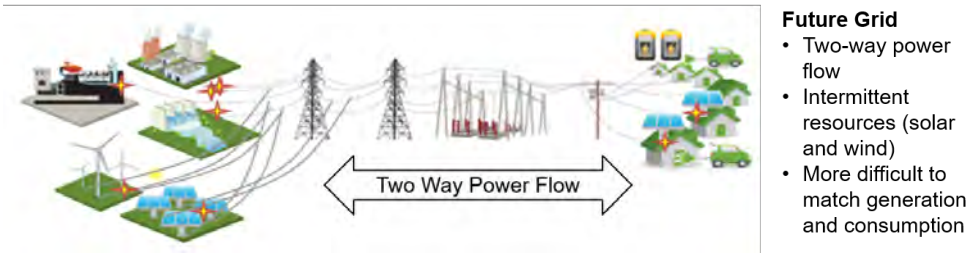


Figure 2: Illustration of the Future Power Grid with the Integration of Utility-scaled Renewable Generation and DER Interconnections.

Therefore, Electric Distribution Companies (EDCs) need to have visibility on their distribution assets to avoid potential risks, e.g., outages and equipment damages, caused by uncertain two-way power flows. For example, EDCs can use the topology information to run analyses and troubleshoot the power grid's real-time operation as well as optimizing planned work. However, a major challenge for EDCs to gain visibility today is on how to develop system-wide models of their distribution systems, where an accurate topology is the basis for improved system-observability and controllability [41,62]. Such modeling challenges exist because a large portion of the infrastructure pre-dates modern communication methods. As a result, there is a significant need for topology identification processes within the EDCs. For topology observability, one idea is to replicate the approaches on the transmission grid. For example, our partner EDC mainly used electronic document repositories and human effort to construct the models on transmission grid. Such a model is subsequently maintained and updated by a human interface based on field measurements to validate or re-estimate the topology, e.g., chi-square Test for topology error process. These approaches have worked well for the transmission grid due to fully constructed monitoring systems and relatively infrequent topology change [12, 24, 31, 59, 66]. For instance, the majority of transmission changes are planned for weeks, months, or even years, allowing the topology model to be updated with high accuracy [65]. Unfortunately, the sensor types and numbers are much fewer in the distribution grid with regular and "unexpected" topology changes [6,58,61], due to routine but unreported reconfiguration [6,18,39,60], e.g., deliberately changing the radial topology to a meshed distribution topology in city networks [5,18]. The dramatic increase of residential DER systems in recent years has accelerated the modernization of the power grid. With the introduction of DER, the distribution grid is becoming less predictable with the addition of the intermittent generation, undergoing multiple re-configurations and upgrades throughout almost each day of operation for many utilities [2,3,64]. When considering "grid modernization" the term applies not only to the physical, but also, digital infrastructures that allow for 2-way power distribution systems. As such, there is a deep need to better understand the modernization of planning and forecasting tools so that decision-making can become as flexible as distributed resources themselves. Fortunately, with the deployment of Advanced Meter Infrastructure (AMI), utilities can employ big data to develop topology

recovery processes with the little manual effort. AMI data and their GPS coordinates provide additional information to accurately recover the topology of the grid, e.g., connections among generation devices (e.g., residential solar panels), load devices (e.g., electric vehicle charging station), and devices with both generation and load capability (e.g., battery). Utilities have to alter their traditional ways of operating and planning to maintain the reliability and safety of the grid. To accomplish the transformation, utilities have to detect and monitor all the DER installations in their territory. One challenge flies in the face is that sometimes the existence of DER systems do not align with the utilities' records, as such, we can focus on future predictive tools to better determine the impact of intermittent renewables on grid reliability. Therefore, there has been significant work done to develop methods to use the AMI and micro-synchrophasor data to recover the system topology leveraging voltage correlations [6, 58, 61]. In this dissertation, we consider the power of a large amount of data, where physical laws are inherently embedded. System topology and accuracy is of the utmost importance in regards to start understanding the implications of DER. If the system is not identified and modeled correctly, then a large amount of data cannot be leveraged to analyze past, current, and future models to make accurate decisions about future infrastructure needs by minimizing cost and maximizing DER benefits. The overall goal will be to understand their impact on the grid in short-term and long-term planning cycles.

1.1 Objective

The key objective of this research is to design data-driven machine learning algorithms to identify system topology and DER interconnections to provide utilities with secure, fast, and scalable processing tools for real-time data.

The first area of study is to develop a physically meaningful, clustering method to roughly separate the data. Second, we use both the voltages and street information to characterize the data for algorithmic solutions to refine the solution.

The second area of study is to introduce a proposed method to detect customers with solar or without solar. We explain why solar detection is urgently needed and why the problem is hard and costly in reality based on our data mining of realistic utility data. Model the solar detection problem in supervised learning, semi-supervised learning (SSL), and one-class classification (OCC) setups. So, future researchers can develop relevant tools based on our problem modeling. And then proposes new SSL and OCC methods based on autoencoders, greatly boosting the power data representation and learning.

The last area of study will be to perform system planning studies using the current utility data available and then the additional extracted data from the AMI meters and compare the results from each to provide insight on the key attributes that are important to ensure the grid is being planned with minimized cost and maximize DER assets.

1.2 Dissertation Organization

This dissertation is organized so that any content can be separated into three primary categories: literature review, proposed work, and summary of completed work utilized in a power system analysis.

Section 2 addresses the background literature review, summarizing the differences between transmission and distribution systems and comparing the availability of a utility distribution system to the transmission system and how a new framework is emerging among utilities seeking to plan for DER integration proactively.

Section 3 covers the work in regards to developing a utility data-driven (AMI voltage data, utility assets GPS, and publicly available GPS) machine-learning algorithm to identify system topology rapidly and effectively on a distribution system.

Section 4 covers the work in regards to developing a utility data-driven (e.g., AMI kWh usage data) machine-learning algorithm to analyze solar or non-solar customer usage data. The algorithm identifies DER interconnections (i.e., solar). This will aid in the implementation of a proactive DER planning process.

Section 5 provides the results of a conducted hosting capacity analysis (HCA) by using the deliverable from the two previous initiatives — the first analysis using the existing data available to the utility planner. The second analysis will use the recovered data from the AMI meters. Then the two sets of results will be compared by analyzing the results to determine impact of the optimized location of the DER in regards to cost and potentially maximize DER benefits.

Section 6 provides an overall conclusion.

2.0 Background

As DER increase, electric utility system planning processes must transition from a discrete process to a probabilistic process where multiple time variable analysis is performed to ensure continued reliability under unknown circumstances. The distribution system currently experiences higher exposure to outages and switching configurations, which makes the requirement of having an accurate system model available for planning and operational purposes. The variability in electrical output from DER impacts the utilities' ability to forecast and respond to real-time overloads and regulate voltage within limits. The tasks mentioned in this research could be utilized to guide the development of a roadmap for developers/aggregators, EDCs, TOs, RTO/ISO, and others on what the roles and responsibilities should be and how to share data of DER to ensure accurate system planning forecasts are developed and utilized. Early planning efforts will enable utilities to minimize the risks and realize the benefits of a distributed energy future. With the permission of a utility in the Mid-Atlantic, the research will be able to utilize real utility data to perform system analyses to provide a public awareness of what the challenges are that a utility faces in regards to planning and implementing infrastructure and advanced technologies while managing the unknown forecasts of future DER within a utility's business model. The utility provides electric service to more than 5590,000 customers.

This challenge is representative of many utilities in the U.S., but also in broader markets where generation and T&D have been divided. At the same time, when considering the need for more resilient power infrastructure and being associated with diversified energy resources, it is now vital for utilities to understand how they can best manage these new resources that are put onto their systems. When considering the pressure for continued high-levels of reliability, it is increasingly vital for utility system operators to be able to understand and manage the power that comes onto the U.S. grid through DER in the residential and commercial sectors, specifically. Without doing so, there is an increased risk that utility system managers will inaccurately manage the grid.

This is not only a problem for system operators, but also for power market participants who may find it increasingly challenging to predict system pricing. For optimal grid operation, they must ensure that 1.) Utilities are able to predict the amount of DER assets coming online within their systems; 2.) The approximated deviation of existing data/modeling uses from actual real-time existing data, and 3.) The impact of this deviation/opportunity for more accurate forecasts and improvements upon current operational costings.

The first section is a summary of the transmission system and design overview for the current transmission system. The second section is a summary understanding the distribution system design aspects. The final section provides a comparison of the availability of a utility distribution system to the transmission system and providing an overview how an inaccurate system topology could limit the design and maintenance of the existing grid in regards to the emerging technologies' impacts.

2.1 Transmission System Overview vs. Distribution System Overview

The first step in understanding the difference between transmission and distribution systems is the fundamental differences between the mathematical equations between the two designs. The difference between the transmission and distribution power flow equations is with transmission, and one must consider the phase angle difference between two sources as the transmission system is connected as a network. However, with the distribution system, it is a radial system that only has one source that drives the calculations, and therefore, one does not have to worry about phasor math during its current design state. The first step will be to state the fundamental make-up and theory of an A.C. transmission line, and when voltage magnitudes and phase angle information is available between two substations, the real and reactive power flow can be calculated. Calculating the power flow includes calculating the M.W. flow on a transmission facility, which is the result of the resistive component (R) and is in-phase with the load being served. The Mvar flow on a transmission facility is the result of the reactive component (X). Mvars supply magnetizing current for inductive loads and charging current for capacitive loads. When calculating the impedance of a facility on a transmission system, the admittance matrix must be defined. Different lines have different values for resistance, inductance, and capacitance depending on the length, conductor spacing, and conductor cross-sectional area. Resistance is the property of the material that opposes current flow real power or watt losses due to I2R heating. The line resistance is dependent on conductor material, conductor cross-sectional area, and conductor length. In a purely resistive circuit, voltage and current are in phase; and the instantaneous power equaling the product of the two. Reactance is the opposition to current caused by capacitance and inductors. Reactance causes current to be out-of-phase with voltages. Inductive reactance (X_L) causes the current to lag the voltage, and capacitance reactance (X_C) causes the current to lead the voltage. Loads containing pure inductance or pure capacitance cause the current to be 90 degrees out of phase with the voltage. Inductance and capacitance depend on the conductor length, conductor cross-sectional area, and distance between phase conductors. Capacitive reactance increases as the cross-sectional area of the

conductor increases and decreases as the conductor spacing increases. Inductive reactance decreases as the cross-sectional area of the conductor increases and increases as the conductor spacing increases. To summarize, the total impedance includes the resistance, inductance, and capacitance and is termed impedance (Z). The reactive component of Z is X and is made up of inductance and capacitance and is typically greater than the resistive (R) component of a line. Also, it is noted that reactive components of impedance are greater for higher voltage lines than for lower voltage lines. As for the power flow on a distribution grid, it is calculated utilizing only one source simplifying the design as it exists today; however, there are significant hurdles that must be championed to prepare the distribution grid for two-way power flow due to the implementation of distributed energy resources (DER).

2.2 Transmission System Overview

Design and Development Principles

Transmission lines are used to connect generation sources to customer loads. In general, transmission lines connect the system's generators to its distribution substations. Transmission lines are also used to interconnect neighboring power systems. Since transmission power losses are proportional to the square of the load current, high voltages, from 69 kV to 765 kV, are used to minimize losses. The design and development of the transmission system are built to ensure it is reliable and robust and to ensure that all critical elements (e.g., substation, line, transformers, etc.) are protected from any disturbance on the power system. A disturbance can be anything from a line circuit breaker operating due to planned maintenance or a line sagging during heavy loading periods into poorly maintained vegetation and cause a fault to occur. The system must be planned to continuously serve all distribution substations to ensure power stays on at the customer level. The transmission system must be studied to ensure all protective relays will coordinate to any variations in the power flow, network bus voltages, machine rotor speeds, generator and transmission voltage regulators, prime mover controls, system loads M.W. and Mvar consumption are all considered when designing the transmission system. To summarize, design of the transmission system is to ensure after a disturbance, and the power system remains stable. For instance, following a disturbance, a power system is stable, then it will reach a new equilibrium point with the system integrity preserved. All generators and loads are connected through a single contiguous transmission system, some generators or loads may be disconnected by the isolation of the fault or intentional tripping to preserve the continuity of the operation of the transmission system. If following a disturbance, a power system is unstable, and then it will result in a run-away or run-down situation. With a progressive increase of angular separation of generator rotors and/or a progressive decrease of system voltages, an unstable condition could lead to cascading outages and a shutdown of significant portions of the power system. As well as ensuring the transmission system is designed to endure multiple outages planned and unplanned, each transmission system is unique from the perspective of system design, geography, performance, and load, but there are similarities in how assets

are evaluated and upgraded. The key objective in developing a reliable grid is to create the correct balance between the investment required to maintain the reliability on the system and construct the necessary upgrades to the transmission infrastructure. The evaluation of degrading asset performance and condition, as well as increased maintenance cost to the transmission system to determine when it's appropriate and cost-effective to replace versus repair. The transmission system is sophisticated and composed of an enormous number of assets that provide specific functionality and must work in unification with each other in the operation of the grid. The intricacies and analyses of the system must be broken down into multiple classifications to ensure the phase angle separation is maintained throughout the contiguous transmission system due to the varying phenomena and variables of impact. For instance, power system stability is broken down into three physical natural/main system parameters: rotor angle stability, frequency stability, and voltage stability. And then, those system parameters can be studied under different guises such as the size of the disturbance and then also the period for which it is studied. For example, determining the system's small-disturbance rotor angle stability and/or dynamic transient stability are studied within the microsecond time frame. Frequency stability is reviewed within the micro-second time frame as well as reviewing it over seconds/minutes. And then lastly, voltage stability is studied varying the degree of the disturbance and well as looking at the microsecond to minutes time frame after a disturbance occurs. Each analysis requires power system analysis tools to be able to run a large number of simulations and data required. To be able to run these analyses there are multiple software packages currently being utilized by the industry (i.e., General Electric (G.E.) Positive Sequence Load Flow (PSLF) and Siemens Power System Simulator for Engineers (PSS/E) are two dominate software programs used in the industry). Note the transmission system assets have been deployed over a long period using engineering principles, design standards, safety codes, and good utility practices that were applicable at the time of installation and have been exposed to varying operating conditions over their life [29].

Therefore, the analysis and focus on the transmission system to ensure it is maintained and developed has been being studied using commercialized software for over 30 years, and there are significantly different variables required to understand when determining how to maintain and develop the transmission grid with reverse power flow taken into consideration.

2.3 Distribution System Overview

Design and Development Principles

The design and development of the distribution system planning start at the customer level versus the power plants. The customer load, load type, and load factor determine the type of distribution system that is required. The customer loads are summed together for service from the secondary lines that are then connected to the distribution transformer that steps it up to the primary voltage and runs the feeder back to the substation (source). The primary distribution system load is then assigned to a substation that steps up to a transmission voltage. The customer loads technically determine the size and location of the substations as well as the routing and capacity of the associated transmission and distribution lines. The design of the distribution grid is an iterative approach. Each step in the process provides input for the step that follows. Note in this process; the planner is typically restricted by regulated voltages values, voltages dips, voltage flicker, as well as service reliability. There multitude of factors that also need to be considered, such as the transformers impedance, insulation levels, availability of spare transformers and mobile substations, and, most importantly, the rates that are charged to the customers [27].

The distribution planning problem is an attempt to minimize the cost of substations, feeders, laterals, as well as the cost of losses. Indeed, this collection of requirements and constraints has put the problem of optimal distribution system planning past a manageable calculation by a single engineer. Even though the distribution planning power flow equations are much simpler than transmission as it is a system designed for one-way power flow. The distribution planning approach will profoundly be altered with the evolution of DERs. The DERs will bring on an additional layer of challenges, and many of the assumptions upon which traditional distribution planning relies upon will be changed. DERs are creating two-way power flows on the distribution system that legacy distribution equipment was not designed for. DERs are also bewildering conventional load forecast methodologies and muddling the modeling of distribution feeders by introducing new kinds of generation sources or altering load profiles.

Refer to Fig. 3 for a snapshot in regards to the difference between the current methods used versus the future methods to be studied. The increase is significant and will take the power of software and machine learning algorithms to help aid in the approach.

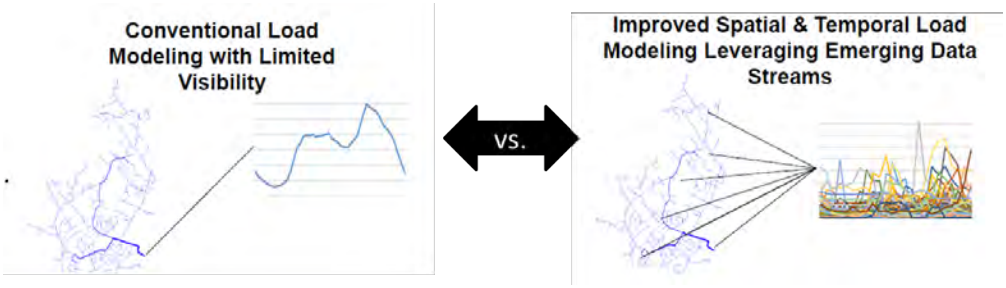


Figure 3: Illustration of the Difference Between Current Planning Methods Versus Future Methods.

2.4 Emerging Technology Driving the Need for Identification of System Topology

The transmission system has been under strict regulatory requirements since 2007, after the fall-out of the 2003 Northeast blackout. After the blackout, the NERC Standards went from becoming best practices to carry monetary fines. Therefore, the industry as a whole has collectively worked and collaborated to ensure that the system topology of the transmission grid is well known and modeled to be able to perform system planning and operational studies on the networked system. Initiated by the U.S. government, the rapidly expanding usage of distributed energy resources (DER) have been proliferating to incorporate emerging technologies that ultimately facilitate renewable energy (i.e., photovoltaic and storage devices). There is a significant amount of discussion within the electric utilities of when an electric utility should start to increase their usage of technology and strengthen their bench strength on how and when to plan for the proliferation of DER especially with a large amount of emerging technologies that are being introduced at a rapid pace and when the EDC should buckle down and work on managing the existing aged infrastructure. Industry-wide, EDCs have not been incentivized to maintain and invest in the distribution system over the last couple of decades as the reliability and resilience of the distribution power grid was enough, and load growth of the customer based allowed the investment to occur naturally. However, with the introduction of DER, Energy Efficiency, and Demand Response, the load growth is not as prevalent and makes the economics behind investment in technology and new assets harder to explain to customers.

A review of five utilities there is key drivers, methodology, and tools that are being investigated to prepare for the proliferation of DER; however, the utilities are at all range of the spectrum. The key drivers include regulatory compliance and operational necessity. The methodology that is being investigated is the timeline for DER planning, incentivizing preferred interconnection locations, and then the cost recovery/rate restructuring. Every utility is located in a separate geographical and political atmosphere, which drives the differences as well as the challenges to understand the best path forward.

Lastly, the toolset that the utilities need most is the visual maps providing preferred interconnection locations, acquiring increased distribution planning software tools that require the system topology to be modeled, and lastly, how the operations center plan to observe and manage the DER once it is connected.

3.0 Data-Driven Machine Learning Algorithms to Identify System Topology

The ongoing deployment of Distributed Energy Resources (DERs), while bringing benefits, introduces significant challenges to the electric utility industry, especially in the distribution grid. These challenges call for closer monitoring through state estimation, where real-time topology recovery is the basis for accurate modeling. Previous methods are either based on outdated maps or an ideal assumption of an isolated sub-network for topology recovery, e.g., within one transformer. This requires field engineers to identify the association, which is costly and may contain errors.

A density-based clustering method is proposed to solve these problems that leverages both voltage domain data and the geographical space information. The goal of such a method is to efficiently segment datasets from a large utility customer pool, after which other topology reconstruction methods can carry over. Specifically, we show how to use voltage and GIS information to refine the connectivity within one transformer. We also show how to improve existing method further for robustness. To give a guarantee, we show a theoretic bound for our clustering method, providing the ability to explain the performance of the machine learning method. Numerical results on both IEEE test systems and utility networks show the outstanding performance of the new method. An implementation is also demonstrated in the field.

The utility under study herein is an excellent example of an electric distribution company (EDC) that is in a position to start preparing themselves for the changing of tides in regards to how the distribution system is planned and operated and why it is crucial for them to have the tool set and system topology modeled in a power system analysis software platform such as Eaton's CYME Power Engineering Software Solutions to analysis the future complicated planning needs of the distribution system.

However, such approaches usually require narrowing down data beforehand. They will fail if data pulling from a large number of buses, e.g., 100K-customer network, are given together with different network devices, e.g., transformers. Even worse, many distribution

assets do not have any measurement data available. For example, a large portion of network nodes are probably unmeasured, e.g., transformers, poles, and underground facilities. This makes the "directly recovered" connectivity among customers meaningless.

For such a purpose, we propose a hierarchical data-driven method to divide the network to avoid extensive human effort and to prepare the network for subsequent algorithms. Such a method integrate theoretic approaches, practical challenges, and GIS information to resolve topology recovery problems robustly against limited measurements and scalability issues [21]. For example, in order to recover the connection between smart meters and transformers, we perform a comprehensive analysis of the families of clustering methods in the voltage space and geographical space.

We discover that density-based methods are well-suited for outlier detection while methods, where the number of clusters is prespecified, are better for associating smart meters to their parent transformer.

Subsequently, we use both the voltages and street information to characterize the subnetwork for algorithmic solutions. For example, we use the street information to classify the data into the streets and further classify into segments of a street for long streets. Furthermore, as Euclidean distance is unable to provide accurate distribution clusters for utility networks, we used Bing Maps API to compute distance along a street to improve the association of smart meters to their parent transformers. Our algorithm provides an easy, fast, and scalable processing tool for real-time data.

Numerical experiments are carried out on the standard distribution test beds, e.g., IEEE 123-bus, and by our partner EDC's local grid with $10,000$ customers. The result shows that the proposed method segments the distribution grids accurately and helps to achieve a highly accurate topology estimate.

First, the research developed a physically meaningful, clustering method to roughly separate the data. Second, both the voltages and street information was used to to characterize the data for algorithmic solutions to refine the solution. In which during such a process, it was observed that the direct application of Euclidean distance contradicts to utility practice of network planning, so the Bing Maps API was used to obtain the new distance along the street.

3.1 System Model

To define the method for topology clustering, we describe time series voltage data given by smart meters and their location data. For example, the latitude-longitude pairs in radians for N smart meters $\mathbf{l}^1, \cdots, \mathbf{l}^N \in \mathbb{R}^{2\times 1}$ are stored as row vectors in matrix $\mathbf{L} \in \mathbb{R}^{N\times 2}$. $\mathbb{R}$ represents the set of real numbers. The voltage time series with T timeslots for N smart meters $\mathbf{v}^1, \cdots, \mathbf{v}^N \in \mathbb{R}^{T\times 1}$ are stored as row vectors in matrix $\mathbf{V} \in \mathbb{R}^{N\times T}$. Therefore, the combined dataset $[\mathbf{L}, \mathbf{V}] \in \mathbb{R}^{N\times(T+2)}$ with row vectors $\mathbf{x}^1, \cdots, \mathbf{x}^N$ for N smart meters. In addition to smart meters, there are k transformers forming k clusters of smart meters in the distribution grid. $Cluster(j)$ represents a vector of indices of all smart meters in the j^{th} cluster. Due to radial configuration, a smart meter $i \in \{1, \cdots, N\}$ is uniquely present in a cluster $j \in \{1, \cdots, k\}$ that is supplied by a common transformer. There exists a many-to-one mapping $f : i \to j$.

For correlating these variables, a distribution system is characterized by buses $\mathcal{V} = 1, 2, \cdots, N$ and by branches $\mathcal{E} = (i, i'), i, i' \in \mathcal{V}$. The voltage measurement data at bus i and time t can be represented as $v_i(t) = |v_i(t)| \exp^{j\cdot\theta_i(t)}$, where $|v_i(t)| \in \mathbb{R}$ denotes the magnitude of the bus instantaneous voltage in per-unit, and $\theta_i(t)$ denotes the phase angle of the voltage in radian. The measurements in $\mathbf{v}^i$ are steady-state voltages over a period according to utility collection speed. We define the problem below.

- Problem: Identify smart meter to transformer connectivity
- Given: Smart meter voltage data and location $[\mathbf{V}, \mathbf{L}]$,
- Find: The mapping rule $f : i \to j$.

3.2 Clustering Methods for Grid Segmentation

3.2.1 Data Sources

In the past, most topology analyses in the distribution grid assumes to use AMI data only, e.g., voltages. While this is a good start for better topological understanding, GIS information is equally important and many utilities have such information for usage. Even if a utility does not have GIS information on smart meters, we can have the house addresses or apartment addresses to use. For example, we can convert the addresses into latitudes and longitudes of the smart meters by using Google Maps API. Therefore, we propose to include both information from geographical space and the voltage space for topology clustering. In the next section, we analyze different clustering methods based on the two data sources.

3.2.2 Metric Evaluation for Clustering Algorithm Design

Data clustering requires one to understand how to group a set of objects based on their similarity of attributes and/or their proximity in the vector space. A key first step is having that basic knowledge about the data that is being used and then to understand how it is being used. For data clustering, there are numerous approaches. Therefore, we investigate the compatibility of of them before using or modifying a method. For these methods, three categories are popular in the machine learning fields. One is to consider the group properties, e.g., calculate the sum of distances within each cluster. The second category is to bound the cluster with a limit, e.g., maximum diameter for clusters. The third category investigates the importance of cluster "density", e.g., the number of data points in a neighborhood of points.

By analyzing their suitability for power distribution data, we analyze the representative ones from the three classes, namely, K-means for average cluster distance, Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) for maximum diameter, and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) for density [67]. Fig. 4 provides the visual ideas of the three categories.

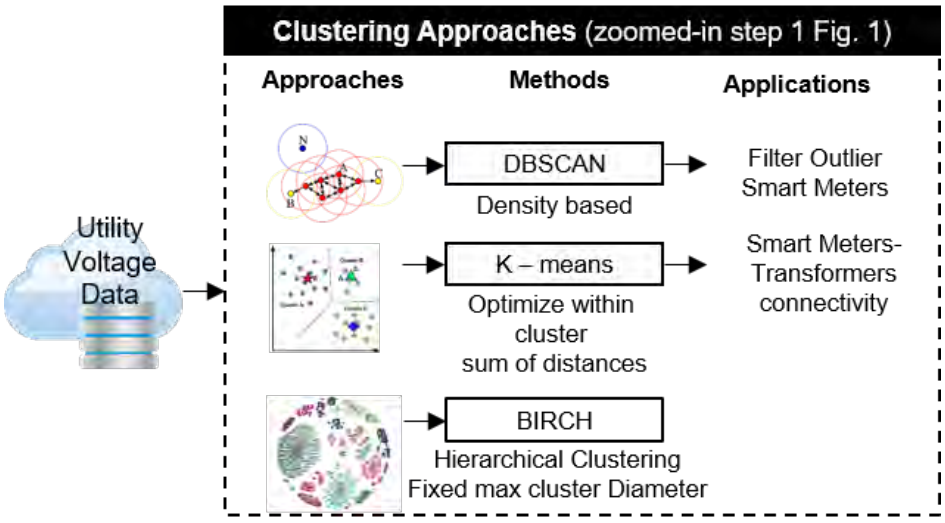


Figure 4: Comparison of Three Important Families of Algorithms for Clustering Based on Both GIS and AMI Data.

3.2.2.1 K-means for Average Distances

One idea for clustering is to consider the average distances for all the members in a group.

For example, K-Means creates k centroids $\overline{x^j} = \frac{1}{n_j} \sum_{i \in cluster(j)} x^i$, where n_j is the number of smart meters in cluster j.

It aims at minimizing the squared error loss, $J = \sum_{j=1}^{k} \sum_{i \in cluster(j)} (||x^i - \overline{x^j}||)^2$, where $||x^i - \overline{x^j}||$ is the Euclidean distance between a point x^i and centroid $\overline{x^j}$ iterated overall points in the j_{th} cluster, for all n clusters.

Drawback for Power Data: While such a method can be used for clustering, it is not good to be directly used in distribution grids. For example, it is quite common for parallel streets to have their own transformers. However, as the span of the street can be quite long, the GIS information can give incorrect clustering decision, making the idea of including more information than voltage to be vulnerable. Even in the voltage space, different locations can have similar voltage ranges. So, it can confuse the K-means.

3.2.2.2 BIRCH for Maximum Cluster Distance

Instead of looking at the grouping effects in K-means, one can also bound the extreme points, where the BIRCH algorithm is well-known.

It requires three parameters, the branching factor B_r, the threshold T, and the cluster count k.

The cluster centers $\overline{x^j} = \frac{1}{n_j} \sum_{i \in cluster(j)} x^i$, where n_j is the number of smart meters in cluster j, and the cluster radii $R_j = \sqrt{\frac{1}{n_j} \sum_{i \in cluster(j)} (x^i - \overline{x^j})^2}$ can then be computed for each cluster.

Every point is assigned to the nearest-center subcluster.

Drawback for Power Data: For distribution systems, the geographical radius can be different depending on the location of the grid and if the grid is newly planned or if the grid is in the urban area. For example, both long feeder and short feeder are common in power domain. Therefore, it is unwise to have a hard limit on the diameter for the GIS space. On the voltage space, the voltage data can have outliers due to measurement error or due to a change of operational points. Therefore, such a method is not preferred.

3.2.2.3 DBSCAN for Local Densities

In the two approaches above, the focuses are on either the group property or on the property of an extreme limit. Another idea is to focus on a subgroup of points and check how the trend is propagating, which is the third category. For example, DBSCAN includes a point or not based on two parameters: (1) a neighborhood region specified by the radius ϵ and (2) the minimum number of data points $minPoints$ in the neighborhood. The algorithm counts the data points in the sphere of radius ϵ around a data point and includes it in the cluster if it exceeds $minPoints$.

Advantages for Power Distribution Grids: As the power distribution grid is planned in a radial structure, the density is relatively predictable in the direction that the grid grows, e.g., along the streets. Therefore, a density-based method can provide correct clusters for GIS information. For voltage, each voltage will have its own dense region. So, working together with the GIS information can easily boost the performance.

3.2.3 Proposed Density-Based Method

In this section, we define the proposed algorithm mathematically. First, we define the distance in both GIS and the voltage space. Then, we show how to evaluate the density. Finally, we show how to use density to conduct clustering.

3.2.3.1 Distance for GIS Data and for Voltage Data

For GIS data, let $\mathbf{l}^1, \mathbf{1}^2 \in \mathbf{L}$ be two latitude-longitude pairs in radians. The distance in km between them on Earth's surface is given using the Haversine formula $d_{\mathbf{L}}(\mathbf{l}^1, \mathbf{l}^2) = 6371 \cdot \{\arccos(\sin(l_1^1)\sin(l_1^2) + \cos(l_1^1)\cos(l_1^2) \cdot \cos(l_2^2 - l_2^1)\}$. For the distance in the voltage domain, we use mutual information to quantify the distance. Specifically, the voltage-distance between two points $\mathbf{v}^1, \mathbf{v}^2 \in \mathbf{V}$ is defined as $d_{\mathbf{V}}(\mathbf{v}^1, \mathbf{v}^2) = \frac{1}{I(\mathbf{v}^1, \mathbf{v}^2)}$ where $I(\mathbf{v}^1, \mathbf{v}^2)$ is the mutual information between $\mathbf{v}^1$ and $\mathbf{v}^2$. The key idea of mutual information-based topology analysis in the past is based on using voltage correlation in a probabilistic way [63]. A distribution system typically has a radial structure. Therefore, we can represent the voltage data in a graphical model via the joint probability density $P_{\mathbf{V}}(v^2, v^3, \cdots, v^N) = P_{\mathbf{V}}(v^2) \cdot P_{\mathbf{V}}(v^3|v^2) \cdots P_{\mathbf{V}}(v^N|v^2, \cdots, v^{N-1})$, where we assign the swing bus as bus 1 with a deterministic value, which is eliminated from the measurements. Based on such a chain rule, mutual information can be used for measuring voltage similarity, e.g., in the discrete-time scenario mutual information is defined as $I(\mathbf{v}^1, \mathbf{v}^2) = \sum_{i=1}^{T} \sum_{j=1}^{T} p_{(\mathbf{v}^1, \mathbf{v}^2)}(v_i^1, v_j^2) \ln\left(\frac{p_{(\mathbf{v}^1, \mathbf{v}^2)}(v_i^1, v_j^2)}{p_{\mathbf{v}^1}(v_i^1) p_{\mathbf{v}^2}(v_j^2)}\right)$.

Essentially, it is a weighted sum measuring the averaged similarity between the joint distribution

$p_{(\mathbf{v}^1, \mathbf{v}^2)}(v_i^1, v_j^2)$ and the products of the individual distributions, $p_{\mathbf{v}^1}(v_i^1) \cdot p_{\mathbf{v}^2}(v_j^2)$. For example, if v_i^1 and v_j^2 are independent random variables, $p_{(\mathbf{v}^1, \mathbf{v}^2)}(v_i^1, v_j^2) = p_{\mathbf{v}^1}(v_i^1) \cdot p_{\mathbf{v}^2}(v_j^2)$, making $\ln\left(\frac{p_{(\mathbf{v}^1, \mathbf{v}^2)}(v_i^1, v_j^2)}{p_{\mathbf{v}^1}(v_i^1) p_{\mathbf{v}^2}(v_j^2)}\right) = 0$, showing no connection between buses i and j. On the other hand, neighboring smart meters sharing a common transformer have similar voltage profiles resulting in a high mutual information.

Based on the distances in the voltage and GIS domain, the combined distance of two datapoints $\mathbf{x}^1, \mathbf{x}^2 \in [\mathbf{L}, \mathbf{V}]$ is given as $d_{[\mathbf{L}, \mathbf{V}]}(\mathbf{x}^1, \mathbf{x}^2) = d_{\mathbf{L}}(\mathbf{l}^1, \mathbf{l}^2) + d_{\mathbf{V}}(\mathbf{v}^1, \mathbf{v}^2)$.

3.2.3.2 Evaluate the Density in the Combined Space of GIS and Voltage Data

To define a notion of density in $(n+1)$-dimensional space $[\mathbf{L}, \mathbf{V}]$, we first consider the two-dimensional space L. For arguments sake, consider two points $\mathbf{l}^1, \mathbf{l}^2 \in \mathbf{L}$. $\mathbf{l}^1 = (l_1^1, l_2^1)$, $\mathbf{l}^2 = (l_1^2, l_2^2)$ in a 2-D space. The Euclidean distance for these two points is $d(\mathbf{l}^1, \mathbf{l}^2) = [(l_1^1 - l_1^2)^2 + (l_2^1 - l_2^2)^2]^{0.5}$. If we fix the distances to be less than ϵ, then we obtain the following: $d(\mathbf{l}^1, \mathbf{l}^2) = [(l_1^1 - l_1^2)^2 + (l_2^1 - l_2^2)^2]^{0.5} < \epsilon$. Squaring both sides yields: $(l_1^1 - l_1^2)^2 + (l_2^1 - l_2^2)^2 < \epsilon^2$. The equation looks similar to the equation of a circle with radius ϵ and center is at the point (l_1^2, l_2^2). Thus, the algorithm counts the data points in the sphere of radius ϵ around a data point and includes it as a core point in the cluster if it exceeds $minPoints$. However, using Euclidean distance is wrong due to Earth's spherical shape, and therefore, we use Haversine distance that gives the distance on the surface of Earth in km.

Definition 1**:** (ϵ-neighborhood of a point) The ϵ-neighborhood of a datapoint $\mathbf{x}^1 \in [\mathbf{L}, \mathbf{V}]$, denoted by $N_r(\mathbf{x}^1)$, is defined by $N_r(\mathbf{x}^1) = \left\{\mathbf{x}^2 \in [\mathbf{L}, \mathbf{V}] : d_{[\mathbf{L},\mathbf{V}]}(\mathbf{x}^1, \mathbf{x}^2) < \epsilon\right\}$.

The ϵ-neighborhood of a point is a notion of the density of points. If $N_r(\mathbf{x}^1) > minPoints$ then $\mathbf{x}^1$ is a *core point*. The points at the boundary of a cluster may not qualify to be a core point. For such points, we cluster them with a core point if they are in ϵ-neighborhood of a core point.

Definition 2**:** (Directly density-reachable) A point $\mathbf{x}^2 \in [\mathbf{L}, \mathbf{V}]$ is directly density-reachable from a point $\mathbf{x}^1 \in [\mathbf{L}, \mathbf{V}]$ with respect to (wrt) ϵ and $minPoints$, if 1) $\mathbf{x}^2 \in N_r(\mathbf{x}^1)$, and 2) $N_r(\mathbf{x}^1) \geq minPoints$ ($\mathbf{x}^1$ is a *core point*).

Directly density-reachability is not transitive. To ease algorithmic development, we need a transitive property.

Definition 3**:** (Density-reachable) A point $\mathbf{x}^2 \in [\mathbf{L}, \mathbf{V}]$ is *density-reachable* from a point $\mathbf{x}^1 \in [\mathbf{L}, \mathbf{V}]$ wrt ϵ and $minPoints$, if there is a sequence of points $\mathbf{y}^1, \cdots, \mathbf{y}^m \in [\mathbf{L}, \mathbf{V}], \mathbf{y}^1 = \mathbf{x}^2$, $\mathbf{y}^m = \mathbf{x}^1$, so that $\mathbf{y}^{i+1}$ is directly density reachable from $\mathbf{y}^i$.

Definition 4**:** (Density-connected) A point $\mathbf{x}^2$ is *density-connected* to a point $\mathbf{x}^1$ wrt ϵ and $minPoints$ if there is a point $\mathbf{x}^3$ such that $\mathbf{x}^2$ and $\mathbf{x}^1$ are density-reachable from $\mathbf{x}^3$.

According to DBSCAN, two points are in the same cluster if and only if they are density connected. Density connectedness is a reflexive, symmetric, and transitive property. Therefore, it is guaranteed to form equivalence classes that are the clusters.

3.2.3.3 Density-based Algorithm

We start with some point, $\mathbf{x}^1$, and check if it is a core point by the condition $N_r(\mathbf{x}^1) \geq minPoints$. Essentially, the distance between $\mathbf{x}^1$ and $\mathbf{x}^2$ is not the usual Euclidean distance but the specific Haversine distance. If $\mathbf{x}^1$ is a core point, we keep it as a starting point for the cluster. If $\mathbf{x}^1$ is not a core point, we put it in the outliers list and randomly select another point and repeat the procedure until we find a core point. In such a case, all of $N_r(\mathbf{x}^1)$ are in the same cluster as $\mathbf{x}^1$. Next, we individually check each point in $N_r(\mathbf{x}^1)$ for core point. All newly discovered core points are inserted in a queue. Next, we repeat the same procedure for each core point in the queue, thereby adding new points to the cluster and the core points queue until the core points queue is empty, making cluster one complete. Now, we randomly start searching the remaining points for a new core point for the second cluster and repeat such a process. Algorithm 1 is different from the original DBSCAN [17] as it considers only the core points. Algorithm 2 is an improved version that is robust against adversarial noise [26].

Algorithm 1: Improved DBSCAN

Input: X, ϵ, $minpts$

$H := \{x \in X : |B(x, \epsilon) \cap X)| \geq minpts\}$.

$G :=$ undirected graph with vertices H edge between $x, x' \in H$ if $|x - x'| \leq \epsilon$.

return connected components of G.

Algorithm 2: Robust DBSCAN

Input: X, ϵ, $\widetilde{\epsilon}$, $minpts$

$H := \{x \in X : |B(x, \epsilon) \cap X)| \geq minpts\}$.

$D :=$ DBSCAN(X, $\widetilde{\epsilon}$, $minpts$).

$\mathcal{C} := \{C \cap H : C \in D\}$.

return $\mathcal{C}$.

3.3 Guarantee of the Density-based Algorithm

In this section, we provide a guarantee for the Robust DBSCAN in Algorithm 2 and show that the algorithm is robust against the addition of new data. In particular, we show that adding l new utility customers with smart meter voltage and location data to the original data does not change the original clusters and the cluster assignments to the original points remain unchanged, i.e., the original points that were clustered together (separate), remain together (separate) after adding new points.

Assumption 1. *For the guarantee, we need the theoretical density $f(\mathbf{x})$ to be differentiable.* This suggests that the density should be smooth, and there should be no outliers in the data. For GIS data, a discontinuous density would mean a single house left alone from other houses. Typically this is not the scenario in utility service areas. For voltage data, a discontinuous density would mean an error in smart meter measuring instrument. Usually, such an outlier can be easily detected by a utility and fixed.

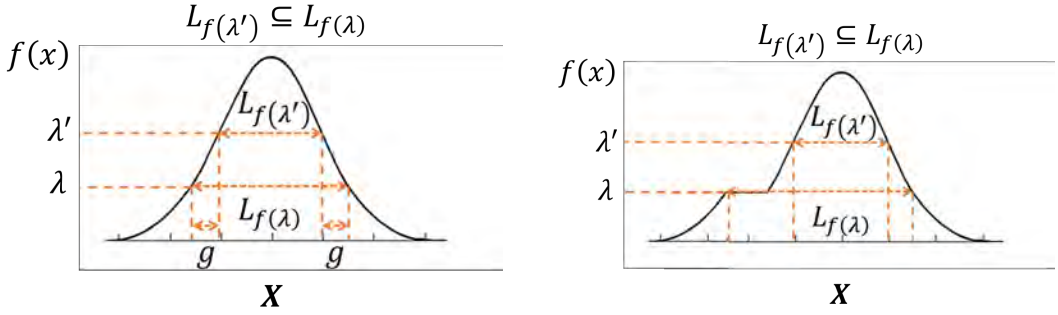
In order to have a mathematical analysis of the density, we need to define *superlevel-set $L_f(\lambda)$ of the density function f corresponding to a given threshold (level) λ* as a set of all points in the dataset $[\mathbf{L}, \mathbf{V}]$ with density equal to or greater than the threshold λ. Moreover, if Assumption 1 holds, the superlevel-sets consist of closed intervals rather than discrete points. The concept of superlevel-set will be useful in Assumption 2 (curvature). Usually, the shape of a density function has one or more overlapping bell curves or some flat regions. Therefore, if we have two levels λ and λ' such that $0 < \lambda \leq \lambda' < \|f\|_\infty$, where $\|f\|_\infty$ represents the peak density, then the Superlevel-set for level λ' is a subset of the Superlevel-set for level λ.

Mathematically, $L_f(\lambda') \subseteq L_f(\lambda)$.

Given a continuous set A, if we "trim" set A from all sides of the boundary by a depth δ, the remaining set is called the *δ-interior of A*. For example, in Fig. 5a, we can "trim" $L_f(\lambda)$ from its boundary by a depth g to make it a subset of $L_f(\lambda')$. Such a concept is also useful for Assumption 2.

To provide a guarantee for robustness of density-based clustering, we need the density function to decay around the cluster boundaries so that the clusters are salient enough to be detected. Simply, we need no flat regions in the density curve. Flat regions in density curve can be avoided in the following way. For Fig. 5a, assume g is an increasing linear function of $(\lambda' - \lambda)$ and assume for all $0 < \lambda \leq \lambda' < \|f\|_\infty$, where $\|f\|_\infty$ represents the peak density, we have $L_f(\lambda) \ominus g(|\lambda - \lambda'|) \subseteq L_f(\lambda')$. This is because there is no flat region in Fig. 5a. However, for Fig. 5b, if we set λ and λ' just below and above the flat region, we do not obtain $L_f(\lambda) \ominus g(|\lambda - \lambda'|) \subseteq L_f(\lambda')$ due to the flat region. Assumption 2 below gives a formal description of this concept.

Assumption 2 (Curvature). *There exists $C_\beta > 0$ and $\beta > 0$ such that the following holds. For any $0 < \lambda < \lambda' < \|f\|_\infty$, we have $L_f(\lambda) \ominus g(|\lambda - \lambda'|) \subseteq L_f(\lambda')$* where $g(r) = C_\beta \cdot r^\beta$.



(a) A Non-Flat Distribution Satisfies Assumption 2.

(b) A Flat Distribution Does Not Satisfy Assumption 2.

Figure 5: Examples of Distributions Based on Assumption 2.

In voltage domain, a density function satisfying Assumption 2 means that the voltage distances (inverse of the mutual information) gradually increase as we move from the center of clusters (houses supplied by the same transformers) to external areas. This assumption ensures that the density function is not flat.

Moreover, this assumption forces sufficient density decay around the superlevel sets so that the superlevel sets are salient and can be detected easily [26].

We can introduce a slightly different density estimator concept than $N_r(\mathbf{x})$ i.e., to keep the number of points k fixed and adjusting the radius $r_k(\mathbf{x})$ to enclose k nearest neighbor points with the sphere called the k-NN density estimator. A lot of literature is based on this approach, formally defined as $f_k(\mathbf{x}) := \frac{k}{n \cdot v_D \cdot r_k(\mathbf{x})^D}$, where v_D is the volume of a unit ball in $d_{[\mathbf{L},\mathbf{V}]}$, $r_k(\mathbf{x})$ is the adjusted radius of the sphere to enclose k points. $v_D \cdot r_k(\mathbf{x})^D$ is the volume of the sphere with radius $r_k(\mathbf{x})$, and $\frac{k}{v_D \cdot r_k(\mathbf{x})^D}$ is the number of points per unit volume. In order to remove the effect of the total number of points n, we divide it by n. Once we have the required assumptions and definitions, now we can go ahead with the proof.

3.3.1 Guarantee of Original Smart Meters Belonging to the Same Cluster after the Addition of l More Smart Meters

We now show robustness guarantees on the core points returned by Algorithm 2. In particular, we show that adding l new utility customers with smart meter voltage and location data to the original data does not change the original clusters. The cluster assignments to the original points remain unchanged i.e., the original points that were clustered together (separate), remain together (separate) after adding new points. That is, when running Algorithm 2 on $[\mathbf{L}, \mathbf{V}]$ vs running it on $[\mathbf{L}', \mathbf{V}']$ with l additional samples, any new core points that appear will be near the original core points.

The k-NN density estimation error can be given by a probabilistic bound between the true density $f(\mathbf{x})$ and the k-NN density estimation $f_k(\mathbf{x})$. Such a bound can be used to identify the upper bound of the theoretical density given the k-NN density estimation via density-based clustering. The upper bound of the true density can be used to provide a guarantee for core points. For measuring $f_k(\mathbf{x})$, if k is very small, it can lead to estimation errors due to less samples within the sphere, reducing the estimation accuracy. Therefore, to provide a confidence level $(1 - \delta)$ for the bound, one needs to have a lower bound on k. The lower bound on k is directly related to the sample size n. Moreover, k is directly related to the confidence level $(1 - \delta)$. Lemma 1 directly follows from Lemma 3 and 4 of [11, 26].

Lemma 1. (k-NN density estimation accuracy). *Let* $0 < \delta < 1$. *Suppose that* f *satisfies Assumption* 1. *Then the following holds for some constants* C *and* C_l *depending on* f. *Suppose* k *satisfies* $k \geq C_l \cdot \log(\frac{1}{\delta^2}) \cdot \log n$. Then with probability of at least $1 - \delta$, the following holds:
$\sup_{\mathbf{x} \in [\mathbf{L}, \mathbf{V}]} |f(\mathbf{x}) - f_k(\mathbf{x})| \leq \left(\frac{\log(\delta^{-1}\sqrt{\log n})}{\sqrt{k}} + \left(\frac{k}{n}\right)^{\frac{\alpha}{D}} \right)$.

Lemma 1 provides the limit to the error in the k-NN estimator accuracy $f_k(\mathbf{x})$. Indeed, the range of error is directly related to the confidence level $(1 - \delta)$. Also, a greater sample size n can lead to greater error if k is small since the number of points within the sphere will be even smaller as compared to the total sample size n. Moreover, higher degree of continuity α of the density function will result in lower error.

From Assumption 1, we have that the density function is continuous. Moreover, from Assumption 2, we have that the density function has a curvature and is never flat. Furthermore, using Lemma 1, we have that the true density will not be much different than measured by DBSCAN. Therefore, the new l points will lie close to the original clusters. In fact, using the above three arguments, we can calculate the probabilistic maximum extension $\widetilde{r}$ from the original DBSCAN clusters. Therefore, the new clusters C' will be bounded by the original clusters extended by the distance $\widetilde{r}$ with a confidence of $1 - \delta$. The lower bound on k remains the same as Lemma 1, however, the total number of points becomes $(n + l)$.

Lemma 2. *Suppose that Assumptions* 1 *and* 2 *hold. There exists constants* C_l *and* C *depending on* f *such that the following holds. Let* $0 < \delta < 1$ *and* k *satisfy* $k \geq C_l \cdot \log(\frac{1}{\delta^2}) \cdot \log(n+l)$, *and* $\widetilde{\epsilon} > \epsilon > 0$. *Let* $\mathcal{C}$ *and* $\mathcal{C}'$ *be the core points returned by Algorithm 2 when run on* $[\mathbf{L}, \mathbf{V}]$ *and* $[\mathbf{L}', \mathbf{V}']$, *respectively. With probability at least* $1-\delta$, *the following holds:* $\mathcal{C}' \subset \mathcal{C} \oplus \widetilde{r}$, *where* $\oplus$ *denotes a tube around a set (i.e.* $A \oplus r := \{x \in [\mathbf{L}, \mathbf{V}] : \inf_{a \in A} |x - a| \leq r\}$*), and there exist* $\widetilde{r} < \infty$. We do not give proof of Lemma 2. It follows from Assumptions 1 and 2, and Lemma 1 [26]. The result $\mathcal{C}' \subset \mathcal{C} \oplus \widetilde{r}$ suggests that the new points lie within the tube of thickness $\widetilde{r}$ around the original clusters. Therefore, if the edges of the original clusters are at a distance $2\widetilde{\epsilon} + 2\widetilde{r}$, then there will not be any original clusters merging to form one cluster. Moreover, if $\widetilde{r} < \widetilde{\epsilon}$, then the new points will not create outliers or form separate clusters.

Theorem 1. *Suppose that conditions of Lemma* 2 *hold. Let* $\mathcal{C}$, $\mathcal{C}'$ *be the output of Algorithm 2 on* $[\mathbf{L}, \mathbf{V}]$ *and* $[\mathbf{L}', \mathbf{V}']$, *respectively and define the minimum inter-cluster distance*

of the returned clusters $R := \min_{C_1,C_2\in\mathcal{C},C_1\neq C_2} \min_{\mathbf{x}^1\in C_1,\mathbf{x}^2\in C_2} d_{[\mathbf{L},\mathbf{V}]}(\mathbf{x}^1,\mathbf{x}^2)$. *If additionally, the following holds:* $\widetilde{r} \leq \widetilde{\epsilon} \leq \frac{1}{2}R - \widetilde{r}$, *then* $|\mathcal{C}| = |\mathcal{C}'|$ *(i.e. the number of clusters does not change) and there exists a one-to-one mapping of the clusters* $\sigma : \mathcal{C} \to \mathcal{C}'$ *such that* $\mathcal{C} \subset \sigma(\mathcal{C})$ *for all* $C \in \mathcal{C}$ *(i.e. original clusters are preserved).*

Proof. Note that all the points appearing in a cluster of $\mathcal{C}$ will also appear in some cluster of $\mathcal{C}'$. By Lemma 1, we have that any newly appearing points in $\mathcal{C}'$ will be at a distance of at most $\widetilde{r}$ from a point appearing originally in $\mathcal{C}$, mathematically $\mathcal{C}' \subset \mathcal{C} \oplus \widetilde{r}$. From the assumption $\widetilde{\epsilon} \geq \widetilde{r}$, we have that the radius hyperparameter for DBSCAN is lesser than $\widetilde{r}$, then such new points will become reconnected to the same cluster in $\mathcal{C}$ since they will be present in the sphere of radius $\widetilde{r}$. Finally, from the assumption $\widetilde{\epsilon} \leq \frac{1}{2}R - \widetilde{r}$, we have that the original clusters are separate by more than $2\widetilde{\epsilon} + 2\widetilde{r}$, this means that no two distinct clusters in $\mathcal{C}$ will become merged in $\mathcal{C}'$.

Table 1: Summary of the Voltage Data Provided by the Partner Utility.

Type	Raw voltage		Cleaned voltage
	Phase A	Phase C	
Total number	$8,975 \times 8,640 = 77,544,000$	$394 \times 8,640 = 3,404,160$	$3,442 \times 8,640 = 29,738,880$
Unique mac addresses	$8,975$	394	$3,442$
Starting time	2016/7/22 4:05:00	2016/7/22 4:05:00	2016/7/22 4:05:00
Ending time	2016/8/21 4:00:00	2016/8/21 4:00:00	2016/8/21 4:00:00
Units	Volt	Volt	Volt

3.4 Numerical Validation

3.4.1 Data Description

The simulations are implemented on the IEEE PES distribution networks for IEEE benchmark systems, such as 8-bus and 123-bus systems. We also implement our algorithm on an utility grid. For benchmark systems, feeder bus is selected as the slack bus. The historical data have been preprocessed by the MATLAB Power System Simulation Package (MATPOWER) and OpenDSS. To simulate the power system behavior in a more realistic pattern, the load profiles from Pacific Gas and Electric Company (PG&E) and "ADRESConcept" Project of Vienna University of Technology [25] are adopted as the real power profile in the subsequent simulation. PG&E load profile contains hourly real power consumption of $123,000$ residential loads in the North California, USA. "ADRES-Concept" Project load profile contains real and reactive powers profile of 30 houses in Upper-Austria. The data were sampled every second over 14 days. Transformers' GPS coordinates were added manually according to power domain knowledges, e.g., from our partner utility grid.

For the utility grid, it is a mid-sized northeast system that includes approximately $600,000$ customers, $7,200$ miles of overhead conductors, $250,000$ poles, $108,000$ transformers, $4,500$ miles of cable, $1,000$ sectionalizers, 400 capacitors, and 500 network protectors. A sample of $10,000$ customers AMI voltage data was used as well as the nearby transformers' GPS coordinates, and the GPS coordinates of the poles. A summary of the voltage information is shown in Table 1.

3.4.2 Robust Clustering

3.4.2.1 Validation on IEEE-123 Test Case System

The IEEE 123-bus system is separated into two random clusters, as shown in Fig. 6. As the IEEE 123-bus system does not provide any GPS coordinates of the nodes, we estimate the location coordinates using the image of the test feeder. The system is disconnected at a bus to create two separate subsystems. The bus location that we choose does not impact

the exercise; therefore, one could choose any part of the system to create subsystems. For this example, the system was split evenly between bus 67 and 68 to provide an even split. On each subsystem, we run load flow analysis on 500 scenarios to generate an example AMI voltage dataset over a typical load cycle, similar to what the utility provided us. Coupled with "GPS coordinates" explained in the setup, we run the clustering algorithms multiple times, changing the hyperparameters for each algorithm to better understand the advantages and disadvantages of each algorithm investigated.

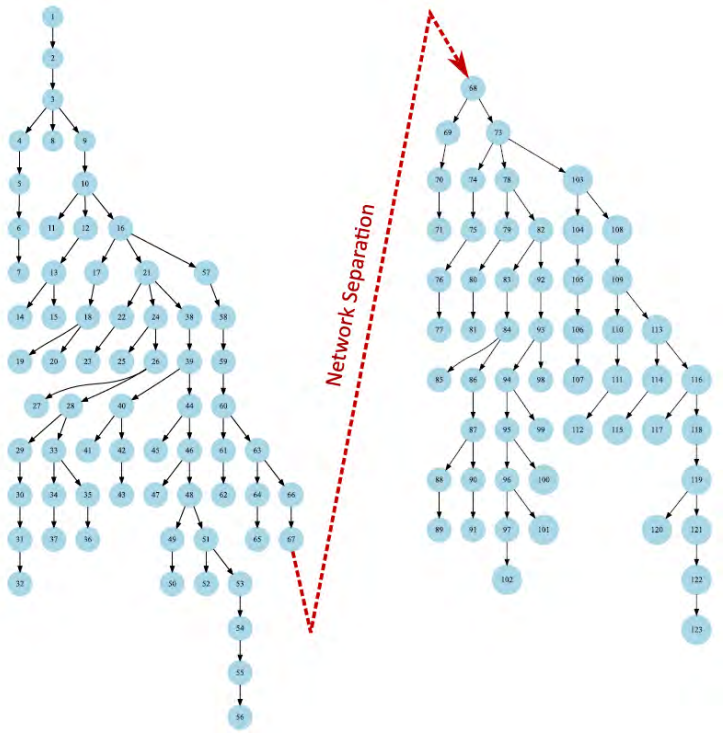
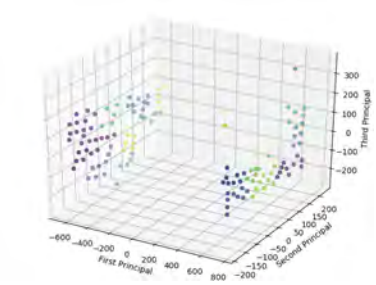
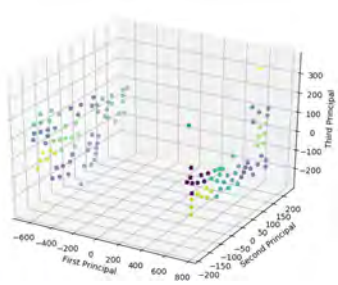


Figure 6: Network Partition. The IEEE 123-Bus System Was Used to Understand the Different Dynamics of the Three Clustering Algorithms for Illustration Purposes.

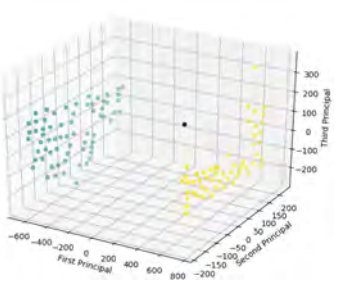
Fig. 7 gives an example of each clustering method on the same dataset, where our proposed algorithm has its algorithm in Fig. 7c. By comparing this figure with the other two on the left, we can see that only our method is clustering consistantly. We observe this throughout our simluation on different loads and topology, showing the power of integrated design of machine learning method with the needs considered in power systems.



(a) Result of Birch clustering algorithm.



(b) Result of Kmeans clustering algorithm.

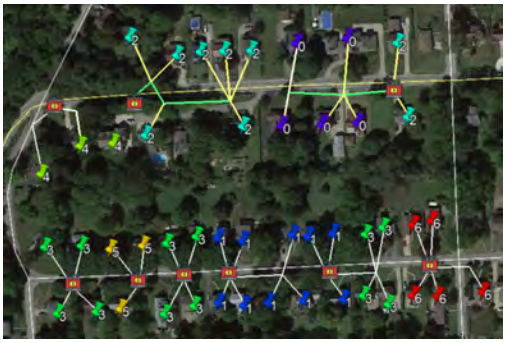


(c) Result of DBSCAN clustering algorithm with voltage mutual informations.

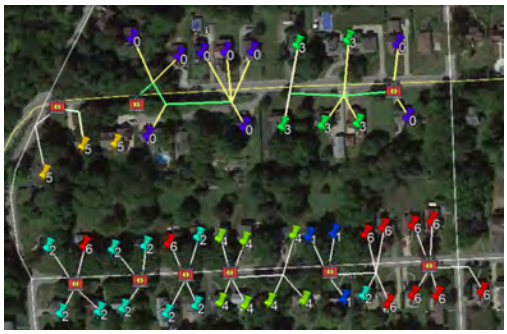
Figure 7: Comparison of the Three Clustering Algorithms Using Voltage and GIS Information on a IEEE-123 Bus Test Feeder.

3.4.2.2 Validation on Real Utility System

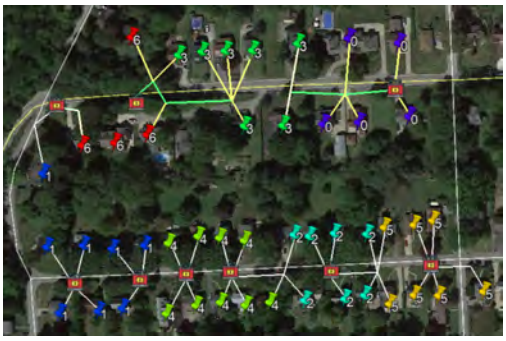
As our utility partner provides GIS information, we also use real GIS data to validation. For example, we show part of the reuslts for six different algorithms, with and without GIS information. For the first five subplots, either the voltage information was not used efficiently or the GIS informaiton was not included. We observe that our proposed method with results in Fig. 8f is the best among all combinations since it can merge the voltage mutual information and ground distance, while other methods including Kmeans and BIRCH cannot directly use the voltage mutual informations.



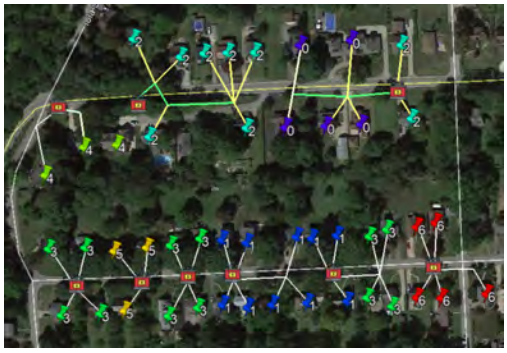
(a) Result of BIRCH Clustering Algorithm Using Voltage Information Only.



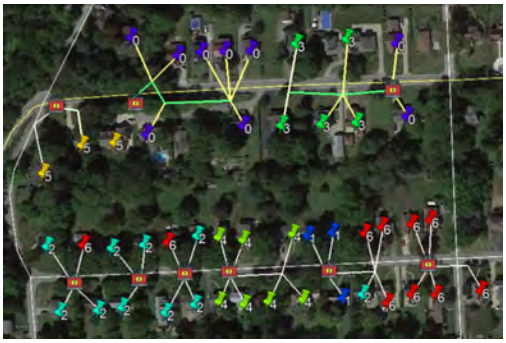
(b) Result of K-means clustering algorithm using Voltage information Only.



(c) Result of K-means Clustering Algorithm using GIS Information Only.



(d) Result of BIRCH Clustering Algorithm Using Voltage and GIS Information.



(e) Result of K-means clustering algorithm using Voltage and GIS information.

(f) Result of DBSCAN clustering algorithm using Voltage and GIS information.

Figure 8: Comparison of the Three Clustering Algorithms Using Voltage and GIS Information on a Sample in our Partner Utility.

3.4.3 Overall Accuracy

To evalute the accuracy, we conduct our algorithm throughout the utility territory in Fig. 8. For methodology, we compare our algorithm with respect to a mutual information method with Chow-Liu algorithm, the BIRCH method, and the k-means method. The results are displayed in Fig. 9, where the proposed method has an accuracy of near 95% over a large number of buses. The result is also quite robust, if the bus number continues to grow.

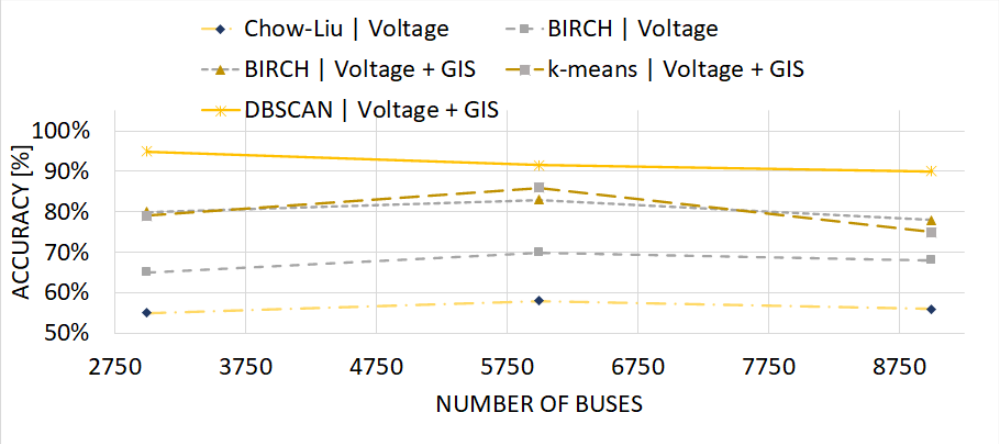


Figure 9: Accuracy of Various Algorithms for the Whole Utility Areas.

3.4.4 Field Deployment

To deploy our method, we need to assign addresses to poles to reverse-geocode each pole, which requires purchasing Google Maps API. Another method is to find the house nearest to a pole and assign its address to the pole. Since poles are near to the houses, the error in the latter method is small, which does not significantly affect the results. Once all poles have addresses, we can select the poles lying in the cluster if their nearest house lies in the cluster. With the parent transformer, poles and the smart meters belonging to the same cluster, we can use a minimum spanning tree to connect them to obtain the overhead secondary connections.

Minimum spanning tree works by connecting the houses with the poles and transformers by minimizing the total length of wire. Such an algorithm is correct as houses are usually supplied from their nearest poles, and also the distribution system has a tree structure.

For underground secondary distribution, there are no poles and we directly connect the transformers to the houses. This is correct as there are mostly separate wires from transformers to all customers so that a wire break affects a single smart meter. Moreover, it is simpler to replace individual wires rather than an underground tree structure of wires. It will be cumbersome for utilities to recover one street at a time; hence, we scaled the algorithm. For any area, large or small, the algorithm will automatically process the data and recover the topology. Fig. 10 shows all the recovered areas together. It is obtained by directly running the algorithm on the sets of data, without any human intervention.

With such a setup, Fig. 10 shows our deployment of our algorithm in the utility and we display the topology recovered based on our algorithm. This visualization shows that such an algorithm is suitable for large scale topology recovery.

Figure 10: Figure Shows Examples of the Topology Recovered for the Utility.

3.5 Conclusion

Electric utilities typically do not have an accurate distribution system topology readily available. With the advent of DERs, the electric utility faces challenges in the distribution grid. These challenges need greater visibility of their distribution system circuits through state estimation, where real-time topology recovery is the basis for modeling. Previous methods are based on either outdated maps or use voltage information only. This paper resolves this challenge by accurately clustering the topology. Specifically, we propose a density-based clustering method that leverages both voltage and geographical space data. And we show how to use GIS information with voltage information to refine the connectivity within one transformer. Finally, we not only show how to improve our method, but also provide an explainable theoretical bound. The proposed method is validated on the IEEE-123 bus system and the real system from our partner utility.

4.0 Data-Driven Machine Learning Algorithms To Identify DER Interconnections

With increasing renewable penetration in the distribution grids, distributed energy resources (DERs) are becoming a "trouble maker" for sub-transmission grid and distribution grid monitoring and control. For example, a different locational configuration of DERs within the power system can drastically impact generation and load forecast models, especially if utilities are unaware of the DER information (as in private home rooftop solar). Therefore this body of work is an essential aspect of being able to have more situational awareness and information about the assets that are connected to the grid. The work being proposed for this research topic is to utilize the system topology identification work and then layer on top the information of where and who has generation connected on the Distribution grid regularly. There are two use cases: 1) identify all customers that at one point installed solar arrays on their property and submitted an application however as time passed the solar arrays are no longer being used, or they are no longer connected to the property and 2) identify any solar connected customers that did not follow the utility process to register their solar arrays. Fig. 11 is an example of the utility space and what is their biggest concern to date which visually shows the importance of understanding which assets are in-service and can provide generation.

Currently, some utilities employ data for their service territory, developing a plan by reviewing each distribution circuit individually, as the system was designed and built for one-way power flows. The use of load curves and excel spreadsheets are some basic tool-sets that are utilized, and the planning process only incorporates 2-4 extreme load curves as a part of their planning process when developing their system re-enforcement plan to ensure capacity in the 5, 10-year forecast. An example approach is shown in Fig. 12 while such an incremental analysis, it does require tremendous time and cost to understand the system infrastructure one by one therefore, determining an alternative approach for system planning will be greatly beneficial.

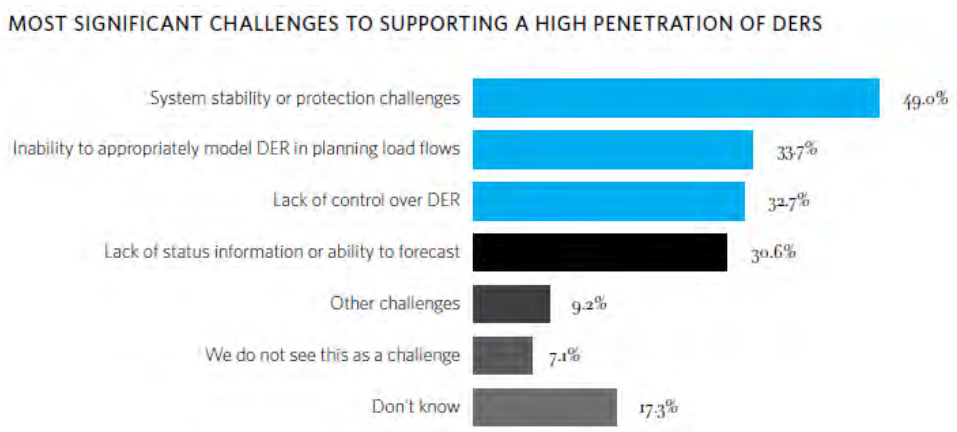


Figure 11: Most Significant Challenges to Supporting a High-Penetration of DER [10].

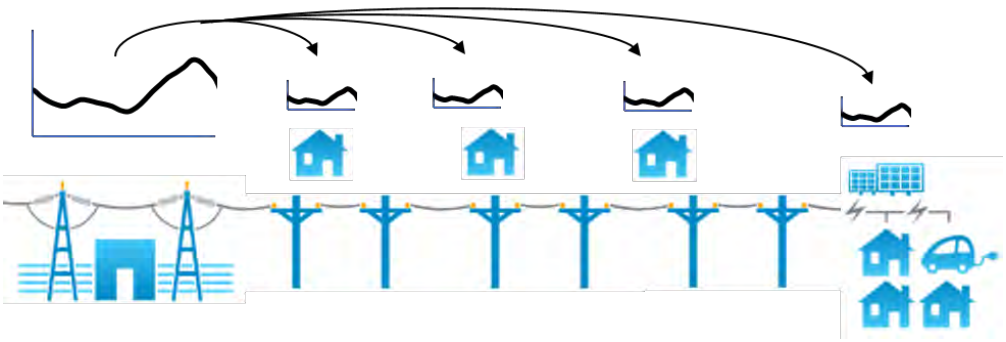


Figure 12: Illustration of Using a Top Down Approach for System Planning Studies.

In this second pillar of research topic, we consider the power of a large amount of data, where physical laws are inherently embedded. To start understanding the implications of DER a large amount of data will be leveraged to analyze past, current, and future models to make accurate decisions about future infrastructure needs by minimizing cost and maximizing DER benefits., this study will investigate the different ways that AMI data can be utilized to analyze the impacts of a generation device (e.g., residential solar panels), a load device (e.g., electric vehicle charging station), and then one type of device that can serve as a generator and load (e.g., battery). One hindrance utilities have is that not all customers

will or are required to report their new interconnected generation or load device. Therefore, it is crucial for the utility to be able to monitor customer usage trends to be able to identify already connected solar, EV charging stations or battery to better prepare for the future reliability of the whole distribution grid. With the increased amount of information available distribution planners will be able to build more statistically precise forecast shapes and perform more accurate load allocations. The next step in the planning process would be to able to use multiple system attributes in regards to the customer usage side and design the future plans using a bottom up aggregation as shown in Fig. 13.

Figure 13: Illustration of Using a Bottom up Aggregation for Future System Planning Studies.

With the increase in installations of residential photovoltaic (PV) systems, it is important for utilities to gain visibility of solar panels [13,14]. This is because the increasing PV systems not only create sustainable and green energy for human society but also build a new type of assets for utilities. To better evaluate the benefits and create new revenues, utilities need to know the locations of these new components to manage the unconventional two-way power flow for sustainable management of distribution grids. For example, detecting and monitoring all the active PV installations in a utility's territory allow the utility to perform accurate hosting capacity analysis (HCA). With HCA, utilities determine the amount of additional distributed energy resources (DERs) that can be "hosted" on the distribution system at a given location and at a given time, without threatening grid safety, reliability, or power quality [51].

Unfortunately, we do not know whether a customer has solar panels or not for sure. Some information can also vary as time passes by. Even worse, some of the solar panel

installations took place without permissions [28]. While a utility can form a team to manually update historical records on active solar locations, it is costly and hard to ensure the solar location are accurate all the time. Without knowing where solar panels are actively producing power, the system operation is prone to over-voltage, back-feeding through substations, or even damage system equipment such as transformers, voltage regulators, and customers' appliances. Therefore, utilities are in need of new methods for providing real-time renewable locations to better plan infrastructure and grid operation.

In the past, a lot of work relies on manual validation on locational information of PV for DER analysis [1, 8, 9, 14]. As manual checks are not scalable, there is work to automate the localization process. For example, [35, 38, 56] propose to use unmanned aerial vehicle (UAV) with different cameras, such as HD cameras, thermal cameras, and the infrared cameras. The goal is to localize different panels and their conditions for fault detection and maintenance. However, these methods typically work on solar farms and it is hard to send UAV across different utilities, which can be geographically large. Therefore, instead of taking photos, [19] and [4] propose to use satellite images to detect solar panels. However, satellite images include many areas without solar systems and there are similar objects that can be incorrectly identified as solar panels. Even worse, such satellite-based approach can not identify active solar users, as there are solar users, who discontinued the solar generation. Luckily, there are smart meter data available. So, [68] aims to detect the solar panels behind the meter data. It uses a change-point detection algorithm to screen out abnormal usage data. However, change-point detection can identify changes that are not due to solar behaviors.

One key drawback of change-point detection is due to its unsupervised nature and simplicity of using any change-point. While we demonstrate in this paper that supervised learning can achieve good performance, such learning requires adequate labels of the inputs and outputs [50]. This is insufficient because a utility may not be able to afford the cost and time for obtaining and maintaining a lot of the labels for solar and non-solar users [22, 47]. Therefore, we propose to use semi-supervised learning (SSL) by only requiring a small part of the labeled data from both classes [44, 69]. When the utility only has labels on one class, e.g., non-solar users, we propose to use one-class classification (OCC) [46, 48].

During the implementation, the direct application of SSL and OCC have relatively low accuracy, as the power system has a high dimensionality in data. For example, each user represents one point in the classification problem, but the user data is the result of vectorizing a long time-series data that can last several days for a clear pattern [20, 42, 53]. Besides, as residential customers have diversified user behaviors, the data of each class lives in a highly non-linear surface [44].

For resolving the issue of dimensionality, one can use principal component analysis, but it is a linear decomposition method [45]. Therefore, we propose to solve the issue of dimensionality and nonlinear representation together by designing new SLL and OCC methods based on autoencoders. Constructed by the two deep neural networks of an encoder and a decoder, an autoencoder is capable of providing a universal approximation of nonlinear and low dimensional space while de-noising [23, 30, 43, 57].

Finally, we use the known public and utility solar data arrays to validate the proposed methods. We use both accuracy and $F1$ score to measure the performance against baseline results. The baseline results were based off of common SSL and OCC methods as well as including common supervised learning methods. Such experiment shows enhanced solar usage detection when compared to the traditional methods. In summary, the contributions of the work are:

1. The work explains why solar detection is urgently needed and why the problem is hard and costly in reality based on our data mining of realistic utility data.
2. The work models the solar detection problem in supervised learning, semi-supervised learning (SSL), and one class-classification (OCC) setups. Future researchers can develop relevant tools based on our problem modeling.
3. The work proposes new SSL and OCC methods based on autoencoders, greatly boosting the power data representation and learning.
4. The work not only validates the methods based on the publicly available synthetic data set but also success with real utility data.

The rest of this section is structured as follows. Section II shows the feasibility of solar detection via data mining. Section III formulates the solar panel detection problem with limited labels. Section IV and Section V show the enhanced SSL and OCC via autoencoder. Section VI provides numerical results, and Section VII concludes the paper.

4.1 Differences between Solar + Non-solar Users

The problem of solar detection via utility data is not widely analyzed. One concern is that the solar users and the non-solar users are hard to tell from each other. For example, it is hard to tell whether solar exists behind a meter if the solar user has a relatively small solar generation when compared to the household usage.

4.1.1 Proof of Feasibility with Realistic Data

To validate this difficulty and provide the reason that differentiation is possible, we conduct data mining over realistic data from our partner utility with $600,000$ meters from a major U.S. city. The meter data range from June 1^{st}, 2019 to June 30^{th}, 2019 which have a one-hour interval between each reading was used for this exercise. We will show that the data is separable no matter how we sample the data, which will be our foundation for the next sections.

To obtain an easy visualization, we use the popular principal component analysis (PCA) tool to visualize the magnitude of eigenvalues of our data in Fig. 14. As the y-coordinate is by taking a log, we can see that only the first few eigenvectors matter and most of the eigenvectors are noises. To illustrate further, we map the data into 2-D and 3-D space in Fig. 15, where we can see that there is a boundary where two different behaviors are separable.

4.1.2 Proof of Feasibility with Synthetic Data

As the data is coming from one specific utility, we also conduct a constructive test to see how robust is this differentiation capability seen from realistic data. For this purpose, we test noise levels to mimic randomness in the residential customers and the environment. As the solar behavior is relatively stable, e.g., generation increases with sunrise and decreases with sunset, we compare two different signals by adding different levels of noises. Specific, we add noises to square waves to represent the electricity usage of the customers who never report their installation of solar panels and sinusoidal waves to represent those with solar panels. We then add different noises to the signal to approximate different users.

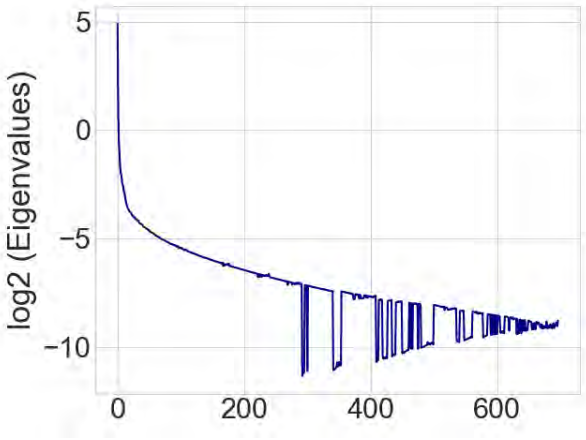


Figure 14: Results from a Popular Principal Component Analysis Tool to Visualize the Magnitude our Data's Eigenvalues Magnitudes.

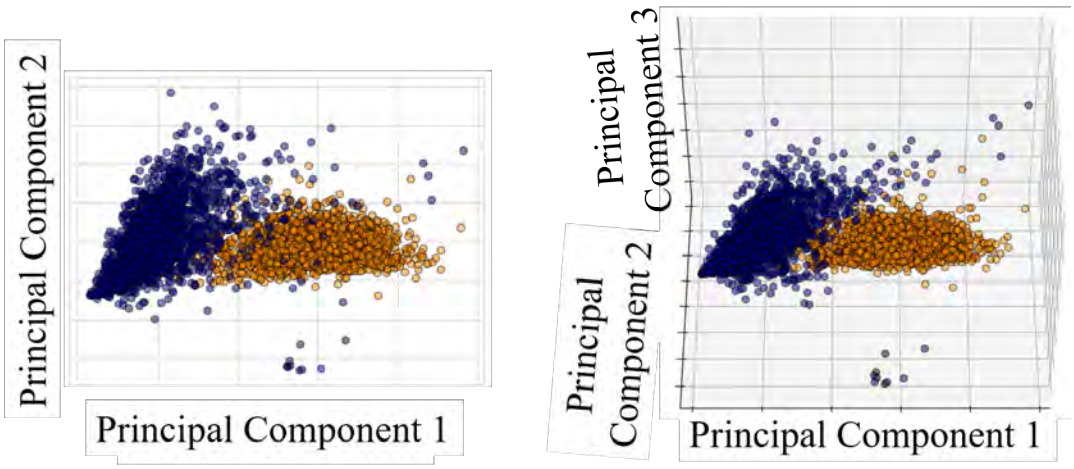


Figure 15: Visualizations of the Principal Components Showing a Boundary Between the Two Different Behaviors Allowing the Data to be Separable.

The synthetic data will be directly fed into typical classifiers such as support vector machine (SVM) and logistic regression to see if accuracy can be preserved with different noise levels.

Fig. 16 presents an example of the data set with different level of noises, the noise level increases from top to bottom. Although it becomes more difficult for us to determine the class of the data, Table 2 shows the classification results is still high when the noise level is much higher than signal level.

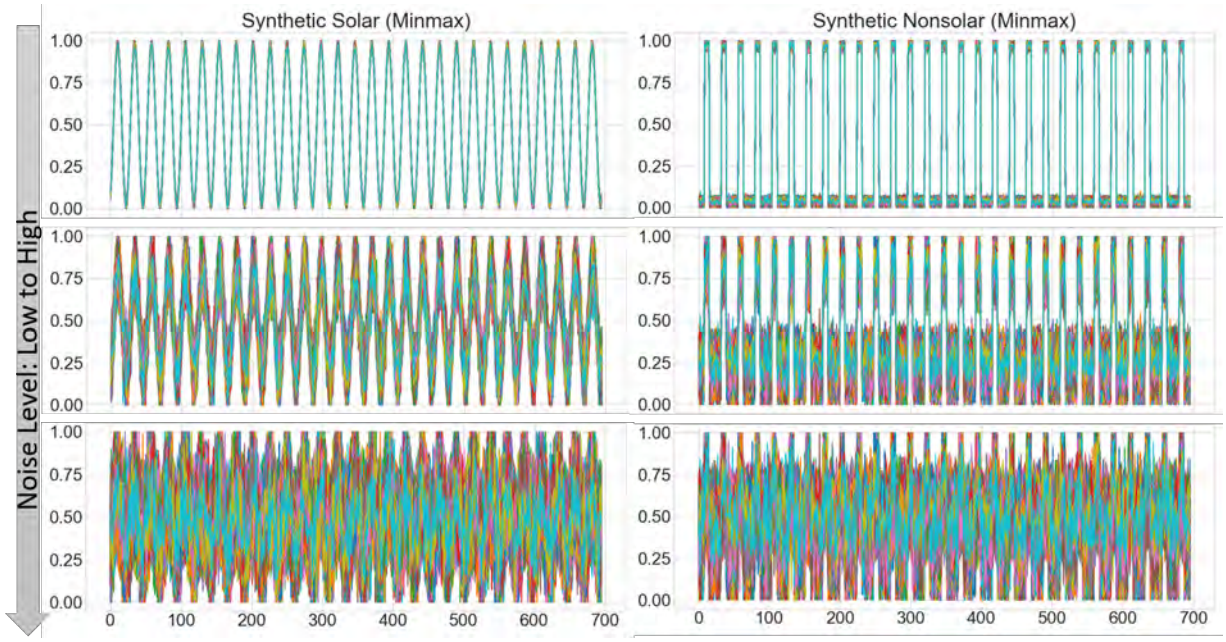


Figure 16: Illustration of an Example of the Data Set with Different Level of Noises to Approximate Different Users, the Noise Level Increases from Top to Bottom.

Table 2: Classification Accuracy (acc) of Different Noise Levels, Which is Normalized with Signal Level.

Classification	SVM (Linear Kernel)	Logistic Regression	Noise Level
Training/Test Acc	100%	100%	$\mathcal{N}(0, 1.0)$
Training/Test Acc	100%	100%	$\mathcal{N}(0, 4.0)$
Training/Test Acc	100%	100%	$\mathcal{N}(0, 7.0)$
Training/Test Acc	100%, 99%	100%, 99.33%	$\mathcal{N}(0, 12.0)$

4.2 Problem Definition

The last section shows the feasibility based on rough visualization and supervised learning of abundant but synthetic data. However, the reality at utilities is that the knowledge of highly accurate labels, solar users and non-solar users can be quite limited. In some utilities, they may have only one class of labels and do not have the time and money to manually label more. Therefore, we define the following two problems based on the scarcity of labels in a data set.

4.2.1 Semi-Supervised Learning (SSL) Problem

- Problem: Solar panel detection via SSL
- Given:
  - Labeled electricity usage data: $(X_m, \mathbf{y}_m) = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$, where m is the number of meter data which have labels showing whether the customer has solar panels or not.
  - Unlabeled electricity usage data: $X_n = \{\mathbf{x}_j\}_{j=m+1}^{m+n}$, where n is the number of meter data which do not have labels, usually $n \gg m$.
- Goal:
  - Find the optimal mapping rule of f_{SLL} so that $\hat{\mathbf{y}}_{SLL} = f^*_{SLL}(\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}, \{\mathbf{x}_j\}_{j=m+1}^{m+n})$.

4.2.2 One-Class Classification (OCC) Problem

- Problem: Solar panel detection via OCC
- Given:
  - Electricity usage data: $X_p = \{\mathbf{x}_i\}_{i=1}^{p}$ which have indicators $\mathbf{y}_p = \{y_i\}_{i=1}^{p} = +\mathbf{1}$ that they belong to the same class, where p is the number of meter data, $\mathbf{1}$ is a vector whose elements are all equal to 1.
  - Electricity usage data: $X_q = \{\mathbf{x}_i\}_{i=p+1}^{p+q}$ which have indicators $\mathbf{y}_q = \{y_i\}_{i=p+1}^{p+q} = -\mathbf{1}$ referring to all the other unknown classes, where q is the number of meter data, $\mathbf{1}$ is a vector whose elements are all equal to 1.
- Goal:
  - Find the optimal mapping rule of f_{OCC} so that $\hat{\mathbf{y}}_{OCC} = f^*_{OCC}(\{(\mathbf{x}_i, y_i)\}_{i=1}^{p}, \{\mathbf{x}_i\}_{i=p+1}^{p+q})$.

4.3 Deep Semi-Supervised Learning

One of the major issues of directly using SSL method from computer science domain is due to the high dimensionality of power data and the need of nonlinear representation. Therefore, we propose to integrate the autoencoder (AE) into the proposed deep SSL method, where we show the expectation-maximization (EM) algorithm below so that we can properly illustrate the AE part afterwards.

4.3.1 Conventional Semi-Supervised Learning Method

EM algorithm relies on mixture models and is a popular way to solve SSL problems and the methods have lots of successful applications in different fields, such as image processing and data classification tasks [33, 37, 54]. As defined in Section 4.2.1, Equation (1) denotes the electricity usage data and their correlated labels. Equation (2) denotes electricity usage data without labels.

$$(X_m, Y_m) = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m} \tag{1}$$

$$X_n = \{\mathbf{x}_j\}_{j=m+1}^{m+n} \tag{2}$$

We let the labels only take binary values (0 or 1), labels with a value of 0 represent the customers who do not have solar panels and labels with a value of 1 represent the customers who have solar panels. Based on this setting, we assume we know the labels $\hat{\mathbf{y}}_{SLL} = \{y_j\}_{j=m+1}^{m+n}$, we are able to compute the likelihood of all the data with respect to the underlying parameters Θ, to be shown in Equation (3).

$$P(X_m, \mathbf{y}_m, X_n, \hat{\mathbf{y}}_{SLL}|\Theta) = \prod_{i=1}^{m} P(\mathbf{x}_i, y_i|\Theta) \prod_{j=m+1}^{m+n} P(\mathbf{x}_j, y_j|\Theta) \tag{3}$$

The EM algorithm iteratively fixes the value of Θ and $\hat{\mathbf{y}}_{SLL}$ to find a suboptimal solution of the maximization of the log-likelihood function over all the data. Specifically, for the t^{th} iteration and in the expectation (E) step, Θ^t is fixed and the EM algorithm optimizes a lower

bound given by the expected log-likelihood $Q(\Theta|\Theta^t)$ in Equation (4).

$$Q(\Theta|\Theta^t) = \mathbf{E}_{\hat{\mathbf{y}}_{SLL}|X_m,\mathbf{y}_m,X_n,\Theta^t}[\log P(X_m, \mathbf{y}_m, X_n, \hat{\mathbf{y}}_{SLL}|\Theta)] \tag{4}$$

In the maximization (M) step, the algorithm maximizes $Q(\Theta|\Theta^t)$ with respect to Θ given in Equation (5). Although the parameters Θ may be highly correlated, the above procedure faces high computational cost as Θ has high dimensionalities [20].

$$\Theta^{(t+1)} = arg\max_{\Theta} Q(\Theta|\Theta^t) \tag{5}$$

4.3.2 Autoencoder (AE) in a SSL Setup

The electricity usage data in the high dimensional space not only have a lot of noises, but also have a highly nonlinear user behavior, therefore we propose to use AE. An AE constitutes an encoder that compresses the original data to a code and then a decoder which reconstructs the data from the code, as shown in Fig. 17. The encoder can be used to reduce the dimension of the data, help the similarity calculation, and extract the most representative information. There is a variety of AEs proposed in previous researches, such as sparse AE [43], denoising AE [57], variational AE [30], and long-short term memory AE [23].

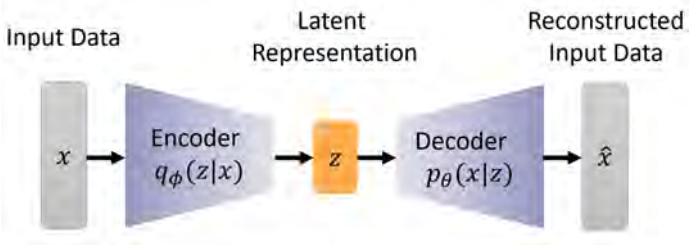


Figure 17: Block Diagram of an AE which Constitutes an Encoder that Compresses the Original Data to a Code and then a Decoder Which Reconstructs the Data from the Code.

To explain the general concept of the AE, we will not add the subscripts of the input and output data that are specific for our problem setup. An AE tries to minimize the error

between the input data $\mathbf{x}$ and the reconstructed output $\hat{\mathbf{x}}$. The reconstruction loss is often determined by the square error and is defined in Equation (6).

$$\begin{aligned} \mathbf{z} &= f_e(W_e\mathbf{x} + \mathbf{b}_e), \\ L(\mathbf{x}, \hat{\mathbf{x}}) &= \|\mathbf{x} - \hat{\mathbf{x}}\|^2 = \|\mathbf{x} - f_d(W_d\mathbf{z} + \mathbf{b}_d)\|^2, \end{aligned} \tag{6}$$

where W_e is the weight matrix between the input vector $\mathbf{x}$ and the latent representation vector $\mathbf{z}$, W_d is the weights matrix between the hidden representation vector $\mathbf{z}$ and output vector $\hat{\mathbf{x}}$. f_e and f_d are the activation functions, $\mathbf{b}_e$ is the bias vector of the encoder, and $\mathbf{b}_d$ is the bias vector of the decoder.

To combine AE with EM, we propose to first input data into an AE to create the latent representation $\mathbf{z}$, shown in Fig. 18. Then, $\mathbf{z}$ with its associated labels $\mathbf{y}$ is fed into a Gaussian mixture model for EM. When EM iteratively finds the solution of maximizing the log-likelihood function, the label of the unlabeled data comes out naturally. The complete structure is shown in Fig. 19.

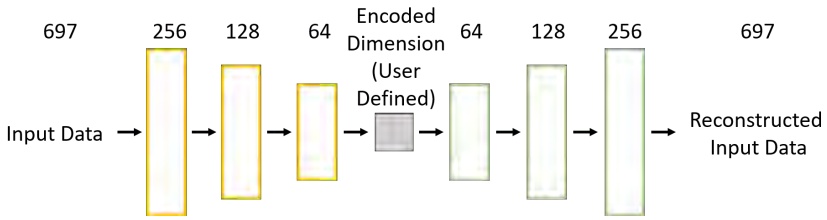


Figure 18: An Example of AE for Power Data.

4.3.3 Steps of the Proposed Algorithm

Let the representation $Z_m = \{\mathbf{z_i}\}_{i=1}^{m}$ coming from the AE be the hidden representations of the labeled data whose labels are $\mathbf{y}_m = \{y_i\}_{i=1}^{m}$. Let the representation $Z_n = \{\mathbf{z_j}\}_{j=m+1}^{m+n}$ coming from the AE be the hidden representations of the unlabeled data whose estimated labels are $\hat{\mathbf{y}}_{SSL} = \{y_j\}_{j=m+1}^{m+n}$. We will assume that labels can only take binary values (0 or 1).

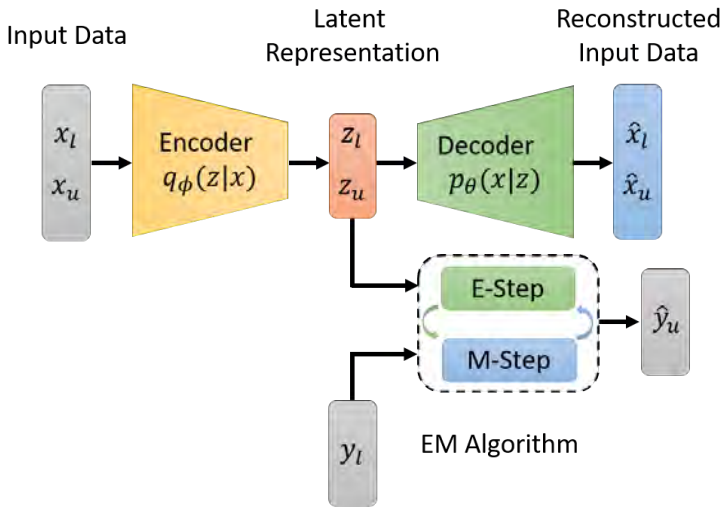


Figure 19: Block Diagram of the Proposed Deep Semi-Supervised EM Approach.

Based on this setting, assume we know the labels $\hat{\mathbf{y}}_{SSL}$, we are able to compute the likelihood of the whole data set with respect to the underlying parameters Θ given in Equation (7).

$$P(Z_m, \mathbf{y_m}, Z_n, \hat{\mathbf{y}}_{SLL}|\Theta) = \prod_{i=1}^{m} P(\mathbf{z}_i, y_i|\Theta) \prod_{j=m+1}^{m+n} P(\mathbf{z}_j, y_j|\Theta) \tag{7}$$

For the t^{th} iteration and in the expectation (E) step, Θ^t is fixed and the EM algorithm optimizes a lower bound given by the expected log-likelihood given in Equation (8). In the maximization (M) step, the algorithm maximizes $Q(\Theta|\Theta^t)$ with respect to Θ.

$$\begin{aligned} Q(\Theta|\Theta^t) &= \mathbf{E}_{\hat{\mathbf{y}}_{SLL}|Z_m,\mathbf{y_m},Z_n,\Theta^t}[\log P(Z_m, \mathbf{y_m}, Z_n, \hat{\mathbf{y}}_{SLL}|\Theta)] \\ &= \sum_{\hat{\mathbf{y}}_{SLL}} P(\hat{\mathbf{y}}_{SLL}|Z_m, \mathbf{y_m}, Z_n, \Theta^t) \log P(Z_m, \mathbf{y_m}, Z_n, \hat{\mathbf{y}}_{SLL}|\Theta) \\ &= \sum_{i=1}^{m} \log P(y_i, \mathbf{z}_i|\Theta) + \sum_{j=m+1}^{m+n} \sum_{y_j \in \{0,1\}} P(y_j|\mathbf{z}_j, \Theta^t) \log P(y_j, \mathbf{z}_j|\Theta) \\ &= \sum_{i=1}^{m} \log P(y_i, \mathbf{z}_i|\Theta) + \sum_{j=m+1}^{m+n} \sum_{y_j \in \{0,1\}} r_{y_j}^j \log P(y_j, \mathbf{z}_j|\Theta) \end{aligned} \tag{8}$$

In the last line of the equation, we define $r_0^j = P(y_j = 0|\mathbf{z}_j, \Theta^t)$, $r_1^j = P(y_j = 1|\mathbf{z}_j, \Theta^t)$, which are our current estimates for the probabilities of each of the labels in the unlabeled

examples. Therefore, in the E step, we compute probabilities r_0^j and r_1^j for all the unlabeled data based on the current Θ^t. In the M step, we maximize the expected log-likelihood (the last term of Equation (8)) for all the data.

4.4 Deep One-Class Classification

When the labeled data are so limited at a utility that only one class of the labels can be obtained, e.g., only the labels of some non-solar users. In such a case, it is impossible to create a classification boundary between two classes like SSL.

4.4.1 Conventional One-Class Classification (OCC) Method

Therefore, one-class classification aims to regularize the descriptive loss, popular in supervised learning and SSL, with an additional loss on compactness. The idea is to evaluate the compactness of data with known labels and with nearby data to form a group, while looking for distinct boundaries that can separate the data into two or more groups. For example, support vector data description (SVDD) is one of the OCC solvers. SVDD tries to define the compactness of the targeted class by constructing a hypersphere, wrapped in a compactness matrix. SVDD defines a hypersphere with center $\mathbf{c}$ and radius $r > 0$ which gathers as many observations from one class as possible in the feature space with the help of the kernel function ϕ_k [55]. The radius measures the compactness of the data, the smaller the radius, the more compact the data. The primal problem of SVDD is defined in Equation (9).

$$\begin{aligned} \min_{r,\mathbf{c},\zeta_i} \quad & r^2 + \frac{1}{\nu n}\sum_i \xi_i \\ \text{s.t.} \quad & \|\phi_k(\mathbf{x}_i) - \mathbf{c}\|^2 \leq r^2 + \xi_i,\ \xi_i \geq 0, \quad \forall i, \end{aligned} \tag{9}$$

where the slack variable ξ_i is introduced to allow a soft margin and the regularization parameter ν controls the relative importance of the volume of the sphere and the penalties ξ_i.

The descriptiveness of the data are maintained in the constraints. Solving the minimization problem given in Equation (9) by using Lagrange multipliers, we can derive that the center $\mathbf{c}$ of the sphere should be a linear combination of some important input data. These input data have a significant influence on the construction of the sphere by describing the boundary of the sphere and are called support vectors.

4.4.2 Proposed Deep OCC Method

SVDD often has poor computational efficiency and scalability due to the structure and manipulation of the matrices and SVDD is prone to failure when the data set is extremely large and the dimension of the data is extremely high. Thus, substantial feature engineering is needed [49]. This makes power data in high dimension hard to capture diversified nonlinear user behavior and remove noise.

Therefore, we propose to use the hidden layers of autoencoder (AE) to extract the nonlinear features for one-class classification. For example, Fig. 20 provides a visual representation of AE's ability of representing highly nonlinear customer data in low dimensional space for our utility data set. The top left figure is the non-solar data plotted in a 3-D plane, whereas the bottom left figure is the solar data.

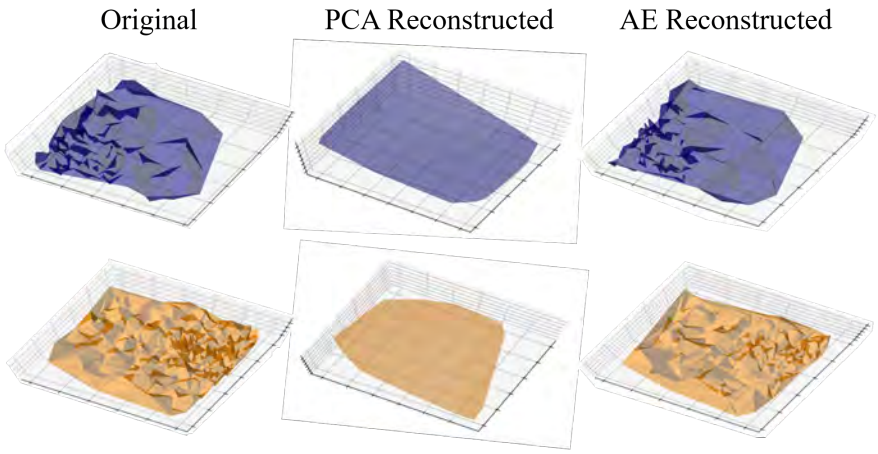


Figure 20: Illustration Comparing PCA Reconstruction versus an Autoencoder for Non-solar (blue) and Solar (orange) Data Set.

The two middle plots top and bottom is the representation of the solar and non-solar data, respectively reconstructed using principal component analysis (PCA). And the right top and bottom figure is the data set reconstructed using an AE and plotted in a 3-D plane. The figure is to provide clarity on the ability of the AE to retain the information more accurately than the PCA. As shown the PCA is not able to reconstruct the data as well as the autoencoder (AE) therefore providing evidence of the high accuracy and advantage to

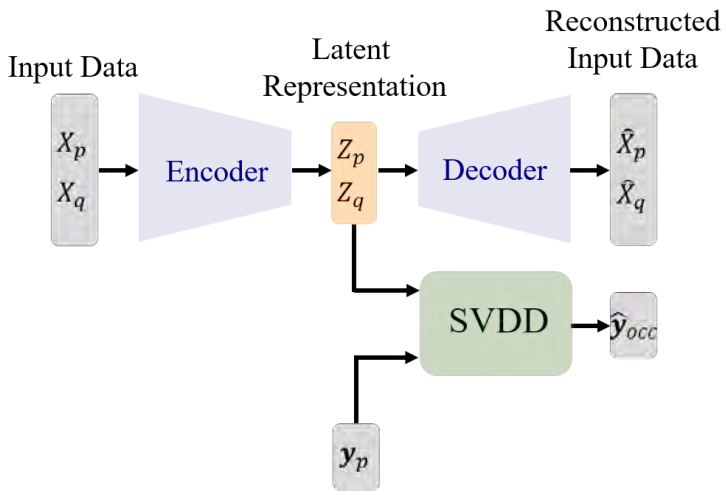


Figure 21: Block Diagram of the Proposed Deep SVDD Approach.

using an AE over a PCA to reconstruct the high-dimensional data for purposes of identifying between solar and non-solar data. The AE can map the original data to a denser area which helps to construct the compactness description of the targeted class. This enhances the design of the OCC. Hence, why the AE will be used in the design for the new proposed method.

The architecture is shown in Fig. 21, where the extracted learned hidden features Z_p for labeled data and Z_q for unlabeled data are fed into the SVDD. Combining the extracted learned hidden features with their labels, the SVDD is able to determine the labels of the unlabeled data. The objective of the problem is to solve Equation (9) after replacing $\mathbf{x}_i$ with $\mathbf{z}_i$.

4.5 Numerical Validation

With the proposed methods in the last two sections, we will validate the performance in this section. The algorithms used are the deep semi-supervised expectation-maximization (Deep-EM) algorithm and deep support vector data description (Deep-SVDD) of this paper. We use both public data sets and the utility data sets to conduct our experiments with traditional common semi-supervised learning and one-class classification algorithms. Principal component analysis is also used when necessary for consistency. As a baseline to our result, we also include the results of supervised learning in our experiments with accurate labeled data sets.

4.5.1 Data Preparation

The public *UMass Smart** data set [52] used in this study contains everyday electricity load profiles, extracted from dataset named "Apartment dataset", from 114 single-family apartments in June 1^{st}, 2015 and to June 30^{th}, 2015 with 15-minute interval between each pair of readings. We take the average of the data to scale the original data to one-hour interval. Therefore, the total number of time indices used in the study is 696, corresponding 29 days. The solar generation data, comes from another dataset named "Solar panel dataset" in the same public data repository, which documents the solar generation data for 50 rooftop solar panels with one-minute interval between each pair of readings. We select 39 solar generation profiles as the other profiles had bad data such as near zero values. Then, we combine them with the aforementioned 114 load profiles to create the electricity usage of the solar users. To mimic the unbalanced data set, we add a number of different noises to the 114 load profiles to create the profiles for non-solar customers for diversity, when compared to 39 solar customers. For example, as the results are similar, we show the case when we add four different noises to the 114 load profiles, leading to 456 non-solar profiles.

The utility data set used in this study corresponds to a set of everyday electricity usage readings from around $600,000$ meters from a U.S. city from June 1^{st}, 2019 to June 30^{th}, 2019 with a one-hour interval between each reading. The total number of time indices used

in the study is 696, corresponding 29 days. Around $1,973$ customers have installed solar panels. Their smart meter readings come from the net meters, which record the household electricity consumption and the PV generation as a whole. The rest of the approximately $598,000$ customers we assume never reported their installations of the solar panels therefore label them as non-solar, we then randomly select $20,000$ from this data set to conduct this study.

To eliminate the influence of different scales of the data, we use min-max normalization methods to scale the data between 0 and 1 throughout the paper.

4.5.2 Performance Metrics

To evaluate binary classification several statistical rates are available to measure performance (i.e., accuracy, $F1$, recall, or precision). For this work we use the accuracy and $F1$ score as our performance measurements. Accuracy is used when the true positives (TP) and true negatives (TN) are important and the data set's class distribution is similar. $F1$ score is used when the False Negatives (FN) and False Positives (FP) are critical and the data set is unbalanced. These metrics are defined as follows:

$$\begin{aligned} &Accuracy = \frac{TP + TN}{TP + FN + TN + FP}, \\ &Precision = \frac{TP}{TP + FP}, \ Recall = \frac{TP}{TP + FN}, \\ &F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}. \end{aligned} \tag{10}$$

We use the $F1$ score, since our data set will most likely have an imbalanced class which will take the precision and recall rate into account which cares for both the majority class and the minority class [40]. We will also use the accuracy metric even though accuracy can be biased when the amount of members in different classes are unbalanced, e.g., the non-solar users are much more than solar users for some utilities. For example, if non-solar users are 99 times more than the solar, a naive algorithm to achieve high accuracy of 99% is simply to label all the customers in the testing set with a non-solar label. We include the accuracy performance metric to observe considering that the synthetic data may not always be imbalanced and therefore should be available to observe any differences.

4.5.3 Baseline of Supervised Learning for Deep SLL and OCC

As a reference for SLL and OCC, we conduct simulations for different supervised learning methods [7, 15, 32, 36]. As the results are similar, we show the the result of support vector machine (SVM) and logistic regression (LR) in Fig. 22. The figure shows that when the provided information is little and the data set is unbalanced, the supervised learning method tends to overfit the data and thus results in bad $F1$ score. A relative high projection dimension helps to improve accuracy and $F1$ score and indicates that more supervision and more information ensure better results. Finally, Fig. 22 also shows that result of the public data set and the utility data set are similar, which is also the case for SSL and OCC. So, we will focus on one dataset for the rest of the visualization work.

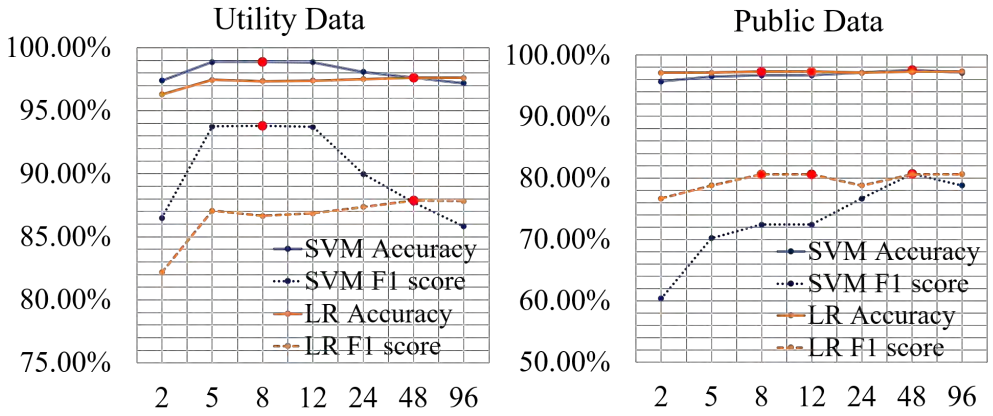


Figure 22: The Supervised Learning Results of the Public Data Set and the Utility Data Set.

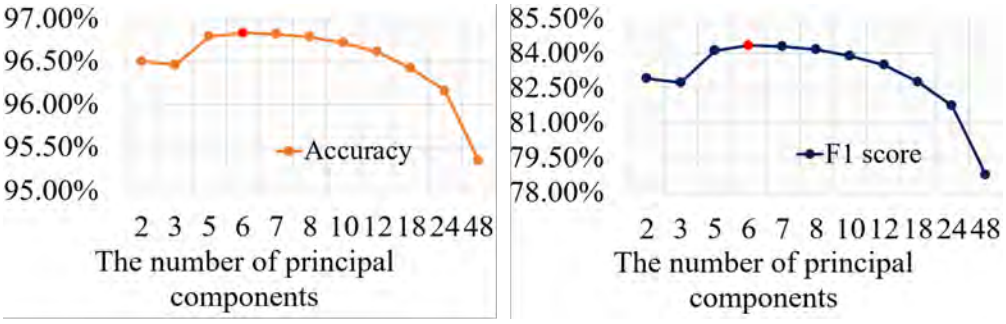
4.5.4 Feature Numbers for Linear and Nonlinear Representation

To understand how many features are needed in nonlinear representation learning of autoencoder, we plot the results in terms of the two performance metrics in Fig. 23, where we also show results of linear representation of PCA for comparison. In the sub-figures, We try to ensure the consistency in the setups for all the learning processes. For the deep semi-supervised learning (SSL) method, we choose to use the first 50 solar data and first

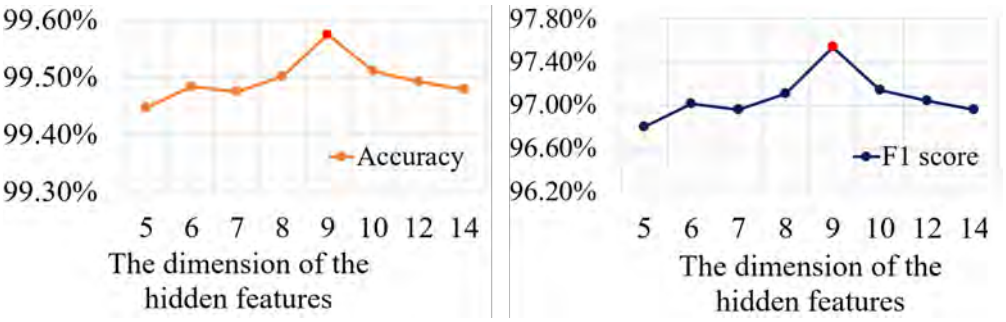
50 non-solar data as the labeled data, all the other 1923 solar data and 19,950 non-solar data as the unlabeled data. The proposed deep SSL method takes all the labeled data and the unlabeled data and infers the labels for the unlabeled data. For the deep one-class classification (OCC) method, we keep the same setup by using the first 50 non-solar data as the given class and all the rest data being unknown classes. The proposed deep OCC method interpret the labels for the rest of the 21,923 data based on the 50 non-solar data.

For the deep SSL method, as can be seen from Fig. 23a, when we increase the dimension of the projected principal components, the $F1$ score and the accuracy increase with a little fluctuation and reach the optimal and finally decrease. The optimal value is reached when we choose 6 projected components. Also shown in Fig. 23b, when we increase the dimension of the hidden representations we extracted, the $F1$ score and the accuracy reach the optimal with a little fluctuation and finally decrease. The optimal value is reached when a 9-dimensional hidden representation is used. Comparing to the results of using principal component analysis (PCA), the results of the autoencoder (AE) are much better, especially for the $F1$ score, which has a more than 10% increase. This also shows that although the accuracy of using PCA and AE is always above 95%, the true performance for the classification for the minor class may not be as good as it seems and $F1$ score successfully distinct the performance.

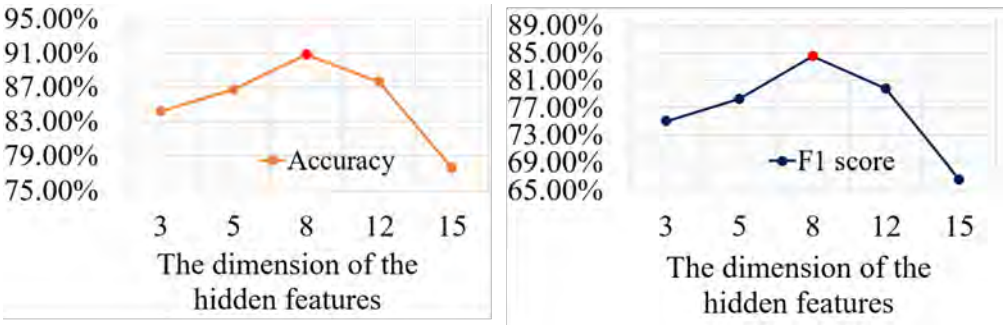
For the deep OCC method, as shown in Fig. 23c, the accuracy and the $F1$ score first increase to the peak and then decrease. The optimal value is reached when an 8-dimensional hidden representation is used. The deep OCC has a reasonable performance reduction in both accuracy and $F1$ score, it's acceptable because less information is provided. All aforementioned results indicate that a relatively low dimension is enough for learning. Higher-dimensional components may contain information that is harmful to the results, i.e., noises and bad data, so the results guide us to experiment on a dimension between 5 to 12 as the representations of the original data. The results also indicate that as PCA is a linear transformation of the input space aiming to find the directions that have higher variances, the projected data have low or close to zero correlation with each other. However, the electricity usage data used in our simulation are highly nonlinear and the features which are different timestamps are correlated with each other. The AE has the advantages of capturing the



(a) The Data after PCA to the Semi-Supervised EM Algorithm. The Red Dot shows the Optimal $F1$ Score and Accuracy is Reach when we Choose 6 Projected Components



(b) The Hidden Representations to the Semi-Supervised EM Algorithm. The Red Dot Shows the Optimal $F1$ Score and Accuracy is Reached when we Choose 9 Latent Representations.



(c) The Hidden Representations to the SVDD. The Red Dot Shows the Optimal $F1$ Score and Accuracy is Reach when we Choose 8 latent representations.

Figure 23: The Optimal Dimension for Each Method.

complex relationship between the features and the nonlinearities of the data. The AE has such abilities due to the introduction of the nonlinear activation function and by propagating the gradient, the AE automatically learns the parameters.

4.5.5 Performance Improvements for Deep SSL and Deep OOC

To better visualize the performance boost of the proposed methods, we plot all the results together in Fig. 24. These results include supervised learning, SSL, OCC, with and without autoencoder components. The left graph illustrates the comparison of accuracy where the right graph is the $F1$ score. The dashed green line shows the performance of the supervised learning method based on support vector machine with radial basis functional (RBF) kernel. The dashed orange and navy line are the results of the classic SSL and classic OCC methods when using the projected data based on principal component analysis (PCA), respectively. The solid orange and navy line are the performance of the proposed deep SSL and deep OCC methods when using the hidden representations extracted from the autoencoder (AE), respectively.

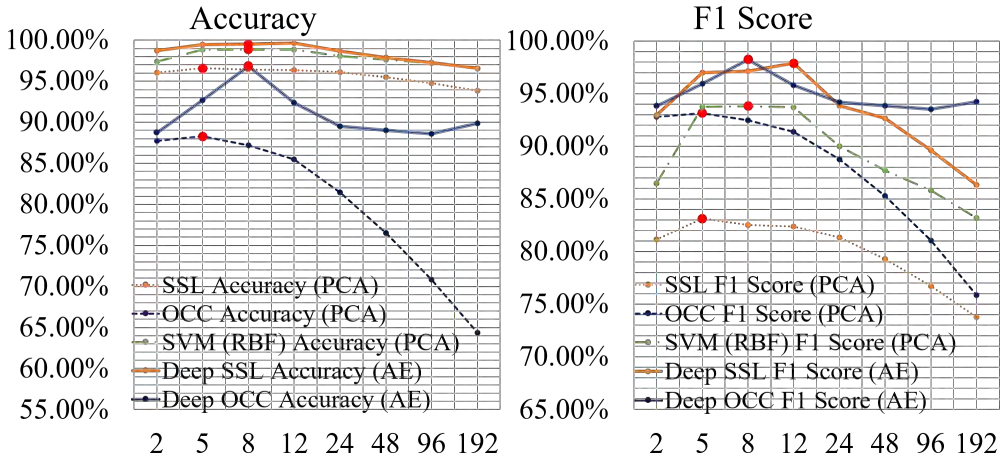


Figure 24: Illustration Providing the Comparison Between the Accuracy and $F1$ Score of the Study Results Between the Baseline Supervised Learning, the Proposed Deep SSL and Deep OCC Methods Utilizing the Projected Data of the PCA and the Hidden Representation Extracted from the AE.

For the three dashed lines, we can observe that the accuracy of supervised learning is always higher than the accuracy than the SSL and the accuracy of SSL is always higher than OCC, if we use the projected data after PCA. We also obtain a similar conclusion for the $F1$ score by ignoring the projection to 2 principal components. The results confirms that more information guarantees better performance.

Next, we focus on the performance of the proposed deep SSL method, which is shown by the orange dash line and the orange solid line in the figure. The performance curves first increase and then decrease as we increase the dimensionality of the projected data, either from PCA or AE. We conclude that a relative low dimension, from 5 to 12 is enough to summarize the characteristics of electricity usage. The figure also shows that the accuracy has a clear improvement with the help of the AE and the $F1$ score has a huge boost of more than 10%. The result indicates that supervised learning tends to overfit the data when given limited information. The unlabeled data helps to improve the performance by providing more complete information on the distribution of the data.

Finally, we look at the performance of the regular OCC and the proposed deep OCC, which are shown by the navy dash line and the navy solid line in the figure. The performance curves first increase and then decrease as we increase the dimensionality of the projected data, either from PCA or AE. While the performance of using the projected principal components has a sharp decline when the dimensionality of the projected data increases, the performance of using the hidden representations from the AE remains stable. This indicates that the nonlinear transformation of the AE guarantees the OCC method to find a good hypersphere in regardless of the dimensionality. The performance of the proposed OCC is slightly worse than the supervised learning in terms of the accuracy, which is acceptable as the provided information is much less.

Overall, the proposed methods with the assistance of the AE show to provide greater accuracy and $F1$ scores than the supervised learning and merely using the principal components from PCA.

4.5.6 Computational Time

Table 3: The Average Computation Time for All the Methods.

Method	Supervised learning	SSL (PCA)	SSL (AE)	OCC (PCA)	OCC (AE)
Average computation time	0.3 s	354.0 s	215.6 s	716.6 s	878.3 s

Table 3 shows the average computation time for each method based on a CPU *Intel(R) Xeon(R) CPU E5-2687W v4 @ 3.00GHz* and 64 *GB* memory. As can be observed from the table, the performance boosts come with a cost of computation time. For example, the AE accelerates the speed of the SSL method but slows down the speed of the OCC method, which is possibly because the AE maps the data to $[-1, 1]$ and saves the computation cost. However, the OCC method has to compute the relative distance of the data, so the AE cannot do much to the computation time. However, the analysis of this work is offline, so the needed computational time is feasible.

In summary, the solar detection is urgently needed and is hard and costly to maintain accurate utility databases with current methods. Electric Distribution Companies need to have visibility of these assets to avoid potential risks of two-way power flow, e.g., outages and equipment damages. In this paper, we proposed a deep semi-supervised learning and a deep one-class classification approach to detect residential PV systems under different scenarios. The proposed methods use the extracted features from the autoencoder and combine them with the original label information to predict the labels for the rest of the data. The proposed methods have been validated on a utility data set and a publicly available data set and have shown their effectiveness and robustness towards the solar panel detection problem.

5.0 Comparative Analysis of System Planning Studies

Once the system topology and DER interconnections are known, the next step is to understand the varying impacts additional DER will have on the current infrastructure. We will examine the feasibility and effectiveness of the current industrial methods for determining feeder upgrades. With the dramatic increase of the residential photo-voltaic (PV) systems, EV charging stations, and/or any distributed generation, utilities need to know the locations of these new components to manage the unconventional two-way power flow for sustainable management of distribution grids. And as stated above, historical records are not up-to-date. It is costly to repeatedly check active DER locations; therefore, this work is pivotal for utilities and other third parties.

Consequently, we will provide insight into the impacts of using the new data-driven topology recovery and awareness methods that were developed to ensure the holistic system perspectives of the granular load is modeled as well as the DER location data. The result of the study can then refine the method of upgrading the current system, delay the unnecessary upgrade, and expedite urgent upgrade, significantly saving utilities' budgets.

5.1 Overview of CYME Power Engineering Software and Benefits to Utilizing AMI Data

With the additional accessibility of the AMI data, the full system model can be verified and confirmed more readily with access to the secondary grid's system loads (i.e., AMI operational data). Refer to Fig. 25 for an overview of the circuit model within the engineering software. Refer to Fig. 26 to visualize the individual load model components that need to be adjusted to ensure the accuracy of real-time scenarios. Each individual triangle represents a system's pole-top transformer. With the work from the topology reconstruction, the meter-to-meter-to-transformer data is available for modeling purposes providing the inputs to define the model with more granularity and overall accuracy. Refer to Fig. 27, which is a screenshot of the CYME software providing the user the opportunity to either adjust the load curves by a global scenario or to input a data file with all of the associated transformers identified load cycle to capture the full range of the customers' usage. Lastly, refer to Fig. 28 providing additional granular access to the data to ensure that the model has the most accurate inputs that can be taken from the AMI usage data once the meter-to-meter-to-transformer relationship is known from utilizing the AMI voltage data. This work is important to determine the results on the ability of not just the main feeder to handle a certain threshold of DER but each individual lateral and customer transformer.

5.1.1 Circuit Model Development

To begin any study, a model must be developed in a platform where multiply scenarios and contingencies can be processed and ran. One software that is being utilized in the industry is Eaton's CYME power engineering software solutions. The connectivity of a circuit model was created from the utility data provided, such as the transformer rating, fuse rating, recloser size, and conductor type, which was obtained from the utility circuit maps. The substation all the way to the high-side voltage of the distribution transformers where the secondary connections (to the meters) are totaled to represent the customer load at each connection. The connectivity model was built utilizing the utility GIS pole and

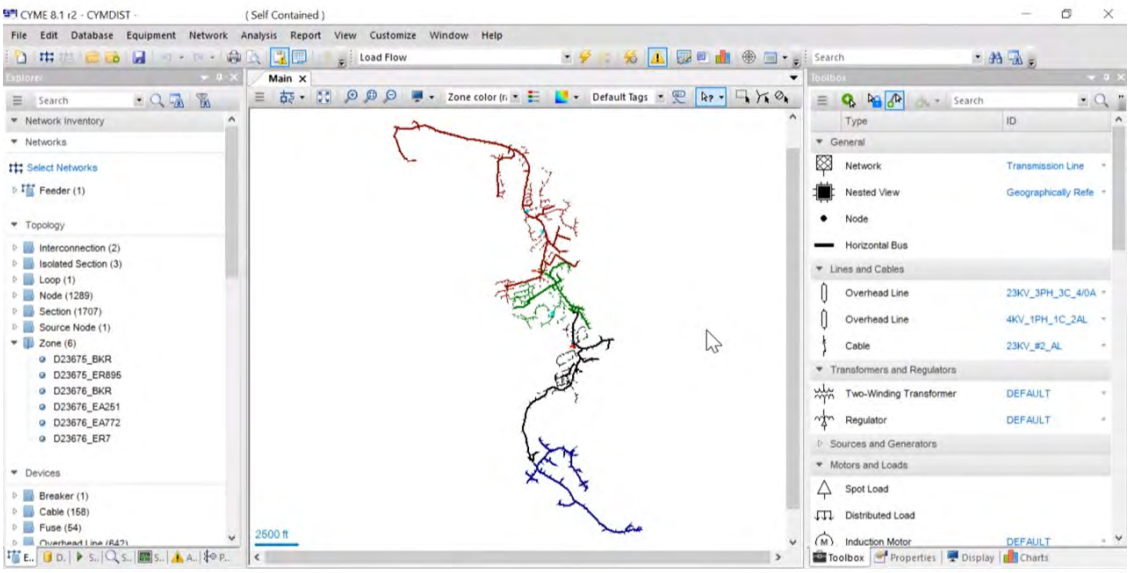


Figure 25: Depiction of a Real-utility Distribution Circuit Modeled in CYME Software.

transformer data to construct the main structure of the model.

Currently, the models are built and verified by limited data and outdated maps, causing uncertainty in the results. However, with the utilization of the machine learning algorithms, the models can be reversed engineered by utilizing the AMI voltage data to identify system topology, identify the DER interconnections, and then build the model in the CYME software. Fig. 30 is an example of a radial distribution feeder that was developed in CYME, and Fig. 31 is an example voltage dataset that was used to build a model with the existing available data. The voltage data available to verify the model is the equipment at the substation breaker as well as the intelligent pole top devices that are labeled "ERXXX" in Fig. 31 as illustrated, providing an example of the limited data available for the study. And provide insight onto the importance of having more granular data available to the utility to better represent the grid and grid's current capabilities.

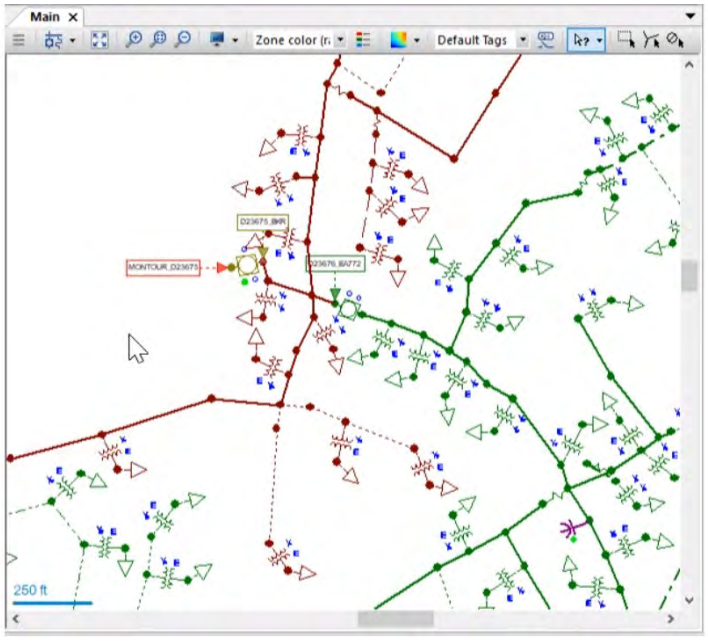


Figure 26: Depiction of a Real-utility Distribution Circuit Modeled in CYME Software (Zoom-in to Depict Load Models).

5.1.2 High-Level Summary of Analyses Ran During a Hosting Capacity Analysis

In this project, we aimed at answering the questions of "what now" and "what if" for situational awareness. The goal is to provide a model based on existing architecture and the expertise of a distribution system planning engineers. For "what now", we would like to know the overloaded area as well as the voltage profile and the stresses of a feeder, so that a timely upgrade is executed only in the area that has a need. For this purpose, we need to know the state of the system, where the topology is the foundation.

The following is a summary of screening metrics that are utilized for a hosting capacity analysis: over voltage, under voltage, voltage deviation, thermal loading, additional fault current, protection reach, synthetic tripping, unintentional islanding, and reverse power flow.

Traditional distribution feeders are designed based on radial conditions with power flow-

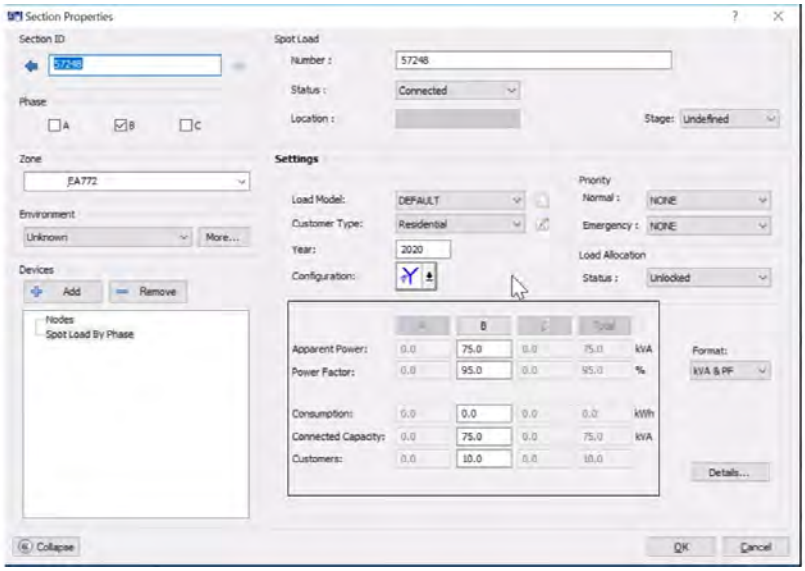


Figure 27: Illustration of the Graphical GUI Representing the Ability to Breakdown the Load Modeling by Individual Transformers.

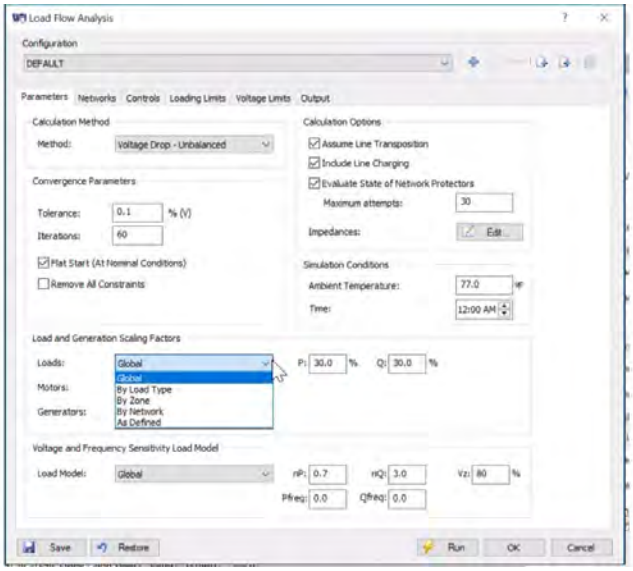


Figure 28: Further Illustration Providing the Capability to Input the Granularity

ing from the substation to load along a feeder which results in a voltage drop. If DER is not located appropriately the voltage can rise above acceptable limits, especially during light load conditions therefore, the maximum generation (discharging DER) capacity that can be installed on each node before reaching any over voltage condition is calculated.

The opposite is true during peak loading conditions exasperating the voltage drop along the feeder resulting in under voltage conditions. And finally, the voltage deviation is calculated in regards to the maximum DER available on the circuit before adversely impacting the voltage regulation set points and equipment. Another study performed in determining the maximum generation of the DER discharging before the specified thermal loading limit is reached on the circuit.

Another consideration to consider is that the distribution circuits often have several protective devices connected, such as circuit breakers, reclosers, fuses, etc. When the available fault current along a feeder changes, the coordination of protective devices may also be impacted. Another analysis is performed, and the percent change in fault current greater than the screening criteria are flagged as a concern to ensure the additional DER does not impact the existing protection coordination.

The last area of study during a hosting capacity analysis is determining the amount of DER installed, which will create a deviation in the substation breakers' fault current that will be higher than the specified protection reduction of reach in the breaker fault current and the calculations are done to determine the amount the DER that is installed at each point on the feeder that would generate a zero sequence sympathetic fault current which would cause sympathetic tripping due to back-feed of the DER's short-circuit contributions. This study also determines the maximum generation capacity that can be installed on each node before generating any reverse power flow in the circuit. It is anticipated the limit will be near the minimum load for the circuit since the primary concern is the reverse flow through the breaker at the substation.

Refer to Fig. 29 lists the screening metrics used to flag potential concerns with the addition of the DER based on guidance from [16].

The following is the approach used for the hosting capacity analysis for this circuit to provide insight on the importance of having an accurate load model.

Category	Criteria	Screening Limit
Voltage	Primary Overvoltage	105%
	Primary Undervoltage	95%
	Primary Voltage Deviation	3%
	Regulator Voltage Deviation (% of Bandwidth)	50%
Loading	Loading Thermal Loading	100%
Protection	Max. Deviation in Feeder Fault Current	10%
	Max. Deviation in Breaker Fault Current	10%
	Sympathetic Tripping (Zero Sequence Current)	150 A

Figure 29: Listing of the Screening Metrics Used to Flag Potential Concerns with DER Interconnection Based on Guidance from [16].

5.2 Overall Approach for Performing a Hosting Capacity Analysis on a Circuit

Base cases were created to bound the results of the simulations under different loading conditions based on historical data representing overall peak and minimum loading conditions for the circuit. The model has 10,000+ customers being represented each having a different unique consumption behaviors therefore it is important to review the response of the circuits performance under 100,000+ scenarios scaling the load under various levels and various levels of DER penetrations. The base case scenarios set are a percentage of the peak loading and light loading to bound the results.

- Heavy Load (30% of connected kVA)
- Light Load (10% of connected kVA)

The study is to determine the upper bounds when looking into the future, so the specific location and rating for the DER is not known. Therefore, cases are examined looking at numerous (e.g., thousands) potential DER location and ratings.

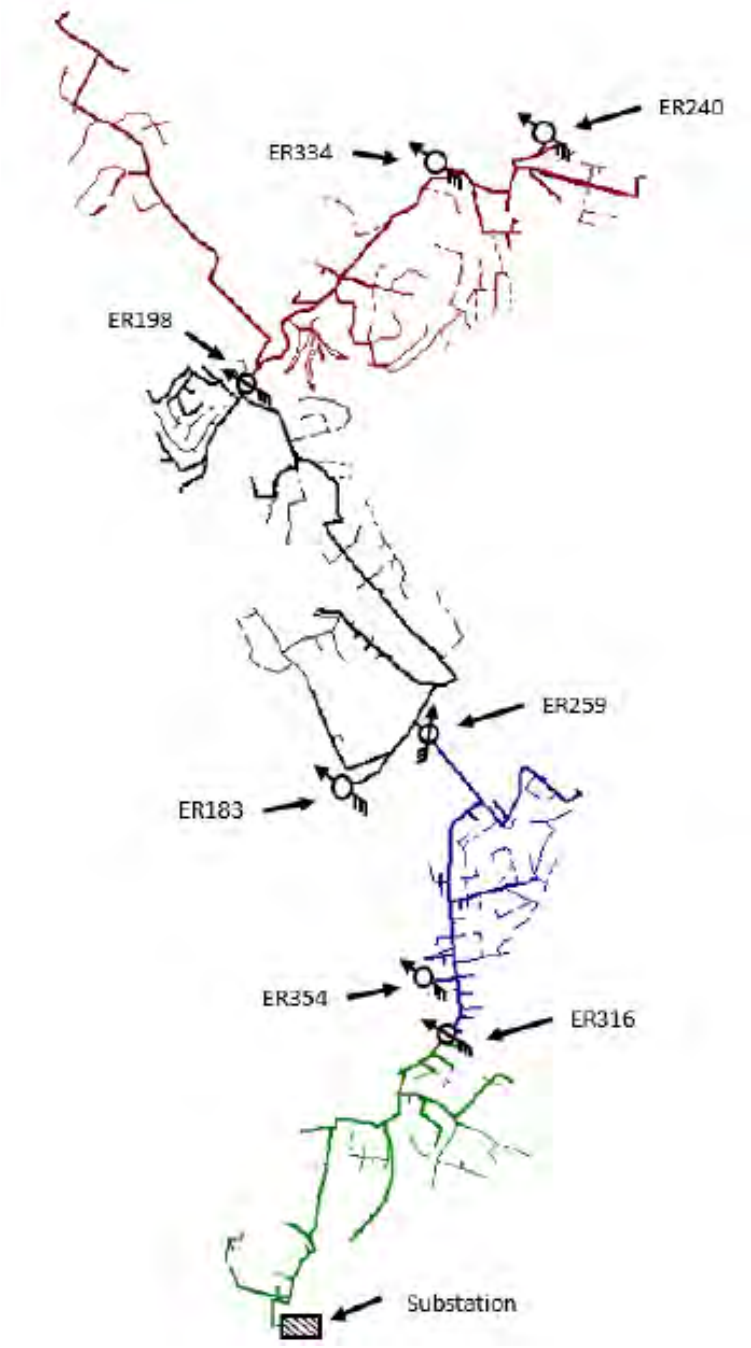


Figure 30: Depiction of a Real-Utility Distribution Circuit Modeled in CYME Software.

For instance, the following is to be examined in this analysis:

- Large (three-phase) and small (single-phase) DER distributed throughout the feeder

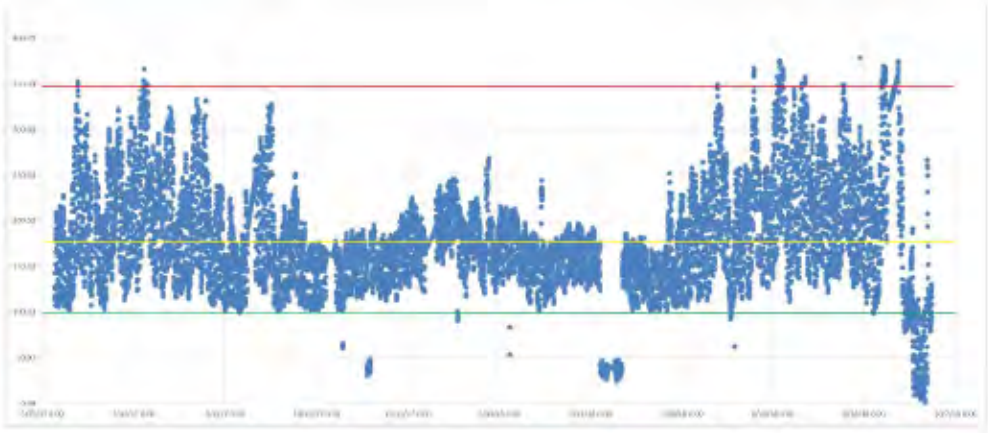


Figure 31: Example Data-Set for a Real-Utility Substation Breaker which are Currently Limited to System Planners.

- Centralized (large three-phase) DER at various locations throughout the feeder

For the hosting capacity analysis, the EPRI DRIVE module was used to screen for any potential concerns in the circuit.

5.2.1 Results

Refer to Fig. 32 and Fig. 33 which are the tabulated results after running the various hosting capacity studies on the circuit modeled. The results are a summation of all 10,000+ scenarios that are run using the CYME EPRI Drive tool.

It was shown that depending on how the individual loads varied point to point would provide a various significant changes to the results. Therefore it became very apparent that having the foundation of the study, i.e. circuit model loading scaled and verified using AMI metered data which provides a granular look at the whole circuits loading instead of a conservative broad brush approach across the whole circuit's loading based only on the feeder head data. The results depending on which criteria was ranked changed as the DER was varied across the circuit in large and small distribution as well as centralized connections. An accurate model will have a significant impact on the utilities awareness of what can and

can not be connected to the grid and ultimately what mitigation needs if any are available to continue to connect DER throughout the distribution grid. With the data available only at the breaker and pole-top switches it limits the utilities ability to provide DER capacity to the laterals and different locations providing an easier transition and interconnection process to all customers that wish to connect to the grid.

Verification of Hosting Capacity Under 10% Global Load	Distributed Large DER		Distributed Small DER		Centralized Large DER	
	Min (kW)	Max (kW)	Min (kW)	Max (kW)	Min (kW)	Max (kW)
Primary Over-Voltage	900	1100	50	50	50	20000
Primary Under-Voltage	8000	8800	50	50	1850	20000
Primary Voltage Deviation	1500	1850	50	50	50	20000
Regulator Voltage Deviation	2400	2400	50	50	350	20000
Thermal Loading - Generation (Discharging DER)	970	10880	970	10880	160	19040
Thermal Loading - Load (Charging DER)	890	6420	890	6420	140	12790
Reverse Flow	---	2420	---	2460	---	3120
Additional Element Fault Current	10870	10870	3100	3100	270	20000
Protection Reduction of Reach	550	550	550	550	550	550
Sympathetic Tripping	20000	20000	3060	3060	20000	20000
Unintentional Islanding	---	2420	---	2460	---	3120
Hosting Capacity	**550**	**550**	**50**	**50**	**50**	**550**

Figure 32: Illustration of the Results of the Study under Light Load Conditions Modeling 10% of the Peak Load.

Refer to Fig. 34 as a visual representation of the results measuring the quantitative results against the metrics as shown in Fig. 29.

Verification of Hosting Capacity Under 30% Global Load	Distributed Large DER		Distributed Small DER		Centralized Large DER	
	Min (kW)	Max (kW)	Min (kW)	Max (kW)	Min (kW)	Max (kW)
Primary Over-Voltage	1350	1650	50	50	50	20000
Primary Under-Voltage	6250	6700	50	50	1500	20000
Primary Voltage Deviation	1500	1850	50	50	50	20000
Regulator Voltage Deviation	2400	2400	50	50	350	20000
Thermal Loading - Generation (Discharging DER)	1000	14170	1000	14170	170	20000
Thermal Loading - Load (Charging DER)	700	700	700	3540	130	6210
Reverse Flow	---	7160	---	7270	---	9390
Additional Element Fault Current	10870	10870	9300	9300	270	20000
Protection Reduction of Reach	550	550	550	550	550	550
Sympathetic Tripping	20000	20000	3060	4680	20000	20000
Unintentional Islanding	---	7160	---	7270	---	9390
Hosting Capacity	**550**	**550**	**50**	**50**	**50**	**550**

Figure 33: Illustration of the Results of the Study under Peak Load Conditions Modeling 30% of the Peak Load.

The results also show the impact of adding an additional 650 kW to the base case as am example of how the results change if a utility was unaware of all of their DER connections. The additional 650 kW causes a change in the overall capacity of the circuit to "host" DER. Reiterating again, the importance of having a verified and accurate model for all laterals as well as the main feeder of the distribution circuit.

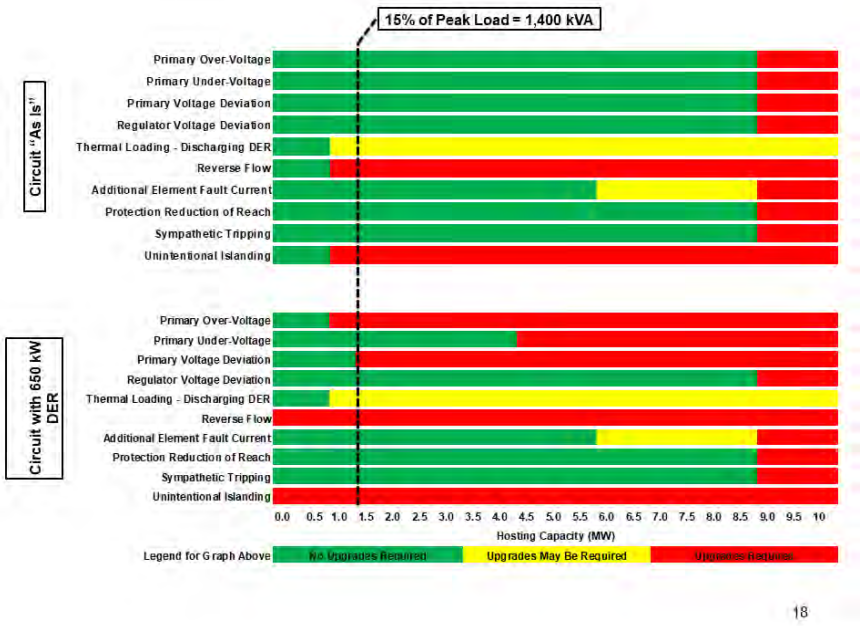


Figure 34: Visual Depiction of Hosting Capacity Results Comparing no DER versus 650 kW Added Throughout Circuit.

5.3 Additional Benefits to an Accurate Topology Model and Hosting Capacity Analyses Capabilities

Such a topology estimator can not only let people know where the trouble is in an event, e.g., an outage but also speed up the repairing process, e.g., sectionalizing the problematic location, phase correction. It will also provide a comprehensive knowledge of potential voltage violations, e.g., houses near the secondary transformer have higher voltages, and secondary transformers close to the primary transformers have higher voltages. Also, the transformer aging speed can be calculated by forecasting the customer usage in an area, for which we will also use the topology information to estimate the potential usage of the transformer soon.

Finally, this project can show insights into how to operate with increasing intermittent generation of solar sites, providing a guideline to supervise the future installation. For

example, we can use the topology to calculate the hosting capacity of a feeder, and the utility will obtain knowledge about the optimal locations of placing the voltage regulators in the distribution grid with a fast-evolving pattern, due to solar panels and EVs.

Specifically, with surging renewable penetration, electric vehicle usage, and increasing power usage in some areas, the grid is aging much faster, and it is essential to gather necessary information for both the planning and operating purposes. To obtain such a situational awareness, existing data in the system serves as an economical and viable way to conduct data analytics rather than hiring expensive consultant companies and upgrading all the components in a utility at once.

6.0 Conclusion

Transmission and distribution (T&D) systems, although they connect as an integrated system, have actually been planned, designed, and operated very differently. Transmission systems have been designed in a network configuration allowing for two-way power flow. Distribution systems are designed radially, allowing for one-directional power flow. Because of these traditional design approaches and higher penetration levels of DER, there must be increased coordination in managing the power flow on existing infrastructure in the US.

To plan and operate the T&D system, T&D planners/operators must collect and share validated data across the transmission-distribution interface that includes a multitude of stakeholders. These stakeholders include developers/aggregators, Electric Distribution Companies' (EDCs), Transmission Owners' (TOs), Regional Transmission Operators/ Independent System Operators (RTO/ISO), state regulators, and the Federal Energy Regulatory Commission (FERC). Understanding the transmission-distribution interface was not an issue when power generation was only connected to the transmission system. These variable locations of DER within the power system will drastically impact generation and load forecast models, especially if utilities are unaware of the DER (as in private home rooftop solar).

RTO/ISO, TOs, and EDCs (e.g., electric utility) presently base all of their system planning forecasts on historical data. As DER is a newer, realizable technology for the electric utility, there is no DER historical data for the planners/operators to utilize.

The increase in DER systems not only create sustainable and green energy for the human society, but also build a new type of assets for distribution utilities. To better evaluate the benefits and create new revenues, utilities need to know the locations of these new components to manage the unconventional two-way power flow for sustainable management of distribution grids.

For example, detecting and monitoring all the active PV installations in a utility's territory allow the utility to perform accurate hosting capacity analyses (HCA). With HCA, utilities determine the amount of additional Distributed Energy Resources (DER) that can be "hosted" on the distribution system at a given time and at a given location, without

threatening grid safety, reliability, or power quality [51].

Therefore utilities must be prepared for the rapidly proliferating industry and redefine their existing forecasting tools which the first step in any software analysis tool is the identification of the topology and the ability to consistently and frequently monitor the changes that are made on the distribution grid. The research herein has provided a methodology to enhance the operation team's ability to analyze and understand the impacts of any array of DER penetration.

6.1 Research Directions and Applications

Two journal papers were submitted for publication for the first two areas of research. The future direction to take this research is to continue to understand and develop additional methods to help aid the transition of the current power grid to the future grid (i.e., data utility grid).

Operational data from the AMI meters will be requiring more and more software and additional machine learning algorithms to maintain the relational and non-relational data to support business processes and event processing. An operational data management solution will be required to organize the collection of data that may consist in multiple formats and be stored in various forms. The methods developed in these projects can evolve and continue to support the need to identify the relational and non-relational data for utilities and external third-party stakeholders for the benefit of the holistic grid.

For today's electric utility, an intelligent and connected infrastructure has unleashed a tidal wave of operational data that can be used to improve operational efficiency, meet customer demands, and anticipate risks to reliability. This will be required to maintain the safe, reliable, and affordable grid to serve all of society. With the combination of data management and analytics the power industry will be able to continue to take a system-wide view of the current operations, allowing them to run more efficiently and at a lower costs with less fossil fuel dependencies. A single view of operational data across the utility will provide a trusted source of operational data used to make decisions, used to support and

expedite various business processes, and map all of the data to all business units to provide actionable insights for all stakeholders. This is where this project and additional projects developed from our learning's will fit in and contribute to the evolution of the future power grid.

Bibliography

[1] Andrés Argüello, José D Lara, José David Rojas, and Gustavo Valverde. Impact of rooftop PV integration in distribution systems considering socioeconomic factors. *IEEE Systems Journal*, 12(4):3531–3542, 2017.

[2] Freek Baalbergen, Madeleine Gibescu, and Lou van der Sluis. Modern state estimation methods in power systems. In *IEEE/PES Power Systems Conference and Exposition*, pages 1–6, 2009.

[3] Mesut E Baran, Jaesung Jung, and Thomas E McDermott. Topology error identification using branch current state estimation for distribution systems. In *Transmission & Distribution Conference & Exposition: Asia and Pacific*, pages 1–4. IEEE, 2009.

[4] Yakoub Bazi and Farid Melgani. Convolutional SVM networks for object detection in UAV imagery. *IEEE transactions on geoscience and remote sensing*, 56(6):3107–3118, 2018.

[5] Saverio Bolognani, Nicoletta Bof, Davide Michelotti, Riccardo Muraro, and Luca Schenato. Identification of power distribution network topology via voltage correlation analysis. In *Conference on Decision and Control*, pages 1659–1664. IEEE, 2013.

[6] Guido Cavraro, Reza Arghandeh, Grazia Barchi, and Alexandra von Meier. Distribution network topology detection with time-series measurements. In *Power & Energy Society Innovative Smart Grid Technologies Conference*, pages 1–5. IEEE, 2015.

[7] Olivier Chapelle, Patrick Haffner, and Vladimir N Vapnik. Support vector machines for histogram-based image classification. *IEEE transactions on Neural Networks*, 10(5):1055–1064, 1999.

[8] Danling Cheng, Barry A Mather, Richard Seguin, Joshua Hambrick, and Robert P Broadwater. Photovoltaic (PV) impact assessment for very high penetration levels. *IEEE Journal of photovoltaics*, 6(1):295–300, 2015.

[9] Zhang Chi, Luo Pandian, Zhang Xueying, Zeng Jie, Zhao Wei, Huang Jiajian, Xie Yu, and Xu Qi. Research on the impacts of grid-connected distributed photovoltaic on load characteristics of regional power system. In *International Conference on Green Energy and Applications*, pages 95–99, 2017.

[10] Andy Colman, Dan Wilson, and Daisy Chung. Beyond the meter planning the distributed energy future. Technical report, Smart Electric Power Alliance and Black & Veatch Holding Company, Report, 2017.

[11] Sanjoy Dasgupta and Samory Kpotufe. Optimal rates for k-nn density and mode estimation. In *Advances in Neural Infor. Processing Sys.*, 2014.

[12] Deepjyoti Deka, Michael Chertkov, and Scott Backhaus. Topology estimation using graphical models in multi-phase power distribution grids. *IEEE Transactions on Power Systems*, 2019.

[13] Maria Carmela Di Piazza, Antonella Ragusa, Gianpaolo Vitale, et al. Identification of photovoltaic array model parameters by robust linear regression methods. In *International Conference on Renewable Energies and Power Quality*, pages 143–149, 2009.

[14] Daniel L Donaldson and Dilan Jayaweera. Effective solar prosumer identification using net smart meter data. *International Journal of Electrical Power & Energy Systems*, 118:105823, 2020.

[15] Harris Drucker, Donghui Wu, and Vladimir N Vapnik. Support vector machines for spam categorization. *IEEE Transactions on Neural networks*, 10(5):1048–1054, 1999.

[16] EPRI. Stochastic analysis to determine feeder hosting capacity for distributed solar pv. Technical report, EPRI, Palo Alto, CA: 2012.1026640, 2012.

[17] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Knowledge Discovery and Data Mining*, 1996.

[18] Omar F Fajardo and Alberto Vargas. Reconfiguration of mv distribution networks with multicost and multipoint alternative supply, part ii: Reconfiguration plan. *IEEE Transactions on Power Systems*, 23(3):1401–1407, 2008.

[19] Vladimir Golovko, Sergei Bezobrazov, Alexander Kroshchanka, Anatoliy Sachenko, Myroslav Komar, and Andriy Karachka. Convolutional neural network based solar photovoltaic panel detection in satellite photos. In *IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications*, volume 1, pages 14–19, 2017.

[20] Klaus Greff, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. Neural expectation maximization. In *Advances in Neural Information Processing Systems*, pages 6691–6701, 2017.

[21] Zhaomiao Guo, Zhi Zhou, and Yan Zhou. Impacts of integrating topology reconfiguration and vehicle-to-grid technologies on distribution system operation. *IEEE Transactions on Sustainable Energy*, 2019.

[22] Mohsen Hajighorbani, SM Reza Hashemi, B Minaei-Bidgoli, and Shabnam Safari. A review of some semi-supervised learning methods. In *IEEE international conference on new research achievements in electrical and computer engineering*, pages 250–259, 2016.

[23] Borui Hou, Jianyong Yang, Pu Wang, and Ruqiang Yan. Lstm-based auto-encoder model for ecg arrhythmias classification. *IEEE Transactions on Instrumentation and Measurement*, 69(4):1232–1240, 2019.

[24] Jing Huang, Vijay Gupta, and Yih-Fang Huang. Electric grid state estimators for distribution systems with microgrids. In *Annual Conference on Information Sciences and Systems*, pages 1–6. IEEE, 2012.

[25] Institute of Energy Systems and Electrical Drives. *Adres-dataset.* "http://www.ea.tuwien.ac.at/projects/adres concept/EN/", 2016.

[26] Heinrich Jiang, Jennifer Jang, and Ofir Nachum. Robustness guarantees for density clustering. In *Inter. Conf. on Artificial Intelligence and Stats.*, 2019.

[27] William H Kersting. *Distribution system modeling and analysis*. CRC press, 2006.

[28] KHON2. HECO customers asked to disconnect unauthorized PV systems, 2014. accessed August 2020.

[29] Kevin Killingsworth and Jomar M. Perez. Aep guidelines for transmission owner identified needs. Technical report, American Electric Power (AEP), Report, 2018.

[30] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[31] George N Korres and Nikolaos M Manousakis. A state estimation algorithm for monitoring topology changes in distribution systems. In *Power and Energy Society General Meeting*, pages 1–8. IEEE, 2012.

[32] Balaji Krishnapuram, Lawrence Carin, Mario AT Figueiredo, and Alexander J Hartemink. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE transactions on pattern analysis and machine intelligence*, 27(6):957–968, 2005.

[33] Nitin Kumar and Suyash P Awate. Semi-supervised robust mixture models in rkhs for abnormality detection in medical images. *IEEE Transactions on Image Processing*, 29:4772–4787, 2020.

[34] Prabha Kundur, Neal J Balu, and Mark G Lauby. *Power system stability and control*, volume 7. McGraw-hill New York, 1994.

[35] SungWon Lee, Kwang Eun An, Byung Don Jeon, Kyoung Yeon Cho, Seung Jae Lee, and Dongmahn Seo. Detecting faulty solar panels based on thermal image processing. In *IEEE International Conference on Consumer Electronics*, pages 1–2, 2018.

[36] Jun Li, José M Bioucas-Dias, and Antonio Plaza. Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning. *IEEE Transactions on Geoscience and Remote Sensing*, 48(11):4085–4098, 2010.

[37] Zhi Li, Liqun Yang, and Zhoujun Li. Mixture-model-based graph for privacy-preserving semi-supervised learning. *IEEE Access*, 8:789–801, 2019.

[38] KC Liao and JH Lu. Using matlab real-time image analysis for solar panel fault detection with UAV. In *Journal of Physics: Conference Series*, volume 1509, page 012010, 2020.

[39] Yizheng Liao, Yang Weng, and Ram Rajagopal. Urban distribution grid topology reconstruction via lasso. In *IEEE Power and Energy Society General Meeting*, pages 1–5, 2016.

[40] Z Chase Lipton, Charles Elkan, and Balakrishnan Narayanaswamy. Thresholding classifiers to maximize f1 score. *Machine Learning and Knowledge Discovery in Databases*, 8725:225–239, 2014.

[41] RL Lugtu, DF Hackett, KC Liu, and DD Might. Power system state estimation: Detection of topological errors. *IEEE Transactions on Power Apparatus and Systems*, (6):2406–2412, 1980.

[42] Xiaoyi Mai and Romain Couillet. A random matrix analysis and improvement of semi-supervised learning for large dimensional data. *The Journal of Machine Learning Research*, 19(1):3074–3100, 2018.

[43] Alireza Makhzani and Brendan Frey. K-sparse autoencoders. *arXiv preprint arXiv:1312.5663*, 2013.

[44] Tom M Mitchell. Learning from labeled and unlabeled data part 2: coupled training. *Machine learning*, 10:601, 2009.

[45] Amir Mosavi, Mohsen Salimi, Sina Faizollahzadeh Ardabili, Timon Rabczuk, Shahaboddin Shamshirband, and Annamaria R Varkonyi-Koczy. State of the art of machine learning models in energy systems, a systematic review. *Energies*, 12(7):1301, 2019.

[46] Pramuditha Perera and Vishal M Patel. Learning deep features for one-class classification. *IEEE Transactions on Image Processing*, 28(11):5450–5463, 2019.

[47] Y CA Padmanabha Reddy, P Viswanath, and B Eswara Reddy. Semi-supervised learning: A brief review. *International Journal of Engineering & Technology*, 7(1.8):81–85, 2018.

[48] M. Ribeiro, M. Gutoski, A. E. Lazzaretti, and H. S. Lopes. One-class classification in images and videos using a convolutional autoencoder with compact embedding. *IEEE Access*, 8:86520–86535, 2020.

[49] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402, 2018.

[50] Addisson Salazar, Gonzalo Safont, and Luis Vergara. Semi-supervised learning for imbalanced classification of credit card transaction. In *International Joint Conference on Neural Networks*, pages 1–7, 2018.

[51] Volker Schwarzer and Reza Ghorbani. Transient over-voltage mitigation and its prevention in secondary distribution networks with high PV-to-load ratio. Technical

Report HNEI-02-15, Hawai'i Natural Energy Institute, February 2015. accessed August 2020.

[52] Smart*. Umass smart* dataset. Accessed July 31, 2020.

[53] Hongchao Song, Zhuqing Jiang, Aidong Men, and Bo Yang. A hybrid semi-supervised anomaly detection model for high-dimensional data. *Computational intelligence and neuroscience*, 2017(8501683):9, 2017.

[54] Manuel Stritt, Lars Schmidt-Thieme, and Gerhard Poeppel. Combining multi-distributed mixture models and bayesian networks for semi-supervised learning. In *International Conference on Machine Learning and Applications*, pages 354–362, 2007.

[55] David MJ Tax and Robert PW Duin. Support vector data description. *Machine learning*, 54(1):45–66, 2004.

[56] Hicham Tribak and Youssef Zaz. Remote solar panels identification based on patterns localization. In *International Renewable and Sustainable Energy Conference*, pages 1–5, 2018.

[57] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

[58] Alexandra Von Meier, David Culler, Alex McEachern, and Reza Arghandeh. Micro-synchrophasors for distribution systems. In *ISGT 2014*, pages 1–5. IEEE, 2014.

[59] Yang Weng, Rohit Negi, and Marija D Ilić. A search method for obtaining initial guesses for smart grid state estimation. In *International Conference on Smart Grid Communications*, pages 599–604. IEEE, 2012.

[60] Yang Weng, Rohit Negi, and Marija D Ilić. Historical data-driven state estimation for electric power systems. In *International Conference on Smart Grid Communications*, pages 97–102. IEEE, 2013.

[61] Yang Weng and Ram Rajagopal. Probabilistic baseline estimation via gaussian process. In *Power & Energy Society General Meeting*, pages 1–5. IEEE, 2015.

[62] Allen J Wood, Bruce F Wollenberg, and Gerald B Sheblé. *Power generation, operation, and control*. John Wiley & Sons, 2013.

[63] Yizheng Liao, Y. Weng, Meng Wu, and R. Rajagopal. Distribution grid topology reconstruction: An information theoretic approach. In *North American Power Symposium*, pages 1–6, Oct 2015.

[64] Jiafan Yu, Yang Weng, Chin-Woo Tan, and Ram Rajagopal. Probabilistic estimation of the potentials of intervention-based demand side energy management. In *International Conference on Smart Grid Communications*, pages 865–870. IEEE, 2015.

[65] Ahmed Samir Zamzam, Xiao Fu, and Nicholas D Sidiropoulos. Data-driven learning-based optimization for distribution system state estimation. *IEEE Transactions on Power Systems*, 2019.

[66] G Zhang, S Lee, Ritchie Carroll, Jian Zuo, Lisa Beard, and Yilu Liu. Wide area power system visualization using real-time synchrophasor measurements. In *IEEE PES general meeting*, pages 1–7. IEEE, 2010.

[67] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. Birch: An efficient data clustering method for very large databases. *SIGMOD Rec.*, 25(2):103–114, June 1996.

[68] Xiaochen Zhang and Santiago Grijalva. A data-driven approach for detection and estimation of residential PV installations. *IEEE Transactions on Smart Grid*, 7(5):2477–2485, 2016.

[69] M. Zhao, R. H. M. Chan, T. W. S. Chow, and P. Tang. Compact graph based semi-supervised learning for medical diagnosis in alzheimer's disease. *IEEE Signal Processing Letters*, 21(10):1192–1196, 2014.