Development of synthetic power distribution networks and datasets with industrial validation

Muhammad Ali

A thesis in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

UNSW
THE UNIVERSITY OF NEW SOUTH WALES

CANBERRA

School of Engineering and Information Technology

University of New South Wales

March 2022

Thesis/Dissertation Sheet

Surname/Family Name	:	**Ali**
Given Name/s	:	**Muhammad**
Abbreviation for degree as give in the University calendar	:	**PhD**
Faculty	:	**UNSW Canberra**
School	:	**School of Engineering and Information Technology**
Thesis Title	:	**Development of synthetic power distribution networks and datasets with industrial validation**

Abstract: (PLEASE TYPE)

This thesis addresses a key challenge for creating synthetic distribution networks and open-source datasets by combining the public databases and data synthesis algorithms. Novel techniques for the creation of synthetic networks and open-source datasets that enable model validation and demonstration without the need for private data are developed. The developed algorithms are thoroughly benchmarked against existing approaches and validated on industry servers to highlight their usefulness in solving real-world problems.

A review using novel techniques that provides unique insights into the literature is conducted to identify research gaps. Based on this review, three contributions have been made in this thesis. The first contribution is the development of a data protection framework for anonymizing sensitive network data. A novel approach is proposed based on the maximum likelihood estimate for estimating the parameters that represent the actual data. A data anonymization algorithm that uses the estimated parameters to generate realistic anonymized datasets is developed. A Kolmogorov-Smirnov test criteria is used to create realistic anonymized datasets. Validation is carried out by collecting actual network data from an energy company and comparing it to anonymized datasets created using the methods developed in this thesis. The application of this method is shown by performing simulation studies on the IEEE 123-node test feeder.

The second contribution is developing a practical approach for creating synthetic networks and datasets by integrating the open-source data platforms and synthesis methods. New data synthesis algorithms are proposed to obtain the network datasets for electricity systems in a chosen geographical area. The proposed algorithms include a topology for designing power lines from road infrastructure, a method for computing the lengths of power lines, a hub-line algorithm for determining the number of consumers connected to a single transformer, a virtual layer approach based on FromNode and ToNode for establishing electrical connectivity, and a technique for ingesting raw data from the developed network to industrial data platforms. The practical feasibility of the proposed solutions is shown by creating a synthetic test network and datasets for a distribution feeder in the Colac region in Australia. The datasets are then validated by deploying them on industry servers. The results are compared with actual datasets using geo-based visualizations and by including feedback from industry experts familiar with the analysis.

The third contribution of this thesis is to address the problem of electric load profile classification in the context of buildings. This classification is essential to effectively manage energy resources across power distribution networks. Two new methods based on sparse autoencoders (SAEs), and multi-stage transfer learning (MSTL) are proposed for load profile classification. Different from conventional hand-crafted feature representations, SAEs can learn useful features from vast amounts of building data in an unsupervised automatic way. The problems of missing data and class imbalance for building datasets are addressed by proposing a minority over-sampling algorithm that effectively balances missing or unbalanced data by equalizing minority and majority samples for fair comparisons. The practical feasibility of the methodology is shown using two case studies that include both public benchmark and real-world datasets of buildings. An empirical comparison is conducted between the proposed and the state-of-the-art methods in the literature. The results indicate that the proposed method is superior to traditional methods, with a performance improvement from 1 to 10 percent.

Copyright Statement

Signed .....

Date .....

Authenticity Statement

I certify that the Library deposit digital copy is a direct equivalent of the final officially approved version of my thesis. No emendation of content has occurred and if there are any minor variations in formatting, they are the result of the conversion to digital format.

Signed .....

Date .....

Originality Statement

I hereby declare that this submission is my own work and to the best of my knowledge it contains no material previously published or written by another person, or substantial portions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institute, except where due acknowledgment is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.

Signed .....

Date .....

INCLUSION OF PUBLICATIONS STATEMENT

UNSW is supportive of candidates publishing their research results during their candidature as detailed in the UNSW Thesis Examination Procedure.

Publications can be used in their thesis in lieu of a Chapter if:

- The candidate contributed greater than 50% of the content in the publication and is the "primary author", ie. the candidate was responsible primarily for the planning, execution and preparation of the work for publication
- The candidate has approval to include the publication in their thesis in lieu of a Chapter from their supervisor and Postgraduate Coordinator.
- The publication is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in the thesis

Please indicate whether this thesis contains published material or not:

☐ This thesis contains no publications, either published or submitted for publication
(if this box is checked, you may delete all the material on page 2)

☒ Some of the work described in this thesis has been published and it has been documented in the relevant Chapters with acknowledgement
(if this box is checked, you may delete all the material on page 2)

☐ This thesis has publications (either published or submitted for publication) incorporated into it in lieu of a chapter and the details are presented below

CANDIDATE'S DECLARATION

I declare that:

- I have complied with the UNSW Thesis Examination Procedure
- where I have used a publication in lieu of a Chapter, the listed publication(s) below meet(s) the requirements to be included in the thesis.

Candidate's Name	Signature	Date (dd/mm/yy)
Muhammad Ali		

List of Publications and Awards

List of Publications

1. **[Journal]** M. Ali, K. Prakash, M. A. Hossain, and HR. Pota, "Intelligent energy management: Evolving developments, current challenges, and research directions for sustainable future," ***Elsevier Journal of Cleaner Production***, vol. 314, p. 127904, Sep. 2021. **IF: 9.297**.

2. **[Journal]** M. Ali, K. Prakash, C. Macana, M. Rabiul, A. Hussain, and HR. Pota, "Anonymization of distribution feeder data using statistical distribution and parameter estimation approach," ***Elsevier Sustainable Energy Technologies and Assessments***, vol. 52, p. 102152, Aug. 2022. **IF: 5.353**.

3. **[Journal]** M. Ali, K. Prakash, C. Macana, MQ. Raza, AK. Bashir and HR. Pota, "Modelling synthetic power distribution network and datasets with industrial validation," (Under-review) ***Elsevier Industrial Information Integration***, Dec. 2021. **IF: 10.063**.

4. **[Journal]** M. Ali, K. Prakash, C. Macana, AK. Bashir, A. Jolfaei, A. Bokhari, JJ. Klemes, and HR. Pota, "Modeling residential electricity consumption from public demographic data for sustainable cities," ***Energies***, vol. 15, no. 6, Art. no. 6, Jan. 2022. **IF: 3.004**

5. **[Conference]** M. Ali, C. A. Macana, K. Prakash, R. Islam, I. Colak, and HR. Pota, "Generating open-source datasets for power distribution network using OpenStreetMaps," Sep. 2020, pp. 301–308, ***IEEE International Conference on Renewable Energy Research and Application***

6. **[Conference]** M. Ali, C. A. Macana, K. Prakash, B. Tarlinton, R. Islam, and HR. Pota, "A novel transfer learning approach to detect the location of transformers in distribution Network," June. 2020, pp. 56–60, ***IEEE International Conference on Smart Grid***

7. **[Journal]** M. Ali, K. Prakash, MQ. Raza, AK Bashir, and HR. Pota, "Load profile classification of buildings using AI-based multi-stage transfer learning approach for sustainable energy future," (Under-review) ***Elsevier Energy***, May. 2022. **IF: 7.147**.

8. **[Conference]** M. Ali, K. Prakash, and HR. Pota, "A Bayesian approach based on acquisition function for optimal selection of deep learning hyperparameters: a case study with energy management data," Apr. 2020, vol. 2, no. 1, ***Science Proceedings Series***, Malaysia.

Co-Authored Publications

1. **[Journal]** K. Prakash, **M. Ali**, C. Macana, A. Hussain and HR. Pota, "Optimal battery energy storage system (BESS) management for ancillary services in low Voltage networks," (Under-review) ***Elsevier Renewable Energy***, Jan 2022, **IF: 8.001**

2. **[Conference]** K. Prakash, **M. Ali**, M. R. Islam, H. Mo, D. Dong, and H. Pota, "Optimal Coordination of Photovoltaics and Electric Vehicles for Ancillary Services in Low Voltage Distribution Networks", Oct. 2021, pp. 1008–1013, *IEEE International Conference on Electrical Machines and Systems (ICEMS)*

List of Awards

1. Received "Outstanding Research Award" from UNSW Sydney, Australia (2021)

2. Recipient of research grant (RG194345) from industry in Canberra to integrate PhD research (value: $\$15,000$)

3. Winner of UNSW 3-Minute Thesis (3MT) competition, 2019 (cash prize: $\$500$)

4. Winner of UNSW 3-Minute Thesis (3MT) competition, 2020 (cash prize: $\$500$)

5. 3rd position in ACT Government innovation network "ZeroCO2 Renewable Energy Sustainability Hackathon" competition – (cash prize: $\$2000$)

6. Our team won best paper award out of 400 papers in IEEE ICEMS conference, 2021

7. Obtained "advanced" level in UNSW communication for higher degree research

8. Received certificate from UNSW innovation pitch night competition

9. Recognition and certificate of merit for completing the Graduate Teaching Training Program at UNSW Canberra

10. Awarded prestigious UIPA scholarship by UNSW Australia for pursuing PhD studies

Other Contributions

1. Contributed to the ARENA-EVOLVE project led by Zeppelin Bend Pty Ltd (Australia) (Details here)

2. Contributed to the data science community. Presently ranked in the top 4% of the stackoverflow data science community, with a reputation of 2,267 (Details here)

3. Founding team member of DataMagic start-up

4. Conference session chair of IEEE SmartGrid 2020 in Paris, France, 2020

Acknowledgement

In the name of Allah, the most beneficent, the most merciful. My heartfelt gratitude goes to my family, the soul of my late father, mother, wife, supervisor, siblings and my relatives, for their unwavering support throughout my PhD journey. I am eternally grateful to them. I dedicate my thesis to them.

Pursuing a PhD is a fascinating journey. Four years ago, I decided to get involved in the topic of this manuscript to dedicate several years of my life and to give up the proximity of family and friends. It has been an adventure full of inspiring challenges and joyful experiences. Diving so deep into a topic can also bring difficulties and harsh times, but I have always been surrounded by wonderful people that stood by me in one way or another. I would like to sincerely thank them all.

I would like to express my heartfelt thanks to my supervisor, Prof. Hemanshu Pota, for his excellent supervision, guidance, encouragement, and patience throughout my research journey. Prof. Hemanshu has been extremely kind and supportive throughout this amazing journey of a PhD. His support and contributions in the form of brainstorming research ideas, comments on papers, assisting me in engaging with industry and other researchers, involving me in student projects, and inspiring me to undertake high-quality research have been tremendously beneficial. My sincere thanks to Prof. Jiankun Hu for his kind support and persistent encouragement. I am also grateful to Prof. Sean O'Byrne, Deputy Head of School, for nominating me for the UNSW research awards in 2021.

I am eternally grateful for the unconditional love, and support from my beloved mother (Nargis Khatoon) and wife (Dr. Saima Suman). This journey was not always smooth and easy, and it was not always possible for me to maintain a perfect balance between the research and our lives. Their love and support were always immutable for me, which helped me overcome the hurdles and uphold my sanity throughout this challenging endeavor. I would like to thank my siblings, cousins, brothers-in-law, parents-in-law and

other close relatives for their years of support and motivation.

I sincerely acknowledge the great support from industry partners for their financial support and development of this thesis. I am grateful to Dr. Carlos Macana and Bill Tarlington (Managing Director, Zepben) for their help and support during my PhD journey. My sincere thanks to the external collaborators, including Dr. Muhammad Qamar Raza and Dr. Akhtar Hussain, for their utmost dedication in improving my work. It has been a splendid experience working under their guidance the entire time. Their honest and constructive feedback, along with your extensive involvement, have played an instrumental role in my intellectual growth.

The environment in the UNSW-ADFA has been fantastic due to the people working there. Thanks to everyone, including Krishneel Prakash, Md Rasel Mahmud and other colleagues in the power system group. It is the family with whom I shared my passion for research and the real industry problems. My sincere thanks to Prof. Hussein A. Abbass, Dr. Saber Elsayed, Dr. Essam Debie, Dr. Aya Hussain, Dr. Heba El-Fiqi for their consistent support throughout my PhD.

I am thankful to my new friends in ADFA: Muhammad Shoaib Khan Niazi, Dr. Aziz Ahmad, Dr. Asghar Ali, Dr. Waqas Haider, Dr. Falak Nawaz, Dr. Khawar Mehmud, Dr. Muhammad Khubaib Khan, Mudasir Ahmed (deceased), Muhammad Haq Nawaz, and Muhammad Qasim, who made my experience in Australia worth remembering.

I am thankful to the School of Engineering and Information technology (SEIT) at UNSW Canberra for providing the required facilities and support to make this thesis work. A special thanks to Ms. Elvira Berra from the Student Administrative Services for her consistent support since my enrollment. My sincere thanks to Denise Russel (native English speaker) for her kind support in proofreading the thesis.

Abstract (701 words)

Problem

This thesis addresses a key challenge for creating synthetic distribution networks and open-source datasets by combining public databases and data synthesis algorithms. Novel techniques for synthetic network creation and open-source datasets that enable model validation and demonstration without the need for private data are developed. The developed algorithms are thoroughly benchmarked against existing approaches and validated on industry servers to highlight their usefulness in solving real-world problems.

Solution, validation and comparison

Three contributions have been made in this thesis. The first contribution is the development of a data protection framework for anonymizing sensitive network data. A novel approach is proposed based on the maximum likelihood estimate for estimating the parameters that represent the actual data. Then, a data anonymization algorithm that uses the estimated parameters to generate realistic anonymized datasets is developed. A Kolmogorov-Smirnov test criteria is used to create realistic anonymized datasets. Validation is carried out by collecting actual network data from an energy company and comparing it to anonymized datasets created using the methods developed in this thesis. The practical application of the method is shown on the IEEE 123-node test feeder.

The anonymization methods developed in the first contribution of the thesis are dependent on the data provided by electrical companies. What if energy companies refuse to provide this data due to the risk of privacy disclosures? To answer this question, the second contribution of the thesis is the development of synthetic networks and datasets by combining open-source data platforms and data synthesis algorithms. New data synthesis algorithms are proposed to create synthetic networks and datasets. The proposed algorithms include a topology for designing power lines from road infrastructure, a method for computing the lengths of power lines, a hub-line algorithm for determining the number of consumers connected to a single transformer, a virtual layer approach based on FromNode and ToNode for establishing electrical connectivity, and a technique for ingesting synthesised network data to industrial data platforms. The practical feasibility of the proposed solutions is demonstrated by creating a synthetic test network and datasets for a distribution feeder in the Colac region in Australia. The datasets are then validated by deploying them on industry servers and the results are compared with actual datasets using geo-based visualizations and by incorporating feedback from industry experts fa-

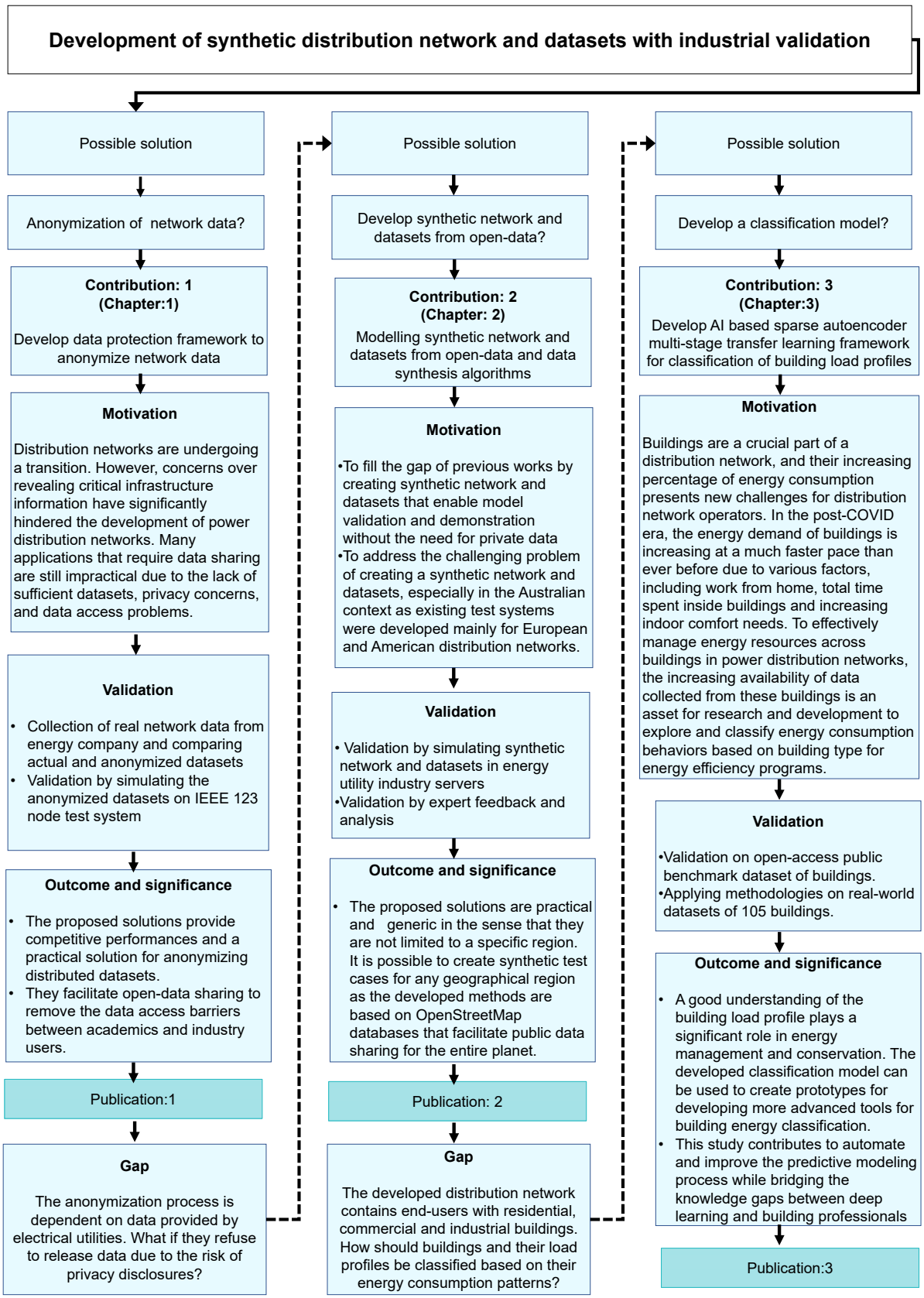
miliar with the analysis.

The crucial part in the creation of the synthetic networks are the buildings (end-users). The changing load profiles (demands) of buildings in the current COVID-19 situation presents new challenges for distribution network operators, as most people work from home and spend the majority of their time in buildings. The third contribution of this thesis is to address the problem of electric load profile classification in the context of buildings. This classification is essential to effectively manage energy resources across buildings in power distribution networks. Two new methods based on sparse autoencoders (SAEs) and multi-stage transfer learning (MSTL) are proposed for load profile classification. Different from conventional hand-crafted feature representations, SAEs can learn useful features from vast amounts of building data in an unsupervised automatic way. The problems of missing data and class imbalance for building datasets are addressed by proposing a minority over-sampling algorithm that effectively balances missing or unbalanced data by equalizing minority and majority samples for fair comparisons. The practical feasibility of the methodology is shown using two case studies that include both public data for benchmarking and real-world datasets of buildings. An empirical comparison is conducted between the proposed and the state-of-the-art methods in the literature. The results indicate that the proposed method is superior to traditional methods, with a performance improvement ranging from 1 to 10 percent.

Outcome and impact

In terms of societal impact, this thesis has three main implications. Firstly, the data anonymization contribution facilitates open-data sharing to remove the data access barriers between academic and industry users. Secondly, the synthetic network and datasets provide practical and generic solutions to have distribution network models that are not limited to a specific region. This approach overcomes the issues related to the dimensions and diversity of distribution systems in different geographic locations. Thirdly, the load profiling classification model can be used in the current era of COVID-19 to improve building energy efficiency, demand flexibility, and building-grid interactions.

Graphical summary and coherence of thesis



"Be kind, for whenever kindness becomes part of something, it beautifies it.
Whenever it is taken from something, it leaves it tarnished."
- Prophet Muhammad (peace be upon him)

Contents

List of Figures

List of Tables

Abbreviations

IEEE Institute of Electrical and Electronics Engineers

SAEs Sparse Autoencoders

MSTL Multi-Stage Transfer Learning

DSOs Distribution System Operators

B2G Building-to-Grid

PES Power and Energy Society

EPRI Electric Power Research Institute

ENTSO European Network of Transmission System Operators

PNNL Pacific Northwest National Laboratory

UKGS United Kingdom Generic System

OPENMOD Open Energy Modelling

OPSD Open Power System Data

OEP Open Energy Platform

ICSEG Illinois Center for a Smarter Electric Grid

MATPOWER Matlab-based power system simulation package

LINES Laboratory for Intelligent Integrated Networks of Engineering Systems

CIGRE International Council on Large Electric Systems

SciGRID Scientific GRID

EMT Electromagnetic Transients

OSM OpenStreetMap

QGIS Quantum Geographic Information Systems

API Application Programming Interface

EWB Energy workbench

CIM Common Information Model

IEC International Electrotechnical Commission

PyQGIS QGIS Python console

SQLite SQL compatible database system

HV High Voltage

LV Low Voltage

DNSP Distribution Network Service Provider

MLE Maximum Likelihood Estimate

K-S test Kolmogorov-Smirnov test

BGL Batch-Geolocation

XML Extensible Markup Language

OSM OpenStreetMap

CIM Common Information Model

ROC Receiver Operating Characteristic

ARENA Australian Renewable Energy Agency

AI Artificial Intelligence

GAN Generative Adversarial Network

KNN K-Nearest Neighbor

SVM Support Vector Machine

FFNN Feed-Forward Neural Network

RF Random Forest

DFT Discrete Fourier transform

SQL Structured Query Language

P-P Probability Plot

NAM Noise Addition Methods

MLDP Machine Learning based Data Perturbation

RMSE Root-Mean-Square Error

QGIS Quantum Geographic Information Systems

EWB Energy workbench

NB Naive Bayes

MOS Minority Oversampling

KL Kullback–Leibler

TN True Negatives

TP True Positives

FP False Positives

FN True Negatives

TPR True Positive Rate

AUC Area Under the Curve

Chapter 1

Introduction

1.1 Motivation

The electricity system is a national asset: it is large, complex, and interconnected. Currently, electrical distribution networks are transforming, and distribution network operators are confronted with three crucial challenges: the increasing penetration of renewable and distributed energy resources; weather instability and storm effects; network and data integrity [2]. Of the three, network and data integrity are the primary concerns for grid operators, owing to the proliferation of new technologies and increasing amounts of data generated by intelligent devices [3]. New technologies are being introduced to upgrade the aging infrastructure of power distribution networks. However, the ability to test and analyze the new solutions and developments in existing networks is limited by a scattered and somewhat incomplete set of public test systems. Privacy concerns in network data and lack of network visibility significantly impede network updates and improvements. There is a scarcity of publicly available networks for use as test networks. This need has prompted the research community and industry users to develop and innovate synthetic test systems that are adapted to modern distribution networks. However, creating such networks and datasets is difficult due to increasing constraints on power distribution networks, such as the continuous expansion of networks, the integration of new low carbon technologies, and the significant penetration of renewable energy resources in the network. A challenge is posed to develop feasible methods for creating synthetic networks and datasets to meet the evolving problems and opportunities, a challenge that this thesis responds to.

1.1.1 Potential of synthetic test networks and datasets

The benefits of synthetic datasets and test systems are becoming increasingly apparent. According to a recent report [4], the value of the Australian government open data is $25 billion per annum, which illustrates the importance of the data created during research. The report shows that a relatively small investment in a combination of data generation and infrastructure development provides a significant increase in value for innovation, research, industry, and the broader economy. Synthetic networks and datasets are at the forefront of the vision to address the challenges of transforming the operations and improving the efficiency and reliability of modern distribution networks.

1.1.2 Can network data from energy companies be used?

Although distribution network data is known to distribution system operators (DSOs), very seldom the topology and details of networks are available for research to stimulate innovation. Due to privacy and confidentiality concerns, electrical utilities refuse to share sensitive network data, which, in turn, affects the validation of new concepts in existing networks. Even if they disclose at some level for testing purposes, privacy regulations and market competition from energy traders hamper this pursuit. This situation is exacerbated further by the large amounts of uncontrollable renewable energy sources connected to distribution systems. In some cases, data is accessed only through strict privacy agreements. However, data-sharing agreements are very slow, taking months on average to establish [5]. As a result, many promising collaborations can fail even before they begin.

A potential solution to address the data sharing problem is the data protection frameworks. The data protection approaches enable open data sharing to remove the data access barriers between academic and industry users. This is beneficial since many applications that require data sharing in distribution networks are still impractical due to the lack of sufficient datasets, privacy concerns, and data access problems. Motivated by these factors, the first part of this thesis examines data protection frameworks. The aim is to propose novel data protection solutions that anonymize sensitive data through data synthesis, offering an effective solution to facilitate open data sharing. A collaboration is established with a distribution network service provider (DNSP) to access the actual distribution network data and evaluate the proposed solutions under different scenarios.

1.1.3 Network synthesis and the role of buildings

Building load profiles have a substantial influence on distribution network operations [6]. To accurately design the synthetic networks, it is important to understand the load profiles of buildings [7]. A distribution network is comprised of a diverse pool of grid components. The distributed energy sources, the energy users (buildings), and the distribution grid need to work together to optimize the system operation. As a result, the system becomes more complex and sophisticated. A classic example of this is the emerging concept of building-to-grid (B2G) [8], which provides seamless interactions between buildings and evolving grid infrastructure. Buildings are crucial part of power distribution networks and play a central role in network operations. Their increasing percentage of energy consumption presents new challenges for distribution network operators. With the extended COVID-19 lockdowns in various countries, the energy demand of buildings is changing, and there is a global debate about stabilizing energy demand, recovering energy-economic resilience, lessons learned, and new prospects. With the fast deployment of smart meters, huge amount of data is generated from network devices. The increasing availability of data collected from smart meters in buildings provide a valuable opportunity for research and development to understand the energy consumption profiles.

A careful analysis of the building load profile is important for energy management and conservation [9]. In this context, data analytics techniques such as deep learning can play an important role in performing benchmarking, enhanced analysis, and classification of buildings based on their energy usage. The last part of this thesis investigates data-analytic techniques based on deep learning to explore and classify energy consumption behaviours based on building type for energy efficiency programs.

This thesis has the three objectives of developing: feasible methods for modeling synthetic networks and datasets; data protection frameworks for confidential network data; and a deep learning solution for classifying the electric load profiles of buildings in distribution networks.

1.2 Background

Based on the abovementioned three objectives, this section explains the background knowledge of the topics and examines their current progress and critical challenges.

1.2.1 Synthetic networks and global debate

In 1991, for the first time, the Institute of Electrical and Electronics Engineers (IEEE) Power and Energy Society (PES) published four reference test systems for the United States (US) distribution system [10]. Now, 30 years later, the number of published test systems is still limited or only applicable to specific geographical regions. Several organizations have contributed to the development of synthetic test systems for scientific analysis and improvements in modern power networks, as shown in Figure 1.1. The IEEE PES released transmission and distribution network datasets in [10]. A group of researchers at Texas A&M University provided synthetic transmission network datasets to study transient stability [11]. The goal of this study was to design dynamic cases for synthetic network models. The Electric Power Research Institute (EPRI) developed synthetic test cases for grid analysis, grid planning, and grid operation management [12]. Their datasets were based exclusively on geographical regions in the US. In [13], a comprehensive illustration of the transmission system network operated by members of the European Network of Transmission System Operators (ENTSO) was presented. Pacific Northwest National Laboratory (PNNL) provided benchmark systems for small-signal stability analysis and control [14]. Scientific GRID (SciGRID) released transmission network datasets for European energy networks. The United Kingdom Generic System (UKGS) provided

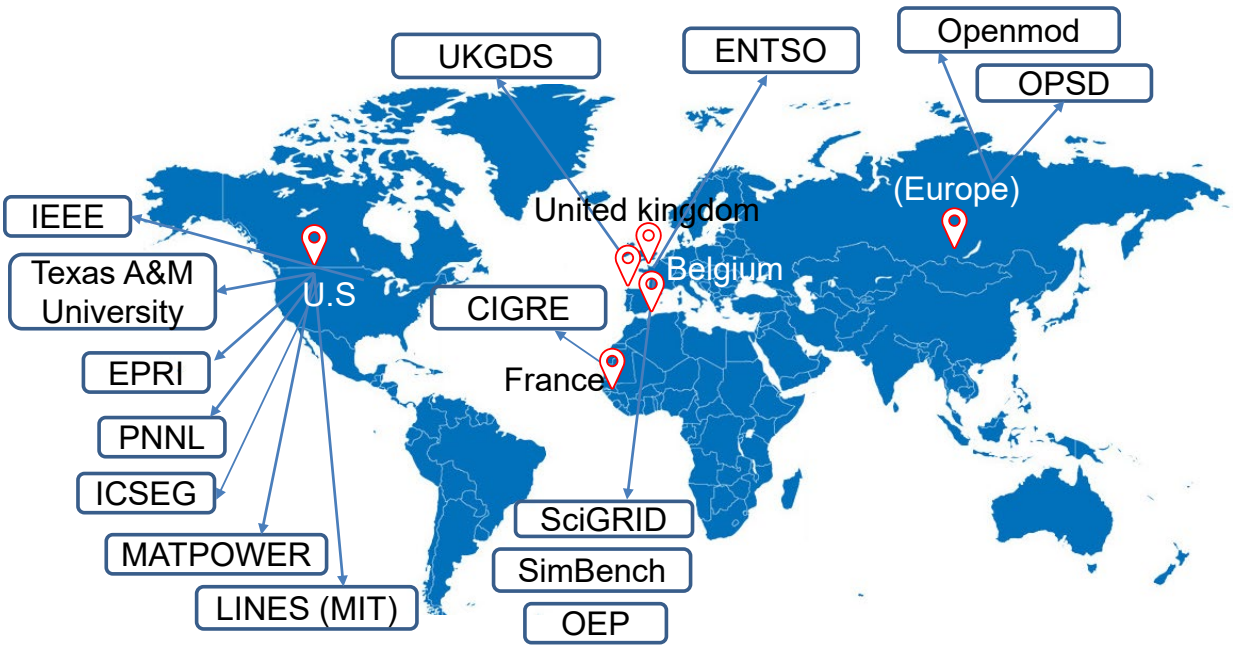


Figure 1.1: Participation of global organizations in developing synthetic test systems

network datasets representing the UK power system [15]. CIGRE published a database including a collection of power system test cases to compare solutions for electromagnetic transients (EMT) [16]. Similarly, Open Energy Modelling (OPENMOD) [17], the Open Power System Data (OPSD) [18], the Open Energy Platform (OEP) [19], the Illinois Center for a Smarter Electric Grid (ICSEG) [20], MATLAB-based power system simulation package (MATPOWER) [21], and the Laboratory for Intelligent Integrated Networks of Engineering Systems (LINES) [22] produced synthetic datasets for the development of novel solutions. In [23], a comprehensive dataset known as SimBench was presented for grid-related solutions. It limited to specific geographical areas, such as Central Europe and Germany.

1.2.2 Data protection frameworks

According to recent research [8], the worldwide market for distribution grid data analytics is estimated to reach US$4.6 billion by 2022. The volume and complexity of data generated in the distribution environment necessitate the development of novel ways to data synthesis, sharing, and visualization. A key element is data privacy that must be addressed before the data can be shared safely.

The confidentiality of critical network data can be achieved using data anonymization methods. There are five techniques for data anonymization that includes generalization [24], suppression [25], perturbation [26], permutation [27], and anonymization through data synthesis [28]. Another popular method is encryption and decryption techniques to preserve data privacy [29]. Cybersecurity techniques have also drawn increasing attention in both academia and industry. False data injection attack methods are used in cybersecurity for protecting data [30]. In [31], extensive research and analyses of data privacy concerns in a distribution grid were presented. The resultant survey [31] revealed the increased need for data protection schemes for distribution networks in recent years. This is due mainly to the geographical distribution of devices and distributed data acquisition for the economical operation of a/the grid. Despite the effectiveness of existing methods, there remain open questions that need further investigation. For instance, a trade-off between privacy preservation and data utility is essential as the loss of crucial information renders the data meaningless. A classic example in [32] demonstrates why it is crucial to preserve correlations between data attributes. Electrical utilities spend a significant amount of money on cybersecurity solutions which can be complex and ineffective, as

indicated in recent literature [33, 34]. There is a need for the simple and user-friendly techniques to reduce the gap between academic research and industry users [35]. Also, there is no standard criteria that define the representativeness of anonymized datasets [36]. For instance, anonymized datasets should be representative of the original ones. This thesis aims to address these challenges by proposing novel solutions and validating them using an actual network.

1.2.3 Network synthesis and buildings

In distribution networks, buildings serve as end-connection points. Figure 1.2 shows an overview of a typical distribution system with buildings. The network is managed by a distribution system operator (DSO), and there are end-users (buildings) that generate data at an unprecedented rate. The data of building load profiles contain meaningful information that enables data-driven decisions such as real-time monitoring of end-users energy consumption. Since loads in buildings change throughout the day, the difficulty is transforming highly heterogeneous datasets into actionable outcomes. In this context, data-predictive analytics combined with visualization can lead to better predictive decisions and situational awareness.

The current trend in building energy modeling is changing from conventional physics-based modelling [37] to data-driven techniques [38]. Physics-based modelling needs en-

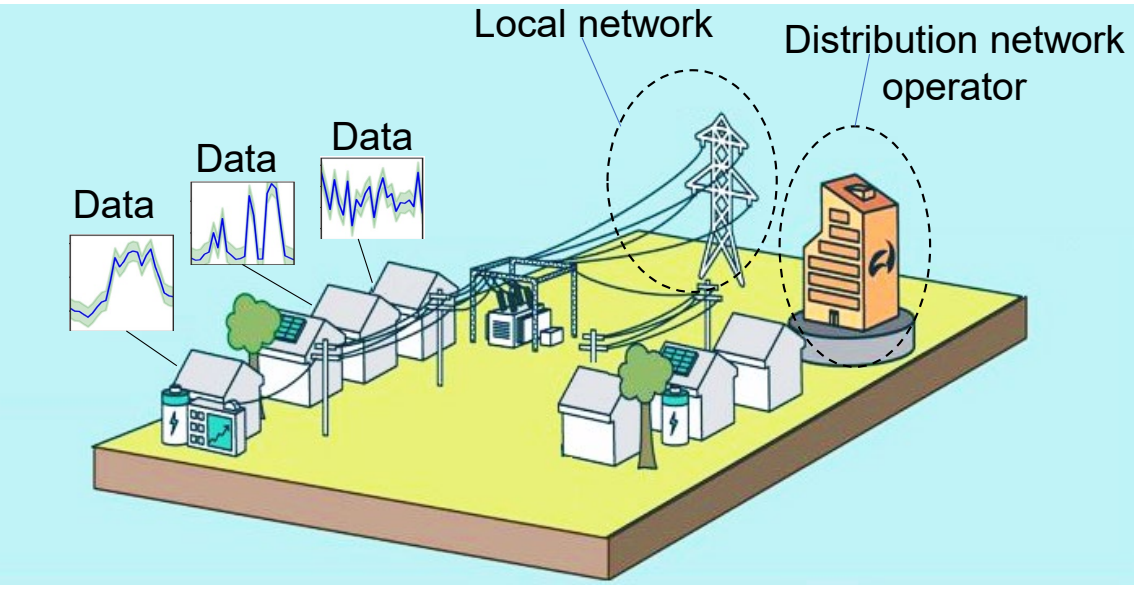


Figure 1.2: Network to buildings operations with data from building load profiles

gineering expertise and effort (with a related cost), which is a major bottleneck for the feasibility of predictive models [39]. Also, benchmarking methods, including clustering algorithms (k-means) [40], fuzzy clustering algorithms [6], and hierarchical clustering algorithms [41], are used to covert the building load profiles into decisions. The k-means algorithm was used with success for the classification of building profiles. However, these methods mostly used manual procedures in clustering analysis. For instance, manual labeling of building load profiles for predictions, hand-designed indicators (features) such as the magnitude of load profiles, and prior assumptions for selecting a suitable number of customer groups in the clustering analysis. The manual procedures are not applicable in real-world settings because buildings exhibit significantly different patterns in their energy consumption. Manual selection of features (key indicators) from building load profiles is not possible due to the vast amount of data generated on daily basis.

The load profile data presents opportunities for insights and improvements. Data-driven techniques are now the most suitable options due to their user-friendly implementations and good prediction performance. Recently, machine learning and deep learning techniques have drawn an increasing attention due to their capability for handling large amounts of generated data, providing new solutions and algorithms to address technical challenges [38]. The most promising machine learning applications in the energy domain are the prediction of building energy demand [42], operational optimization of buildings [43], the evaluation of occupants influence on building energy consumption [44], detection and commissioning of building equipment operational status and failures [45], energy benchmarking analysis [46], and the characterization of building energy demand [47]. This thesis investigates the classification of electric load profiles of buildings into three types of buildings: residential, commercial, and industrial by developing a novel deep learning framework. Classification of load profiles from all sectors (e.g., residential, industrial, and commercial) will benefit grid services, minimizing energy loss and assisting in peak shaving [48]. The classification of load profiles of buildings not only provides reliable energy benchmarks [49], it also helps in the creation of demand response programs and energy management initiatives [50].

1.3 Challenges and research opportunities

Based on the above discussion, the current challenges are summarized below:

- The lack of network visibility and data constraints are the major problems in developing innovative strategies for improving current power networks. Figure 1.3 illustrates the present issues associated with network modernization. The absence of synthetic test systems and network visibility makes it difficult for researchers to understand where restrictions exist in networks or are likely to arise in the future [51].

- Existing solutions mainly focused on synthetic transmission systems. The difficult problem of developing synthetic LV distribution networks was not addressed.

- An important distinction of a distribution network that was not adequately addressed in previous studies is geospatial information of network elements. The significance of geographic information is suggested in [52, 53]. Geographic information is crucial for the planning, modeling, and management of distribution network assets. For instance, the geographical locations help planners to install new assets in existing networks.

- There is no standard criteria and guideline for establishing electrical connectivity between network components. The difficulty is to develop a method that meets the stan-

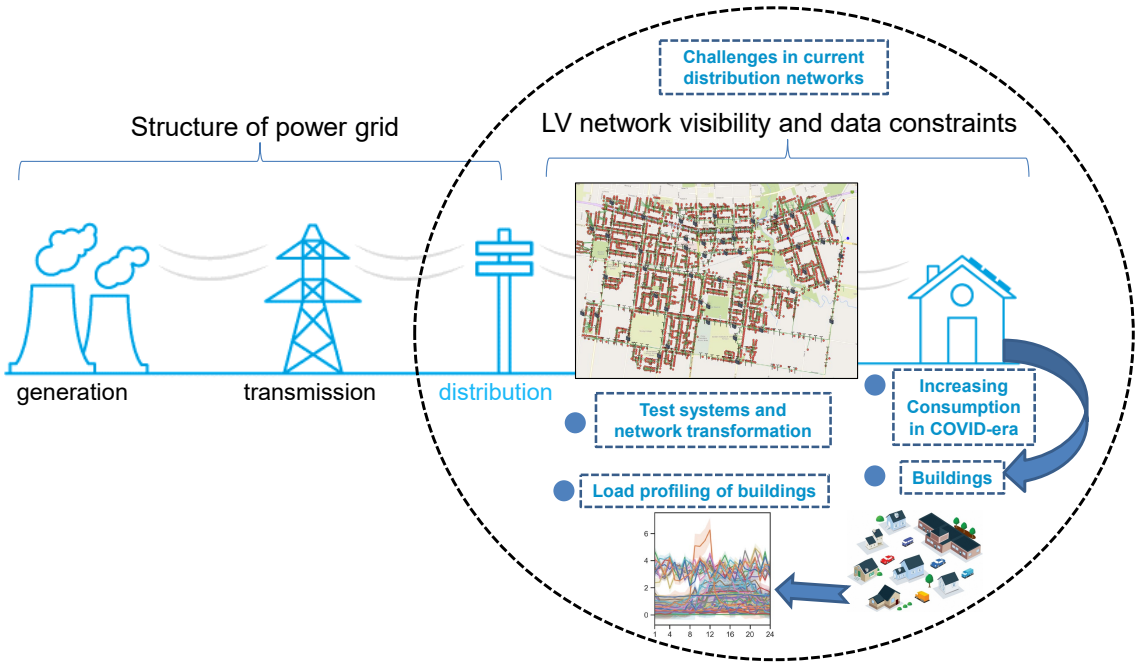


Figure 1.3: Challenges of transforming existing networks

dards of distribution network operators for practical implementation. Also, the power lines in a network are added to the datasets manually, making them unsuitable for real-world applications.

- Most current solutions were designed for European and US networks, with no test cases or data representations developed for Australian ones. Different from existing literature, this thesis focuses on Australian distribution networks.

- There is no investigation in the current literature that confirms the validation of synthetic network and datasets by replicating them on real-world industrial servers. The study in [1] validated its methods through statistical, and operational analysis. It is essential to validate the developed techniques with industrial tools to ensure that the obtained synthetic distribution test systems are similar to the real ones.

- Another critical aspect that is not adequately handled for the existing networks is energy consumption in buildings. Buildings are an integral part of distribution networks, and their increasing share of energy consumption presents new challenges for distribution network operators, particularly in the current COVID-19 environment, where work from home, total time spent inside buildings, and indoor comfort demands are significantly increased. In this regard, data analytic techniques such as deep learning can help improve building energy efficiency and demand flexibility by classifying building load profiles according to their energy usage.

1.4 Contributions and solutions

Based on the motivation and research opportunities previously discussed, the following contributions are made in this thesis.

1. Anonymization of network data to synthesize similar systems.

    - A novel statistical distribution and parameter estimation approach is proposed for anonymizing distribution network data. An algorithm based on the maximum likelihood estimate (MLE) is proposed to estimate the statistical distribution parameters that represent the actual data. Then, a data anonymization algorithm that uses the estimated parameters to generate realistic anonymized datasets is developed.

- Hypothesis testing based on Kolmogorov-Smirnov (K-S) test is conducted to make the datasets realistic
- A practical demonstration is provided by obtaining actual (real) datasets from a local distribution network provider. These datasets are then anonymized and a comparison is made between actual and anonymized methods.
- Validation on benchmark test systems such as the IEEE 123-node test system.
- The methodology is experimentally proven by comparing it to the benchmark data anonymization methods in the literature.
- The solutions provided alleviate network visibility concerns by introducing a new, open, and flexible data anonymization framework.

2. Modeling synthetic networks and datasets from open-source public databases and data synthesis algorithms with industrial validation

   - Data synthesis algorithms are proposed to obtain the network datasets for the power distribution system. For the first time, the topology of power distribution lines is developed using public road infrastructure. The proposed method simplifies the design of power lines by using the concept of nodes and edges.
   - A new method for identifying the number of energy consumers supplied by a transformer in a distribution network is developed. For example, the hub-line algorithm is proposed that connects the energy consumers based on their nearest spatial distance to a transformer. A standard cut-off distance from the transformer to households is maintained by adhering to the guidelines in CIGRE publications [54].
   - A distinctive characteristic of distribution networks that has not been properly addressed in existing studies is the geographical structure of the system. This thesis includes the geospatial locations of network elements by implementing a batch-geolocation (BGL) algorithm.
   - A standard way of representing electrical connectivity between two or more equipment's in the network is provided by proposing the fromNode and ToNode concept. Connectivity is established by defining the start and end points of the network's elements.

- A new way of obtaining substation data is developed by creating overpass XML queries in OpenStreetMap (OSM) databases.
- A technique for ingesting synthetic data to industrial data platforms is developed. The developed method converts synthetic data into a Common Information Model (CIM)-based format, which is a widely used data standard in the power industry [55]
- The practical feasibility of the proposed algorithms is demonstrated by an illustrative case study of the Colac region in Australia. Synthetic datasets are created for the distribution feeder, and the datasets are deployed in the industry servers. The results are then compared to the original feeder datasets to verify the applicability of the proposed techniques.
- The solutions are tested using a two-step validation process. In the first stage, solutions are validated by replicating them on real-world industrial servers, and in the second stage, solutions are verified using expert comments and validation. This method contributes to expanding the utility of synthetic networks and datasets beyond university researchers to industry users.
- Interactive maps are created to visualize the synthetic network and datasets. This allows users to manage the key assets in an existing power distribution infrastructure.

3. Developing an AI-based based multi-stage transfer learning for building load profile classification in the COVID-19 era.

    - A deep learning solution is proposed for addressing the problem of electric load profile classification in the context of buildings.
    - Two new methods based on sparse autoencoders (SAEs) and the multi-stage transfer learning (MSTL) are proposed. Different from conventional handcrafted feature representation, SAEs can learn useful features from a large number of buildings data in an unsupervised automatic way. This is important since each building has unique electrical load patterns, and manually extracting the key features of every building is not possible in practical situations. A MSTL approach is applied to enhance the classification accuracy by combining sequential unsupervised and supervised learning.

- A minority oversampling (MOS) algorithm is proposed that effectively balances missing or unbalanced data by equalizing minority and majority samples for fair comparisons.
- Two case studies are presented to validate the methodology. In case study 1, the techniques are evaluated on public benchmark datasets of buildings. In case study 2, the results are validated using real-world datasets of 105 buildings (35 residential, 35 commercial, and 35 industrial).
- An empirical comparison is made with the benchmark methods in the literature. A performance improvement ranging from 1 to 10 percent has been achieved. Standard performance criteria such as a confusion matrix, receiver operating characteristic (ROC) curves, recall, F1-score, specificity, and precision are used to compare the findings. For a fair assessment, an average percentage performance improvement obtained by the proposed method over traditional methods is computed.

1.5 Thesis organization

The structure of the thesis is shown in Figure. 1.4, and is organized as follows.

- Chapter 2 provides a comprehensive overview of previous works in the area of data anonymization, synthetic network modeling, and benchmarking practices for load profiling classification of buildings in distribution networks. The proposed methods for solving this problem are described.
- Chapter 3 presents the details of the developed data protection framework for anonymizing network data. The strengths and limitations of previous works are reviewed, and the advantages of this research are briefly summarized.
- Chapter 4 describes the proposed methods for synthetic networks and datasets. A case study of the Colac region in Australia is presented. Also, details of the experimental validations of the proposed methods on industrial servers, such as the evolve ARENA data platform, are provided.
- Chapter 5 details the procedure for developing AI-based SAEs and the MSTL framework. It explains how the unsolved problem of building load profile classi-

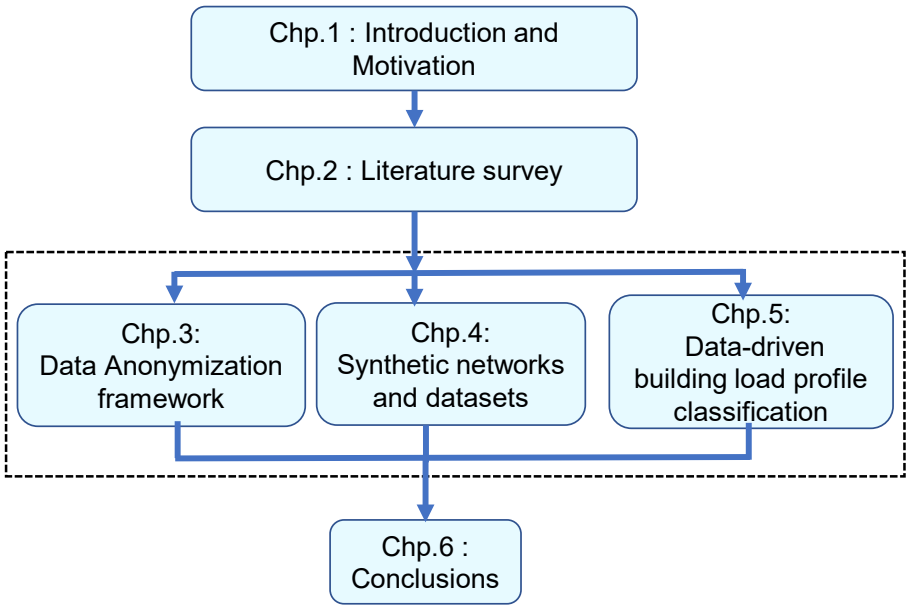


Figure 1.4: Structure of the thesis

fication in the distribution network is addressed. The practical feasibility of the proposed approach is demonstrated by presenting two case studies.

- Chapter 6 summarizes the findings, and recommendations for future research directions are presented. The practical challenges considered in this thesis and recommended future extensions of this study are provided.

Chapter 2

Literature review

The work presented in this chapter is published in the following article:

1. **[Journal]** M. Ali, K. Prakash, and HR. Pota, "Intelligent energy management: Evolving developments, current challenges, and research directions for sustainable future," *Journal of Cleaner Production*, Sep. 2021. **IF: 9.297**

Summary: *This chapter provides the essential background that links this work to other studies in the literature. A literature review of existing studies is conducted to justify the need for this research. A novel strategy for systematically surveying relevant studies by converting the scattered literature into visual presentations is proposed. The process begins by identifying relevant studies from high-impact journals and filtering them based on their relevance. Using VOSviewer analysis, the relevant literature is transformed into visual representations. This analysis identifies the current research gaps and explores existing issues, methods, and findings. It also critiques existing countermeasures and their limitations.*

2.1 Introduction

In Chapter 1, the motivation for using synthetic networks and datasets is discussed. There have been significant developments in the field of power distribution networks, with various methods and new solutions proposed for creating synthetic test systems. An important issue is to determine key insights from the scattered academic literature in the age of digital publishing. In this Chapter, a review using novel techniques that provides unique insights into the literature is conducted to identify research gaps.

2.1.1 VOSviewer approach to identify key research gaps

VOSviewer provides clustering solutions with different colors to indicate the most important occurring topics in scarce literature. In the age of big data, scholarly information usually contains thousands of raw data, such as papers, books, or scientific reports. Analyzing hundreds of scientific publications can be a tedious and troublesome task. One way to overcome this problem is to introduce clustering solutions. Clustering techniques identify related publications or journals by clustering each publication and developing a citation network. A total of ninety-one papers published in high-impact journals are identified based on their scientific soundness and relevance, and a VOSviewer analysis is applied. The findings of collected articles are then reported with VOSviewer experiments.

In a clustering technique, publications are assigned to clusters by maximizing a quality function, as defined by Waltman [56] and Nees [57].

$$Q(x_1, \ldots, x_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} \delta\left(x_i, x_j\right) \left(a_{ij} - \frac{\gamma}{2n}\right) \tag{2.1}$$

where n represents the total publications, a_{ij} shows the relatedness of publication i with publication j. γ is a resolution parameter, and x_i denotes the cluster to which publication i is assigned. The term $\delta\left(x_i, x_j\right)$ equals 1 if $x_i = x_j$ and 0 otherwise. The relatedness of publication i with publication j is given by

$$a_{ij} = \frac{c_{ij}}{\sum_{k=1}^{n} c_{ik}} \tag{2.2}$$

where c_{ij} shows the citations of publication i and j. It will be equivalent to 1 when publication j refers to publication i or either publication i refers publication j. Otherwise, it is 0 . Thus, if there is a citation link between publication j and i, the relatedness a of publication i with publication j is equal to citations of publications i and j divided by the total number of citations of publication i.

2.2 Results and identified gaps

In this section, the key research themes identified from VOSviewer experiments are discussed. A network visualization map is constructed from the collected articles, and results are presented with the visual clusters. The visualization results obtained from

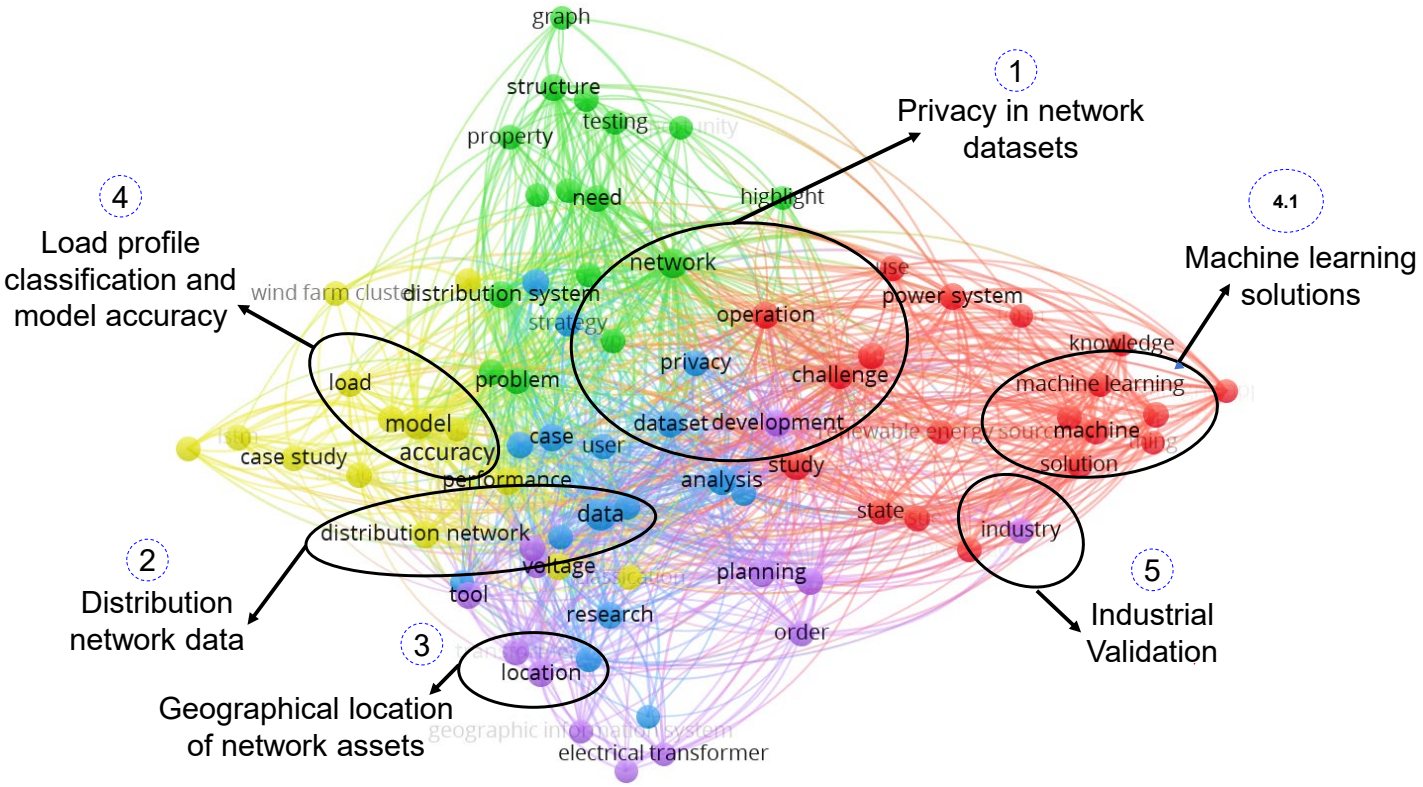


Figure 2.1: VOSviewer experiments to find key insights from highly cited papers

VOSviewer experiments are shown in Figure. 2.1. The key emerging areas are presented with six visual clusters in black circles. Clusters with different colors indicate the relatedness of topics and publications in the respective fields. The size of a cluster represents the number of publications that belong to each cluster. The colored lines connecting clusters show their relatedness, with line width reflecting the number of citations between clusters. The description of each cluster is summarized in Table 2.1. Cluster 1 (row 1) demonstrates that data development and network privacy are the primary problems for distribution network operators. Cluster 2 (row 2) indicates a lack of distribution network data for updating current networks. Cluster 3 (row 3) demonstrates the missing information about the geographical locations of network assets in existing network datasets. Clusters 4 and 4.1 (rows 4 and 5) show rising energy consumption in buildings and the need for load profile classification in energy management and conversion. This cluster also points out the problem of low classification model accuracy and suggests new and advance machine learning methods as a possible solution.

The VOSviewer experiments reveal the existence of three research gaps in the recent literature: data integrity and privacy concerns in distribution network datasets; lack of

distribution network datasets to test new solutions in existing networks; and the impact of load profiles of buildings (end-users) across power distribution networks. In the next subsections, the breakthroughs of previous research contributions on these topics, as well as the main challenges and potential solutions, are explained. The aim is to identify research gaps and opportunities and provide novel insights from scientific and practical perspectives.

Table 2.1: Key research words in terms of clusters

Total Research works	Cluster	Observed keywords	Explanation
91	①	dataset, development, privacy, network, operation, challenge	Privacy in the network dataset
	②	distribution network, data	Distribution network data
	③	location	Geographical location of network assets
	④	load, model, accuracy	Load profile classification of buildings and accuracy of the model
	④.1	machine learning, machine, solution	Machine learning solutions for load profile classification

2.2.1 Literature review on data protection frameworks

This subsection explains the contributions of past research in data protection frameworks to address data confidentiality challenges in distribution network datasets. There are five fundamental techniques for data anonymization, which include generalization [24, 58], suppression [25, 59], perturbation [26, 60] permutation [27, 61], and anonymization through data synthesis [28, 62]. Different approaches have been thoroughly investigated to mitigate the privacy risks in data. [63, 64]. In [62], a machine learning approach to preserve the privacy of data with a utility function of electronic health records of hospital patients was proposed. The authors in [65] proposed a statistical clustering procedure to create prototypical feeders without revealing private information. Its limitation is that it requires manual selection of the parameters for each cluster, which is not possible in practical situations. In [66], an information-masking mechanism was proposed for hiding the original information by transforming it into another form. The proposed solution obfuscates the targeted information by adding additive noise to the sensitive parts of the data. In [25], the perturbation approach was used to anonymize the key attributes of data by adding noise to the datasets. However, this technique fails to retain the key statistical characteristics or trends of the data, resulting in the loss of critical information. The

inclusion of noise can generate fake trajectories in the data that do not correspond to the realistic scenarios [67].

Data masking techniques such as suppression [25] are also implemented to ensure privacy. Often, the simple masking or removal of identifiers may not be sufficient to ensure privacy [68] and masking reduces the amount of information available by suppressing some of the data or decreasing the level of its details. In [69], a data encryption system based on cryptographic techniques was proposed. The cryptographic techniques used to prevent information leakage are computationally expensive [70, 71]. A data aggregation scheme based on local differential privacy was provided in [72]. A lightweight privacy preserving data aggregation technique to address the problem of complexity and computation costs in existing methods was proposed in [73]. Other common approaches for data anonymization are k-anonymity [74], l-diversity [75,76], and t-closeness [77,78] for data anonymization. The random-data perturbation techniques do not entirely protect privacy [68] while the t-closeness limits the amount of useful information released and destroys the correlations between key attributes and confidential attributes [75].

A multi-level reversible data anonymization technique for obfuscating the sensitive parts of documents was presented in [79]. The authors developed a data hiding technique by using a compressive sensing theory. In [80], an entropy-based measure for quantifying the real privacy provided by anonymous privacy-preserving smart metering methods was proposed. A concept of the generative adversarial network (GAN) to anonymize the sensitive data was proposed in [81]. Despite their effectiveness, GANs have limitations, including instability and mode collapse during model training, which significantly impact the quality of anonymised datasets [82]. In [83], a high-degree noise addition method for enhancing data privacy was provided. An innovative differential privacy algorithm for ensuring the privacy of the smart meter data in power distribution networks was proposed in [84]. The impacts of the proposed algorithm on the operations of distribution grid data were thoroughly investigated. A Gaussian noise approach based on artificial intelligence for maintaining data privacy was proposed in [85]. A machine learning model was trained to ensure both demand-side management and consumer data privacy. In [86], a unique decentralized privacy-preserving technique for anonymizing sensitive information from distribution network data was presented.

2.2.1.1 Findings and research gaps:

The literature presents many solutions and recommendations. However, there exist open questions that need further investigation, with the most important listed in Table 2.2. This chapter proposes a data anonymization through data synthesis to address the common limitations of traditional anonymization methods. The frameworks and objectives of several works in the existing literature are quite different from those of ours. The goal is not to implement encryption and decryption techniques, rather a simplified anonymization methodology for addressing the problem of data sensitivity in critical power infrastructures. The idea is to search for statistical patterns that emerge in the data of distribution feeders and their properties and use them to synthesize similar systems. Using the proposed solutions, electric utilities and academic researchers can generate anonymous datasets with minimal expert knowledge. Traditional methods mainly focused on swapping and noise addition techniques to de-identify sensitive data. However, recent research has shown that de-identification is not sufficient for preserving privacy in data [62]. With technological advancements and computer experts, it is easy to re-identify (reverse engineer) hidden information [87]. In the case of noise addition, it is difficult to preserve the main statistical characteristics or patterns of the data, resulting in a loss of pivotal information [88]. Instead of de-identifying private information through swapping and noise addition methods, there is a need for techniques that mimic the properties of original datasets, which is the motivation for this study. In this thesis, an attempt is made to reduce the gap between academic research and industry regarding privacy concerns, as many collaborative projects are based on sensitive and privacy-encumbered data.

2.2.2 Literature review on synthetic test systems

Synthetic power networks are fictitious test cases created for research, development, and demonstration purposes. Their distinguishing feature is systematic validation to verify that they accurately replicate the properties of actual grids. Free from confidential data, the synthetic test cases can be widely shared and published. In the existing literature, there are research efforts to create synthetic networks and datasets for distribution networks [91, 92]. The first distribution dataset for the investigation of radial distribution feeders was published in [93]. A taxonomy of prototypical radial electrical distribution feeders for analyzing smart grid technology models was provided in [94]. The test dataset

Table 2.2: Summary of knowledge gap in the existing literature

Recent studies	Findings
Yoon et al., 2020 [62]	• Most existing solutions focus on data-masking or hiding sensitive data without considering the key relationships or statistical patterns in data. This leads to the loss of important information and a decrease in data usability [79, 89]. The addition of new techniques considering statistical patterns will improve the reliability of data utility. • Electrical utilities spend a significant amount of money on cybersecurity solutions which can be complex and ineffective, as indicated in recent literature [33, 34]. How can data privacy and security be protected in the absence of expert knowledge of cybersecurity techniques? • Unavailability of simple and user-friendly techniques that can reduce the gap between academic research and industry, as suggested in [35]. Collaborative projects are based on sensitive and privacy-encumbered data and many promising collaborations can fail even before they begin due to data-sharing problems. • There is a lack of hypothesis testing to confirm the representativeness of anonymized datasets [36]. For instance, anonymized datasets should be representative of the original ones.
Xin et al., 2018 [66]	
Shaham et al., 2020 [67]	
Bassoo et al., 2019 [68]	
Belguith et al., 2019 [70]	
Minello et al., 2020 [74]	
Raymond et al., 2021 [75]	
Yamac et al., 2020 [79]	
Zang et al., 2020 [85]	
Dondeti et al., 2021 [88]	
Yan et al., 2022 [90]	

was created for several regions in the United States. A fictitious synthetic power system network to capture the functionality, characteristics, and topology of the actual system was created in [95]. In [96], a synthetic electric grid was created to consider the reactive power planning requirements. A methodology for bus level static load modeling in synthetic electric grid test cases was presented in [97].

A clustering approach for creating synthetic power system test cases was proposed in [98]. The methodology demonstrated various real-world examples; however, the concept was built for synthetic transmission systems. A random topology method to generate synthetic data for power transmission grids was proposed in [99]. Synthetic electric grid datasets describing high-voltage transmission grid was created in [100]. The datasets covered the large geographic regions of North America. A method of generating realistic power system steady-state scenarios using synthetic transmission networks and time

series was developed in [92]. Their findings indicate that specialized strategies are required to accelerate the convergence of a power flow solution and to prevent low voltage solutions. In [101], complex-network techniques were presented to generate synthetic transmission networks for European grids. The topological properties of high-voltage electrical power transmission networks were studied in [102]. A two-phase methodology for creating synthetic transmission grid was proposed in [103]. A comprehensive cyber-physical model for a synthetic electric grid was created in [104].

Recently, distribution systems have been regarded as random graphs for constructing synthetic distribution datasets [105], with the findings tested by Monte Carlo simulations. In [106], a bottom-up approach was used to develop a synthetic distribution dataset. A conic-based optimization model for developing a synthetic medium-voltage network for a geographical region of Singapore was presented in [107]. The problem of phase-selection in synthetic systems was addressed in [108]. This study proposed two algorithms that determine the number and sequence of phases at each feeder section. The methodology in [109] integrated dynamic models for renewable resources into synthetic electric grids. In [110], a methodology was proposed to synthesize and validate bus-level load time series in the existing synthetic power systems. The need for synthetic representative networks and the challenges faced by distribution utilities were briefly discussed in [111].

National and international organizations have contributed to the development of synthetic test systems and datasets for scientific analysis and improvements in existing power networks. Table 2.3 summarizes current test systems, and their common limitations. For decades, researchers have used a limited set of standard networks such as IEEE transmission and distribution test cases [21, 94, 112]. The major concern is the restricted number of standard test systems that are only suitable for specific regions [48]. In [11], the Texas AM University researchers published synthetic transmission network datasets to explore transient stability. The goal of this study was to provide dynamic examples for synthetic network models. EPRI created synthetic test cases for grid analysis, grid planning, and grid operation management [12]. These datasets were created solely for geographical areas in the United States. A comprehensive illustration of the transmission system network operated by ENTSO members [13]. PNNL provided benchmark systems for small-signal stability research and control [14]. SciGRID published transmission network datasets for European energy networks [113]. UKGS offered network datasets describing the UK power system [15]. CIGRE developed a database including a variety of power system

Table 2.3: Summary of knowledge gap in the existing literature

Network data repository	**Findings**
IEEE Test Cases [10]	• Mainly focused on synthetic transmission systems. • Topological network data, including geospatial information, is not included. • Designed primarily for North American electrical distribution networks, which differ significantly from European and Australian systems. • There is no information provided for creating an electrical connection between network elements. • There is no standard guideline or strategy for designing power lines in a distribution network. In some cases, power lines are built using manual techniques that are not suitable for real-world applications. • Synthetic networks and datasets are created from paid softwares which are not easily accessible or reproducible for the research community. • Geographical verification and visualizations are not available to monitor the electrical assets in an organized manner. • The essential element of buildings (end users) in the distribution network and their increasing share of energy consumption in the post-COVID-19 period is not examined. • Validation of developed solutions using real industry servers is seldom provided.
Texas A&M University [11]	
EPRI [12]	
ENTSO [13]	
PNNL [14]	
SciGRID [113]	
OPENMOD [17]	
UKGDS [15]	
CIGRE [114]	
OPSD [18]	
OEP [19]	
ICSEG [20]	
MATPOWER [21]	
LINES [22]	
SimBench [115]	

test cases that may be used to compare solutions for electromagnetic transients [16]. Similarly, OPENMOD [17], the OPSD [18], the OEP [19], the ICSEG [20], Matlab-based power system simulation package (MATPOWER) [21], and the LINES [22] produced synthetic datasets for the development of novel solutions. In [23], an extensive dataset known as SimBench was released for grid-related solutions. The created dataset was restricted to specific geographical regions, such as Germany and Central Europe.

2.2.2.1 Findings and research gaps:

Despite the proven success of previous works, some questions remain unsolved, as described below:

- *Limited to power transmission network:* Previous works [116–118] mainly focused on generating synthetic transmission networks, where the effects of individual customer decisions were neglected. Compared to the transmission network, research on synthetic distribution network and datasets is still at preliminary stages. The development of synthetic distribution test cases with more detail, including geographical information will contribute to this topic and enable cross-validation of developed techniques.

- *Missing geospatial information:* Topological network data, including geospatial information, is not included. Conventional design techniques focused primarily on reducing costs such as investment, operation, maintenance and energy losses, while ensuring the safety and reliability of the network. Geographical information regarding distribution networks, such as the location of transformers, substations, power lines, and energy customers is critical for managing distribution network assets. Geographical information gives a valuable topological representation of the network layout. In the event of a network failure, these topological aspects have a substantial impact on network expansion planning and future modernization techniques.

- *Differences in network topologies and appropriate representations*: Most of the existing network datasets were developed for American and European systems. Distribution systems around the world are characterized by different standards and their topologies are constituted by different ranges of line lengths and different transformers that deliver electricity to consumers. A clear difference between European and U.S. distribution systems is the number of phase connections and low voltage power lines that link each power transformer in the street, as shown in Figure 2.2. In Europe (a), all the systems are three-phase (represented by black). In the USA (b), primary feeders are made up of one and three-phase sections that deliver electricity to specified coverage regions. In the USA, the number of consumers served by a low-voltage distribution transformer is considerably smaller than in Europe. For

Figure 2.2: Differences in distribution networks (a) Europe (left) and (b) US (right) [1]

this reason, low-voltage network lengths in the USA are shorter than those in Europe. Considering these differences in networks, a representative dataset is essential for different regions.

- *Electrical connectivity between network elements:* There is no standard way defined to create electrical connection between network elements. The term standard relates to how well it satisfies the needs of distribution network service providers. In [119], a minimum spanning tree solution was proposed to establish electrical connectivity in synthetic transmission network. The solution was not tested in practical environments such as industry servers and neither followed the industry standards or even examined by industry experts to validate its practicality.

- *Standard guideline for creating power lines in a network:* There is no standard guideline or strategy for designing power lines in a distribution network. In some cases, power lines are constructed using manual techniques that are unsuitable for real-world applications. For example, in [120], a random electric topology was proposed to create power lines using the random set of transmission gauge ratios.

- *Datasets for specific applications:* As most network datasets were designed for addressing a specific technical or economic operational issue, they were often insufficient for use in other kinds of applications or problems due to a lack of relevant information. For instance, the test network provided for the distributed generation protection analysis [121].

- *Scalability:* A network scale is crucial as it affects the validation of the overall network [1]. Currently, test datasets are only accessible for limited segments of

distribution networks, such as a single feeder. The largest IEEE test system, for instance, has just 8,500 nodes [122]. However, real-world systems contain millions of nodes. Detailed descriptions are required to assess the efficacy of large-scale solutions provided by new algorithms.

- *Validation and metrics:* In the extant literature, no research has validated its approaches using real-world industrial servers. The study in [1] validates the method through statistical, and operational analysis. It is essential to validate the developed techniques with industrial tools to ensure that the obtained synthetic distribution test systems are similar to the real ones.

2.2.3 Literature review on load profiling classification

Load profiles that are neatly presented and accurately classified are crucial for network planning, demand response, resource allocation, and load forecasting [7]. According to [6], electric load profiles assist utility businesses to develop better marketing tactics, increase energy savings, improve existing operational facilities and reduce forecasting mistakes. Three types of methods were used in the literature to create an effective and efficient load profiling classification. As shown in Figure 2.3, the methods include clustering techniques, artificial intelligence approaches, and time-frequency domain methods. Clustering techniques are widely used for load profile classification. In [123], a comprehensive review of clustering algorithms for classifying electric load profiles was provided. A k-means clustering method was proposed in [124] to classify the daily load profiles of academic buildings. A density-based spatial clustering method was reported in [125]. A clustering strategy based on a Gaussian mixture model for classifying building load profiles was provided in [126]. A shape-based clustering method for pattern recognition of residential electricity consumption was investigated in [127]. Recently, two classification studies utilizing k-means clustering approaches for the study of building operating data were published, discussing relevant applications in this area [124, 128]. Using clustering methods, it is difficult to specify the k-value at the beginning of investigations [129, 130]. Also, clustering techniques assume that each cluster has roughly equal numbers of observations in the dataset [131, 132]. However, in real life, residential buildings have more observations than industrial and commercial. The clustering algorithms classify the load patterns by grouping them into different clusters. Grouping load profiles into specific

clusters is ineffective for real-time evaluations since loads constantly change throughout the day.

Artificial intelligence approaches have also received increasing attention in load profiling classification. For instance, a support vector machine (SVM) approach for classifying the electric load profiles of buildings based on the climatic conditions and building characteristics was proposed in [133]. A classical feed-forward neural network (FFNN) model to investigate the load profiles of a commercial building was implemented in [134]. The results revealed that the FFNN model performed better than echo state networks on an aggregated load. In [135], an ensemble learning method based on a random forest (RF) was implemented for the characterization of non-residential buildings. This work [135] showed that feature selection is important for lowering calculation costs in large-scale deployments, minimizing model over-fitting, enhancing interpretability, and improving the accuracy. The concept of the k-nearest neighbor (kNN) was adopted in [136] to classify customer load profile. A model which was a variant of the decision tree technique was developed in [137]. The characteristics of residential building were classified with a precision of 82 percent and recall of 81 percent. Besides conventional artificial intelligence algorithms, more advanced technologies such as deep learning and adversarial learning can be used to achieve continuous breakthroughs in data resolution, learning, and computing ability, which will have broad application prospects in the research of electric

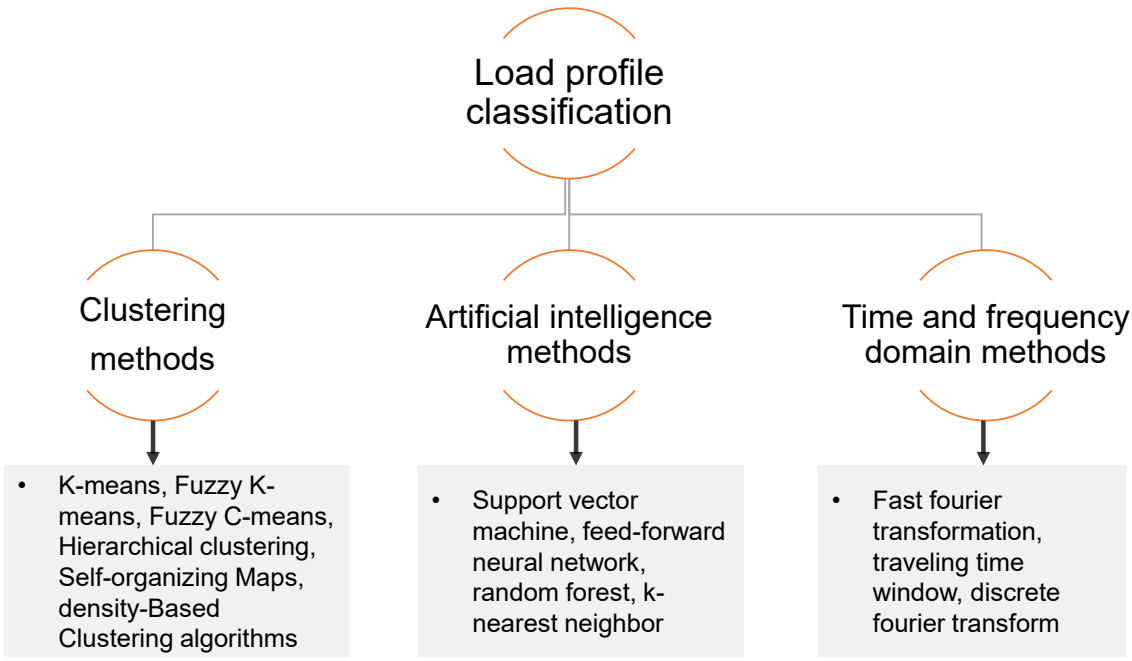


Figure 2.3: An overview of load profile classification methods

load clustering [138]. As complexity in load profiles increases, the incorporation of deep learning technologies in electric load classification will become increasingly important. It entails sophisticated real-time decision making that is backed up by data-driven models. Deep learning approaches have the potential to overcome the limitations of conventional electric load classification algorithms by providing more precise information about the load profile. The new techniques can help improve the operation of the distribution grid and the planning of future networks [138].

Load profiles were also investigated using time-and frequency-domain methods [50, 139]. These methods [50, 139] classify the electric load profiles by transforming the time-based load profile data into their frequency domain representation. The periodic patterns in the load profile were identified using signal processing techniques such as discrete Fourier transform (DFT) In [41], hierarchical classification of load profiles was proposed by using the load profile characteristic attributes in the frequency domain. In [140], household consumption load profiles were analyzed using Fourier and Wavelet transform. The authors in [141] proposed a frequency domain load profile descriptor for load profile characterization. In [142], a unique frequency–domain approach for classifying the typical load patterns of consumers was provided. This study [142] used harmonic analysis to extract key features from the load patterns. In [143], frequency-domain features were extracted using a discrete wavelet transform for classifying household load profiles. A load identification algorithm based on load decomposition was proposed in [144]. A comparative research of time-and frequency-domain algorithms for electric load pattern classification was presented in [145].

2.2.3.1 Findings and research gaps:

Following research gaps are observed during the review of related works:

1. Most of the efforts presented in the literature are devoted to developing classifiers (i.e., the pattern classification phase), often neglecting the role of the feature extraction process in learning important representations from the data. How to extract useful knowledge or features from the vast number of buildings data was not fully investigated [146]. For example, in [147, 148], the useful patterns or features such as magnitudes of daily load curves were extracted from data using manual techniques or general assumptions that are not applicable in real-world settings. In [41], load profile characteristics are studied using frequency and time-domain pattern charac-

teristics, such as base load, peak load, rise time, fall time, and high-load duration. Qualitative manual labeling is used to classify load profiles and to capture the periodic characteristics of load profiles.

2. The most widely used approach for classification in the current literature is k-means clustering. Despite the demonstrated efficacy of clustering algorithms, recent research [138] shows that these methods have three major drawbacks: it is difficult to specify the number of clusters in clustering algorithms; the methods rely on manual parameter adjustment; the validation of these methods is limited to specific datasets (not generic); and the methods assume that the current state is only related to the previous state, which is not true in the case of electric load profiles because load patterns change continuously based on energy consumption behaviors. With the integration of distributed energy resources, load profiles are becoming increasingly unpredictable. To address the growing challenges, more flexible and reliable load profiling approaches are required.

3. The problem of class imbalance and missing data values in building datasets are not properly handled. This is crucial since real-world data is often incomplete or inconsistent and it is likely to produce error-prone results in developed models [149].

4. Existing studies considered mainly residential households [150,151]. Non-residential buildings such as industrial and commercial buildings have seldom been considered.

5. A problem of low classification accuracy due to high variations in building load profiles [152, 153].

6. The existing results are mostly validated on simulated datasets or only one type of building [154]. Insights based on real datasets are seldom provided. This research contributes by integrating computational intelligence with building energy evaluation and management, which in the end facilitates building owners and policymakers to optimize energy utilization and minimize carbon emissions for the development of green buildings [155].

Chapter 3

Anonymization of distribution network data using statistical distribution and parameter estimation approach

The work presented in this chapter is published in the following article:

1. **[Journal]** M. Ali, K. Prakash, C. Macana, M. Rabiul, A. Hussain, and HR. Pota, "Anonymization of distribution feeder data using statistical distribution and parameter estimation approach," ***Elsevier Sustainable Energy Technologies and Assessments***, vol. 52, p. 102152, Aug. 2022. **IF: 5.353**.

Summary: *Based on the research gaps in Chapter 2, this chapter provides a novel method for anonymizing distribution network data using a statistical distribution and parameter estimation approach. The statistical patterns of real distribution feeders are examined by accessing the confidential database of a local distribution network service provider. An algorithm based on the maximum likelihood estimate (MLE) is applied to estimate the statistical distribution parameters that represent the actual data. Then, these statistical distribution parameters are used to generate anonymized datasets that are realistic. A Kolmogorov-Smirnov (K-S) test is conducted to confirm the effectiveness of anonymized datasets, and the results are compared with the actual feeder datasets. Validation is carried out with existing methods and comparisons are shown on the different portions of the datasets (25 percent, 50 percent, 75 percent, and 100 percent). The comparison results indicate superior performance over traditional methods, with a performance improvement ranging from 1 to 13 percent. The practical application of the method is demonstrated by performing simulation studies on the IEEE 123-node test feeder. The*

method achieves consistent results on voltage profiles, with a maximum difference of 0.420 percent between actual and anonymized datasets.

3.1 Research gaps and contributions:

This chapter introduces a data anonymization technique that can be used by utilities and researchers to generate anonymous datasets with minimal expert knowledge. The frameworks and objectives of several works in the existing literature are quite different from those of ours. The goal is not to implement encryption and decryption techniques rather a simplified anonymization methodology for addressing the problem of data sensitivity in critical power infrastructures. The idea is to search for statistical patterns that emerge in the data of distribution feeders and their properties and use them to synthesize similar systems. This chapter aims to reduce the gap between academic and industry users regarding privacy concerns as many collaborative projects are based on sensitive and privacy-encumbered data.

3.1.1 Contributions

- An effective solution for data anonymization is developed by leveraging the statistical distribution techniques and parameter estimation approach.
- An algorithm based on the MLE is proposed for estimating the parameters that best represent the data. Then, a data anonymization procedure is established that uses the estimated parameters to generate anonymized datasets that are realistic.
- A new way to verify the representativeness of anonymized datasets is provided with Kolmogorov-Smirnov (K-S) hypothesis test.
- The quality of anonymized datasets is assessed by simultaneously considering the outlier detection and identifying the missing values, as suggested in [156].
- The effectiveness of the presented approach is validated with the key attributes of distribution feeders, such as feeder line length and results are compared with the existing data anonymizers.
- The practical feasibility of the proposed method is demonstrated by simulating on an IEEE 123-node test system and results are compared with original datasets.

3.2 Proposed data anonymization framework

Figure 3.1 provides an overview of the proposed framework. Confidential dataset from a distribution company's database is the input to the framework and the output is the anonymized dataset preserving the privacy. This framework consists of different mechanisms that work together to anonymize feeder datasets, e.g., an SQLite database input, distribution-fitting modeling in Python, an investigation of statistical characteristics, estimations of distribution parameters, hypothesis testing via the K-S test, data quality management, and comparison of results with popular anonymization methods. A short description of each mechanism is provided in the following paragraphs.

- *Phase 1:* In the first stage, the feeder data, which contains the key characteristics of distribution feeders, such as their line length, connected loads, capacitors, etc., is obtained from the distribution operator in Canberra, Australia. The data are originally available in a relational database such as SQL (Structured Query Language) and SQL queries are generated to retrieve the desired data from database tables. For instance, the SQL query `SELECT Line.L FROM database` is used to select all the feeder line lengths.
- *Phase 2:* In the second stage, the statistical characteristics of the feeder data are investigated by distribution fitting modeling. The motivation of this task is to understand the key characteristics of original data and identify their best distributions for study and analysis.
- *Phase 3:* After finding the best distribution, the statistical parameters of each distribution

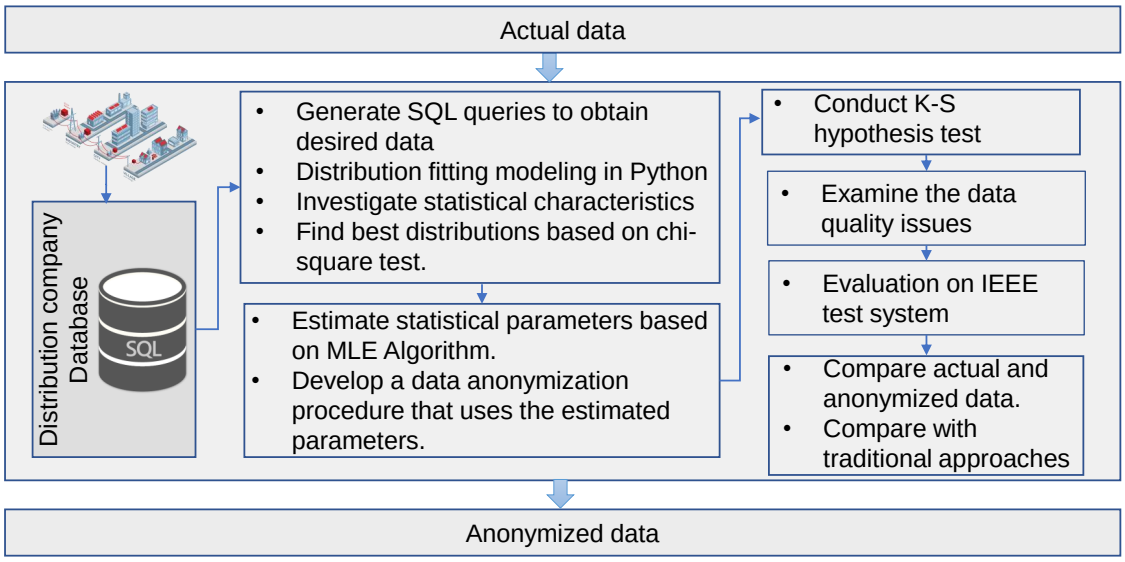


Figure 3.1: Proposed framework.

are estimated. The estimated parameters are then used to generate anonymized datasets.

- *Phase 4:* To confirm the representativeness of anonymized datasets, hypothesis testing based on the K-S test is conducted, with the hypothesis either selected or rejected according to its significance level and K-S static value. The quality of the data or loss of information in the anonymized datasets are assessed using box and whisker plots, as indicated in [156].
- *Phase 5:* The original and anonymized datasets are compared to evaluate the variations among them. The proposed approach is evaluated on an IEEE 123-node test feeder. In addition, the results are compared with recently developed anonymization methods.

3.2.1 Verification of actual and anonymized datasets

The choice for determining how well the anonymized datasets resemble the real data is the K-S test [157], a non-parametric goodness of fit test that compares a sample with a reference probability distribution (one-sample K–S test), or two samples (two-sample K–S test). The test compares a known hypothetical probability distribution of real data with the anonymized data distribution. A null hypothesis is formed to verify that the data samples are drawn from the same distribution to a certain degree of significance. Failing to accept the null hypothesis indicates that they come from different distributions.

3.2.2 One-sample K-S Test

In the one sample K-S test, we are given a sequence of data samples $(z_1, z_2, \ldots, z_N)$ with unknown distribution F. The underlying cumulative distribution function (cdf) is denoted by $F_1(z)$, and a hypothesized distribution by cdf $F_0(z)$. The null hypothesis is tested as

$$H_0 : F = F_0 \quad \text{vs.} \quad H_1 : F \neq F_0 \tag{3.1}$$

The empirical cdf formed by the K-S test from the data samples is

$$\hat{F}_1(z) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}\left(z_n \leq z\right) \tag{3.2}$$

where $\mathbb{I}(\cdot)$ is an indicator function that represents the value 1 if the input is true and zero otherwise. The maximum difference between the two cdf's $F_1(z)$ and $F_0(z)$ is estimated

using the K-S static,

$$D = \sup_{z \in \mathbb{R}} \left| F_1(z) - F_0(z) \right| \tag{3.3}$$

and, in practice, it is computed by

$$D_n = \max_i \left| \hat{F}_n\left(z_i\right) - \hat{F}_0\left(z_i\right) \right| \tag{3.4}$$

The decision of the hypothesis (δ), i.e., acceptance or rejection, can be determined by the decision rule

$$\delta = \begin{cases} H_0: & D_n \leq D_{\text{crit}} \\ H_1: & D_n > D_{\text{crit}} \end{cases} \tag{3.5}$$

The threshold D_{crit} is dependent on the level of significance α and is found from the condition

$$\alpha = F\left(\delta \neq H_0 \mid H_0\right) = F\left(D_n \geq D_{\text{crit}} \mid H_0\right) \tag{3.6}$$

Since the distribution of D_n can be tabulated for each n under H_0, the critical value (threshold) $D_{\text{crit}} = D_{\text{crit},\,\alpha}$ is approximated from the statistical tables [158, 159]. If the level of significance is taken as $\alpha = 0.05$, then the critical value is estimated as

$$D_{\text{crit},0.05} = \frac{1.36}{\sqrt{n}} \tag{3.7}$$

The hypothesis H_0 is accepted at the significance level α if $D_n < D_{\text{crit}}$. Also, the H_0 is tested with P_value and significance level α. The P_value is computed by

$$\text{Prob}(D > D_n) = 1 - 2\sum_{i=1}^{\infty}(-1)^{(i-1)} e^{\left(-2i^2 z^2\right)} \tag{3.8}$$

The hypothesis H_0 is accepted at significance level α if P_value$> \alpha$.

3.2.3 Two-sample K-S Test

If the hypothesized cdf F_0 is not available, and the data samples are drawn from another sequence $F_0, \xi_1, \xi_2, \ldots, \xi_{N_0}$, the empirical cdf $\hat{F}_0(\xi)$ is formed as

$$\hat{F}_0(\xi) = \frac{1}{N_0} \sum_{n=1}^{N_0} \mathbb{I}\left(\xi_n \leq \xi\right)$$

The K-S statistic is now

$$\hat{D} = \max_{1 \leq n \leq N} \left| \hat{F}_1\left(z_n\right) - \hat{F}_0\left(z_n\right) \right|$$

If the level of significance is $\alpha = 0.05$, then the critical value is

$$D_{\text{crit},0.05} = 1.36 \sqrt{\frac{1}{n_x} + \frac{1}{n_y}} \tag{3.9}$$

3.3 Experiments and results

For analysis, the real data provided by the distribution system operator in Australia is used. Table 3.1 presents a summary of the dataset statistics and these datasets are used to investigate the techniques developed in this study. Table 3.2 lists the steps taken to obtain the required feeder data from the electrical company database which is provided in SQLite files, the proprietary format of the SQLite database software. SQL queries are performed for easier manipulation and to obtain the data required for further operations. The feeder line length is considered in this paper as it is an important factor for the voltage profile of the distribution networks [160]. Investigations are carried out on the feeder line length dataset, and then the data is further clustered into HV and LV line lengths to test the proposed solutions under different scenarios. A summary of the dataset statistics is presented in Table 3.1 and these datasets are used to investigate the techniques developed in this study.

3.3.1 Scenario 1: Experiments on feeder line length dataset

In this section, the results obtained from the feeder line length dataset are analyzed to identify the key statistical distributions in the data that can be exploited in a process for

Table 3.1: Experimental datasets and scenarios used in this study

		Statistics (percentiles)							
Scenarios	Experimental datasets	Total samples	Mean (m)	Std (m)	25%	50%	75%	85%	95%
1	Feeder line length	8276	18.96	20.73	8.23	17.77	22.52	27.44	45.17
2	HV line length	463	45.23	59.31	1.00	31.58	62.81	84.42	157.16
3	LV line length	7816	17.40	14.28	8.44	17.62	21.97	25.69	37.33

Table 3.2: SQL queries to obtain data from an electrical utility database

Query	SQL Statement
Query:1	SELECT Line. 1 FROM Database
Description	Select feeder line length
Query:2	SELECT Line. I FROM Database WHERE VoltageLevel= low
Description	Cluster low voltage line length from line data
Query:3	SELECT Line. I FROM Database WHERE VoltageLevel = high
Description	Cluster high voltage line length from line data

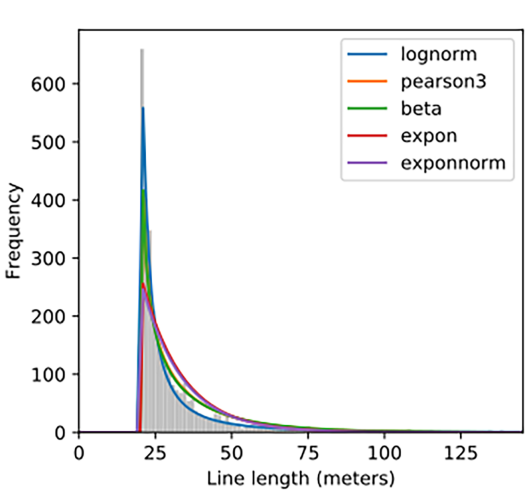
data synthesis. The distribution patterns are identified by statistical distribution modeling using the Python programming language. The purpose is to assess, understand and analyze the fitting of the distributions about the data. Figure 3.2(a) shows these patterns as well as a fit line that follows the lognormal distribution,

$$f(x;\mu,\sigma^2) = \frac{1}{sx\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln x-\mu)^2}{\sqrt{2\sigma^2}}} \tag{3.10}$$

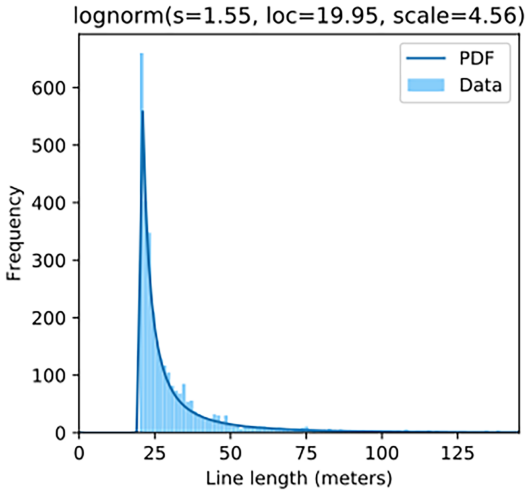
where $x > 0, s > 0$ and $f(x;\mu,\sigma^2)$ is the lognormal probability density function (PDF). This equation represents the parameters of a lognormal distribution, that is, the mean (μ) and standard deviation (σ) with s as a shape parameter. The obtained distribution characteristics are shown in Figure 3.2(a). From the figure, it is clear that the feeder line length follows a lognormal distribution which means that, if the feeder data (x) has a lognormal distribution, $Y = ln(x)$ has a normal one. This distribution is justified with the non-negative feeder line length and the distributions are right-skewed curve. Of the several distributions, the best fits for the data provided are determined using the chi-squared test [161] which sums the relative squared error between the observed and expected frequencies of data, and it is mathematically defined as

$$\text{chi-square test } = \sum \text{((observed - predicted) }^2 / \text{ predicted)} \tag{3.11}$$

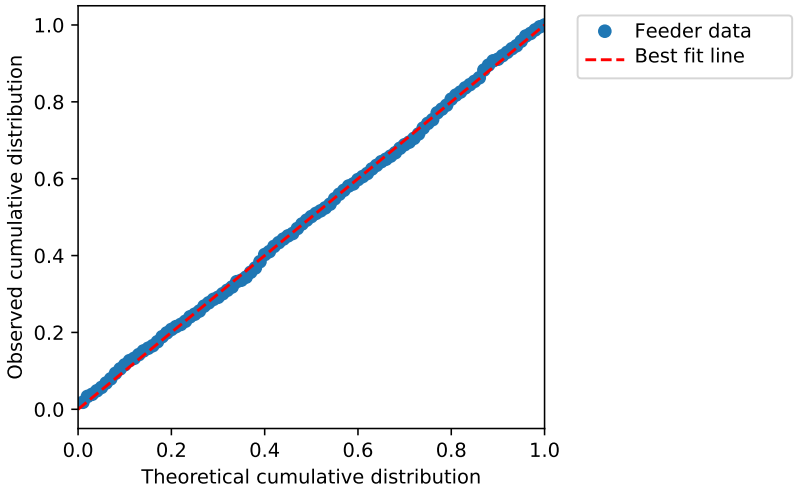
The lower the value obtained from this test, the better the fit. For example, the lognormal



(a) Illustrations of distribution patterns



(b) Best fits obtained from chi-squared test



(c) P-P plot showing the observed and theoretical cumulative distributions

Figure 3.2: Distribution characteristics of feeder line length dataset

distribution has the lowest chi-squared value, as shown in Table 3.3, and, based on these values, the best-fit distributions are plotted in Figure 3.2(b). The distributions with the lowest chi-squared values are ranked first. The pearson3 is the second best distribution followed by beta, expon, exponnorm, and norm distributions, respectively. Besides the chi-squared test, the distribution fits of the data are also validated by probability (P–P) plots. They are used to compare the probability distributions of the observed (empirical) data with those from a specified theoretical distribution such as a lognormal one, as shown in Figure 3.2(c). In order to compare distributions, we verify if the data points lie on a 45° line (x=y). If they deviate, the distributions differ while, if the fit is perfect, the data appears as a straight diagonal line. In the P-P plot, it can be seen that the feeder line data fits the theoretical lognormal distribution perfectly.

Table 3.3: Chi-squared test to assess the distribution fitting with PDFs and estimated parameters for specific distributions.

Rank	Distribution pattern	Chi-Squared test	Probability density function (PDF)	Estimated parameters
1	lognorm	69.41	$f(x;\mu,\sigma^2) = \frac{1}{sx\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln x-\mu)^2}{\sqrt{2\sigma^2}}}$	(s=1.54, loc=19.95, scale=4.55)
2	pearson3	1055.63	$f(x,\text{skew}) = \frac{\lvert\beta\rvert}{\Gamma(\alpha)}(\beta(x-\zeta))^{\alpha-1}\exp(-\beta(x-\zeta))$	(skew=2.60, loc=32.89, scale=16.79)
3	beta	1221.28	$f(x,a,b) = \frac{\Gamma(a+b)x^{a-1}(1-x)^{b-1}}{\Gamma(a)\Gamma(b)}$	(a=0.55, b=140, loc=20, scale=3384)
4	expon	6808.78	$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$	(loc=20, scale=13.21)
5	exponnorm	6824.54	$f(x,K) = \frac{1}{2K}\exp\left(\frac{1}{2K^2} - x/K\right)\text{erfc}\left(-\frac{x-1/K}{\sqrt{2}}\right)$	(k=2380.38, loc=20, scale=0.0055)
6	norm	42720.14	$f(x) = \frac{\exp(-x^2/2)}{\sqrt{2\pi}}$	(loc=33, scale=27.13)

Once the distribution patterns are identified, the parameters of a specific distribution fit are estimated by the MLE technique [162–164]. The MLE involves a likelihood function to find the probability distributions and parameters that best explain the observed data. The rationale is to find model parameters that can be used later in the data anonymization process. For instance, if the model parameters are denoted by θ = (μ, σ) for the given data $X_i, i = 1, 2, \ldots n$. The objective of MLE is to define the likelihood function for the relevant distribution and search for the parameter values by maximizing the data likelihood $\mathcal{L}\left(\mu, \sigma^2 \mid X\right)$. The steps for parameter estimation are shown in Algorithm 3.1 and the results are presented in column 5 of Table 3.3. In the case of the lognormal distribution, the three parameters are location, scale and shape parameter as denoted with μ, σ, and s respectively. The shape parameter is presented with $s = sigma(\sigma)$, location with $loc = mean(\mu)$, and scale with $scale = exp(\mu)$. The overall objective of these estimations is to use these parameters at a later stage in the data anonymization process.

3.3.2 Scenario 2: Experiments on HV feeder line length dataset

In this section, the results obtained for the HV feeder line length dataset are presented. The same process as for the feeder line length dataset is repeated for this one except that it has different distribution characteristics, the patterns of which are shown in Figure 3.3(a). The results show that the HV feeder line dataset follows the exponnorm distribution. This means that this dataset inherits the characteristics of both normal and

Algorithm: 3.1 Parameters estimation procedure.

Input: Data samples $X_i (i = 1, 2, \ldots n)$ and distribution model

Output: Estimated model parameters θ.

Consider the unknown model parameters are denoted by μ and σ^2 i.e θ= (μ, σ^2)

/*The first phase*/

$$\mathrm{L}\left(\mu, \sigma^2 \mid X\right) = \prod_{i=1}^{n} \left[f\left(X_i \mid \mu, \sigma^2\right) \right] \text{ //}\textit{Define Likelihood function}$$

$$= \prod_{i=1}^{n} \left(\left(2\pi\sigma^2\right)^{-1/2} X_i^{-1} \exp\left[\frac{-\left(\ln\left(X_i\right) - \mu\right)^2}{2\sigma^2} \right] \right) \text{ //}\textit{Density function}$$

$$= \left(2\pi\sigma^2\right)^{-n/2} \prod_{i=1}^{n} X_i^{-1} \exp\left[\sum_{i=1}^{n} \frac{-\left(\ln\left(X_i\right) - \mu\right)^2}{2\sigma^2} \right]$$

/*The second phase*/

Take the natural log of the likelihood function:

$$\mathcal{L}\left(\mu, \sigma^2 \mid X\right) = \ln\left(\left(2\pi\sigma^2\right)^{-n/2} \prod_{i=1}^{n} X_i^{-1} \exp\left[\sum_{i=1}^{n} \frac{-\left(\ln(X_i) - \mu\right)^2}{2\sigma^2} \right] \right)$$

/*The third phase*/

Find MLEs of μ and σ^2 which are $\hat{\mu}$ and $\hat{\sigma}^2$, To do this, take the gradient of $\mathcal{L}$ with respect to μ and σ^2

$$\frac{\delta\mathcal{L}}{\delta\mu} = \frac{\sum_{i=1}^{n} \ln\left(X_i\right)}{\hat{\sigma}^2} - \frac{2n\hat{\mu}}{2\hat{\sigma}^2} = 0 \text{ //}\textit{Solving with respect to } \mu$$

$$\implies \hat{\mu} = \frac{\sum_{i=1}^{n} \ln\left(X_i\right)}{n}$$

$$\frac{\delta\mathcal{L}}{\delta\sigma^2} = -\frac{n}{2}\frac{1}{\hat{\sigma}^2} - \frac{\sum_{i=1}^{n} \left(\ln\left(X_i\right) - \hat{\mu}\right)^2}{2} \left(-\hat{\sigma}^2\right)^{-2} \text{ //}\textit{Solving with respect to } \sigma^2$$

$$\implies \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \left(\ln\left(X_i\right) - \hat{\mu}\right)^2}{n}$$

$$\implies \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \left(\ln\left(X_i\right) - \frac{\sum_{i=1}^{n} \ln(X_i)}{n}\right)^2}{n}$$

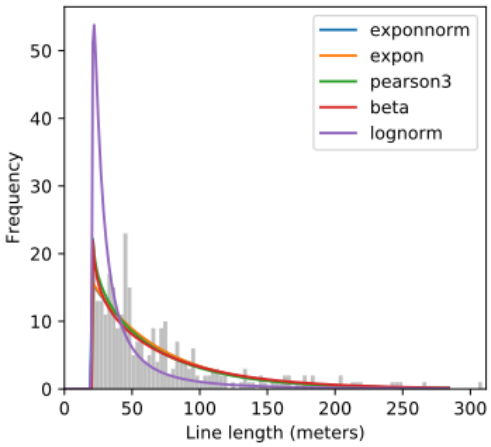
Return the parameters θ

exponential distributions as the exponnorm is a mixture of them, with its PDFs as

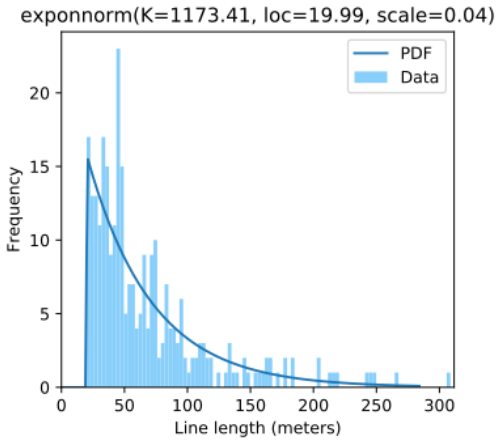
$$f(x, K) = \frac{1}{2K} \exp\left(\frac{1}{2K^2} - x/K\right) \operatorname{erfc}\left(-\frac{x - 1/K}{\sqrt{2}}\right) \tag{3.12}$$

where x is a real number and $K > 0$ is a shape parameter. The best-fit distribution for the data provided is determined by a chi-squared test and the results are plotted in Fig. 3.3(b). The probability distributions of actual (observed) data and the data coming from a specific theoretical distribution like exponnorm is compared in P-P plots of Fig. 3.3(c). It can be seen that both fall along a diagonal line which indicates good fits despite some differences

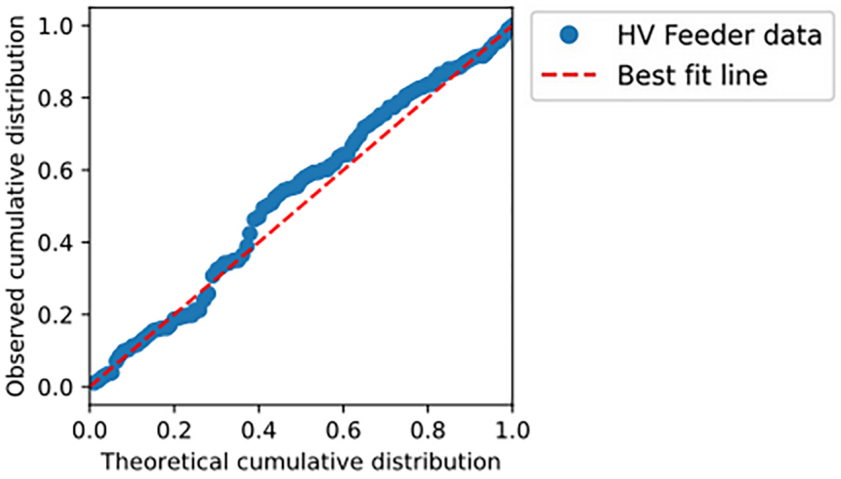
(deviations) in the middle of the two distributions. The parameters that explain the HV feeder line length dataset are estimated by Algorithm 1 and results are shown in Table 3.4.



(a) Distribution patterns



(b) Best fit with chi-square



(c) P-P plot

Figure 3.3: Distribution characteristics of HV feeder line length dataset

3.3.3 Scenario 3: Experiments on LV feeder line length dataset

In this section, the results are extended for the LV feeder line length dataset. The purpose is to test the presented solutions under different scenarios. As in Sections 3.3.1 and 3.3.2, the results are examined from three aspects, the distribution patterns, goodness-of-fit values using a chi-squared test, and P-P plots, to compare their probability distributions. Figure 3.4(a) shows the distribution patterns as well as a fit line that follows the lognormal

Table 3.4: Distribution fits with parameter estimations

Rank	Distribution pattern	Chi-Squared test	Estimated parameters
1	exponnorm	23.76	(k = 1173.41, loc = 19.99, scale = 0.043)
2	expon	23.90	(loc = 20.044, scale = 51.34)
3	pearson3	52.49	(skew = 2.14, loc = 69.40, scale = 52.89)
4	beta	55.97	(a = 0.834, b = 40.63, loc = 20.04, scale = 2721.77)
5	norm	1992.32	(loc = 71.38, scale = 62.52)
6	lognorm	2761.51	(s = 1.36, loc=19.90, scale = 11.19)

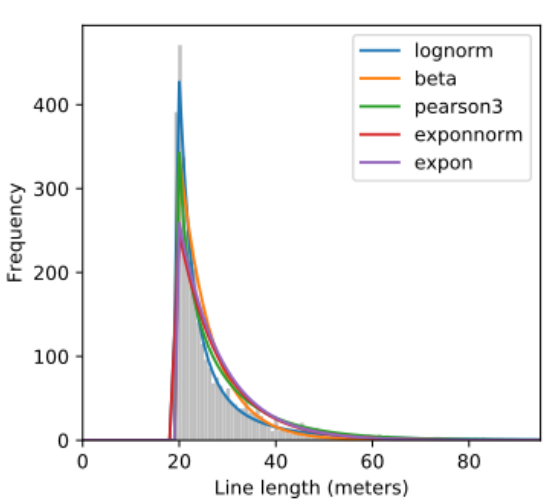
distribution

$$f(x;\mu,\sigma^2) = \frac{1}{sx\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln x-\mu)^2}{\sqrt{2\sigma^2}}} \tag{3.13}$$

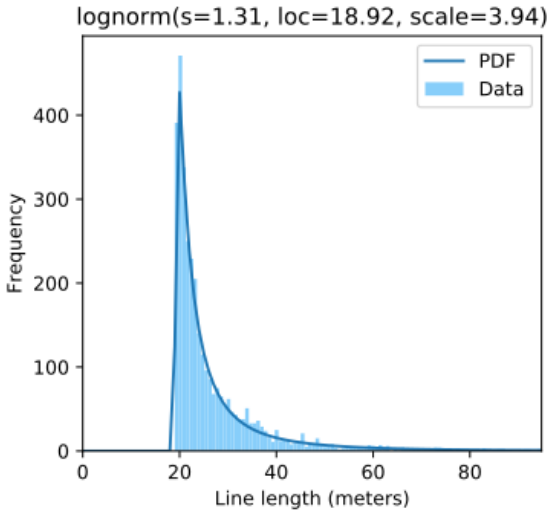
where $x > 0$,$\mu \in R$, $\sigma > 0$ and $f(x;\mu,\sigma^2)$ is the lognormal probability density function. It can be seen from Figure 3.4(a) that samples from the LV feeder line length dataset have lognormal distributions which means that they have positive real values. Of the multiple distributions, the best fits are determined from the values of the chi-squared test and plotted in Figure 3.4(b). The comparisons of the observed and theoretical distributions are made with probability plots in Figure 3.4(c). The plot shows a perfect fit of data along the diagonal line indicating that data distributions are consistent with a lognormal model. As the plots follow the 45° diagonal quite well, it can be considered that the fitted lognormal distributions are reasonably good at describing the LV feeder line length data. The best fit parameters are calculated using the MLE technique defined in Algorithm 1, and the results are shown in Table 3.5.

Table 3.5: Estimated parameters and distribution patterns

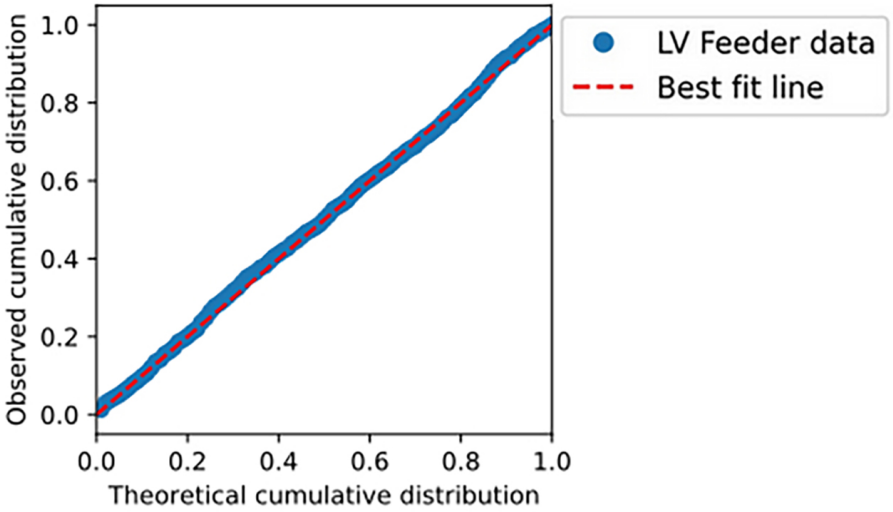
Rank	Distribution pattern	chi-squared test	Distribution parameters
1	lognorm	31.09	(s = 1.31, loc = 18.92, scale = 3.94)
2	beta	658.87	(a = 1.003, b = 41.38, loc = 19, scale = 2714.83)
3	pearson3	1378.48	(skew = 2.463, 10c = 27.744, scale = 10.76)
4	exponnorm	3521.25	(k = 1463.87, loc = 19, scale = 0.006)
5	expon	3533.45	(loc = 19, scale = 8.816)
6	norm	38731.09	(loc = 27.81, scale = 15.14)



(a) Distribution patterns



(b) Best fit with chi-square



(c) P-P plot

Figure 3.4: Distribution characteristics of LV lines dataset

3.3.3.1 Summary and findings from distribution patterns

The key observations from the statistical patterns obtained in section 3.3 are summarized as follows.

- The best distributions are identified and justified by conducting a chi-squared test. The test sums the relative squared error between the observed and expected frequencies of data, and it is mathematically presented in equation (3.11).
- In section 3.3.1, the distribution patterns of feeder line length datasets are shown and the results of the chi-squared test are presented in Table 3.3. The lower chi-

squared values indicate a better data fit. The tabular results show that the best distribution fit for feeder line length is lognormal as it has the lowest value of 69.41 among other distributions.

- In section 3.3.2, the distribution patterns collected from the HV feeder line length dataset are illustrated. The chi-squared test shows that the data samples have characteristics of an exponnorm distribution. The dataset inherits the characteristics of both normal and exponential distributions as the exponnorm is a mixture of them. As shown in Table 3.4, the exponnorm has the lowest chi-square value of 23.76, indicating that the statistical correlations of HV correlate with the exponnorm distribution. The second distribution fit is exponential, followed by person3, beta, and lognormal distributions.
- In section 3.3.3, the results indicate that the dataset of LV feeder line lengths forms the lognormal distribution. The distribution patterns show that the LV data is lognormally distributed and has only positive real values. As shown in Table 3.5, the lognorm has the lowest chi-square value of 31.09, indicating that the statistical correlations of LV correlate with the lognormal distribution. The second distribution fit is beta, followed by person3, exponnorm, exponential and normal distributions.

3.3.4 Data anonymization procedure

The data anonymization process is established from the parameters estimated by Algorithm 1. The procedure is formally presented in Algorithm 3.2, and it begins by taking the actual data and then initializing the required variables. For instance, the required variables are the best distribution parameters which are denoted by $P_d = \{S_p, L_p, S_i\}$. The algorithm searches for the distribution patterns in data in order to obtain the required variables and decides the best distribution based on the results of the chi-squared test in Section 3.3.1. Once the distributions are identified, the parameters are estimated using the MLE technique in Algorithm 3.2. The estimated parameters are then used to generate an anonymized dataset. The benefit of this approach is to construct representative anonymous datasets that can be used for research purposes without accessing the confidential data. In the next step, the representations of the anonymized datasets are verified through K-S test simulations to determine whether anonymized and original data samples have the same distribution characteristics. To make this decision, the K-S test establishes the following two hypotheses: (1) a null hypothesis (H_0) which considers that two datasets

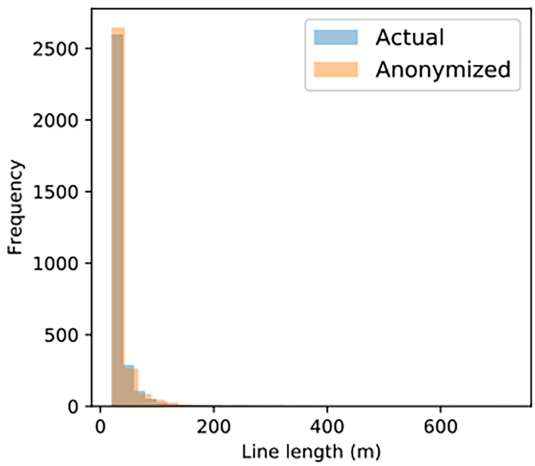
Algorithm 3.2 Generate anonymized dataset

Input: Data samples
Required Variables :$P_d = \{S_p, L_p, S_i\}$ ▷ Get the best parameters
for Data samples $X_i(i = 1, 2, \ldots n)$ **do**
    Get the distribution patterns
    Collect the distribution names
    Determine the best distributions with Chi-square test.
    Rank the distribution by Chi-square value.
    Estimate the parameters with Algorithm:1
    Return the best parameters
end for
Use the estimated parameters to generate anonymize dataset
Check the representativeness of anonymized dataset with KS test
for Hypothesis testing **do**
    Estimate (P_values), (α), (D_n), (D_{crit})
    Accept or reject based on (P_values) verses (α) and (D_n) verses (D_{crit})
end for
Compare the characteristics of original and anonymized dataset
Store the results in output variable
Output: Anonymized dataset

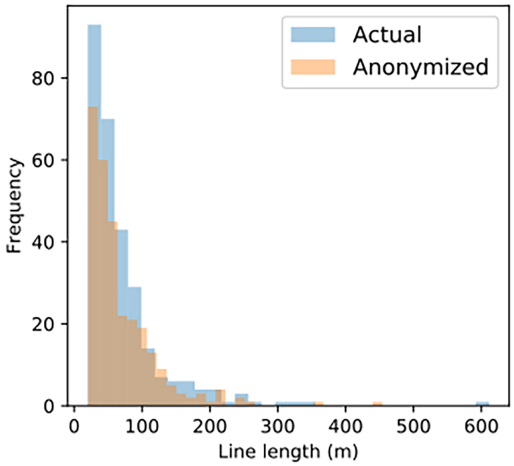
values are from the same distribution; and (2) another (H_1) that they are from different distributions. The test for accepting or rejecting a hypothesis can be carried out using two criteria: (a) comparing the K-static value obtained from the K-S test with the critical value in the K-S table [158, 159]; and (b) comparing the P_value of the K-S test with the level of significance which is 0.05 in our case. Then, the characteristics of the original and anonymized datasets are compared and results obtained are stored in the output variable. The output of the algorithm is an anonymized dataset that represents the original data.

The comparison results of actual and anonymized datasets are evaluated using three different datasets as shown in Figures 3.5(a), 3.5(b), 3.5(c), respectively. The comparisons are drawn with the overlapping histograms and comparing the datasets with different colors. A visual inspection suggests that anonymized data reflects the similar patterns or trends as observed in actual data and shows strong statistical consistency with the real distribution data.

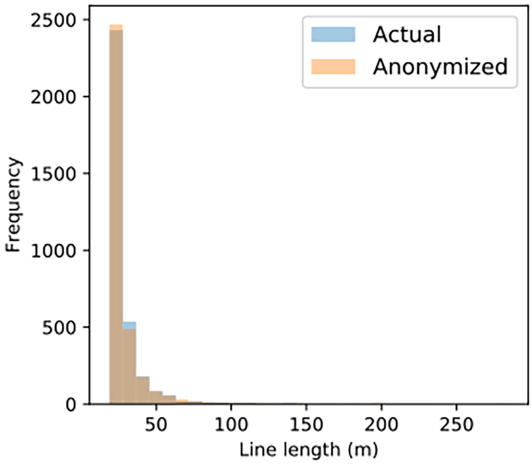
After a visual comparison of the actual and anonymized datasets, the next step is to assess the representativeness of the anonymized datasets. The anonymized datasets should reflect the statistical properties of the original data. To confirm this, a K-S test is carried on the anonymized datasets and decided to accept or reject them based on the two



(a) Feeder line length



(b) HV feeder lines



(c) LV feeder lines

Figure 3.5: Comparisons of actual and anonymized data on three different datasets

standard criteria mentioned in Table 3.6. In the first criteria, the P_value obtained from the K-S test is compared with the significance level $(\alpha = 0.05)$ which corresponds to a $1 - 0.05 = 0.95$ or 95% confidence interval. In the second criteria, the K-S_static (D_n) is compared with the critical value (D_{crit}) obtained from the K-S table [158, 159]. If the P_value is greater than the significance level (0.05), the hypothesis is accepted otherwise it is rejected. Its acceptance indicates that the anonymized datasets are drawn from the same distributions and have the same characteristics as the original one. While its rejection demonstrates that the anonymized datasets are not drawn from the same distribution and have different characteristics than the original dataset. From Table 3.7 (row 1), it is evident that the anonymized P_value is greater than the significance level which means that the hypothesis that the anonymized dataset has the same distribution (lognormal) as the real

one is accepted. The other distributions fail to satisfy this criteria as their P_values are less than the significance level. To test using the second criteria, the K_static (D_n) in Table 3.7 (row 1) is compared with the critical value (D_{crit}) of K-S table. From the results, it is apparent that the D_n is less than D_{crit} which indicates that this hypothesis is accepted and the anonymized datasets have the same statistical properties as the original one. The results obtained from the anonymized datasets are presented in Table 3.8. The results show that the hypothesis criteria is accepted for all datasets, indicating that the anonymized datasets have same representations (statistical compliance) to real data.

Table 3.6: Criteria for acceptance or rejection of anonymized datasets

Criteria:1	**P_value and significance level**	
	Accept	If P_value $>$ significance level (α)
	Donot Accept:	If P_value $<$ significance level (α)
Criteria:2	**K_static and critical value**	
	Accept	If K_static (D_n) $<$ Critical value (D_{crit})
	Donot Accept:	If K_static (D_n) $>$ Critical value (D_{crit})

Table 3.7: Results obtained from K–S test on anonymized datasets

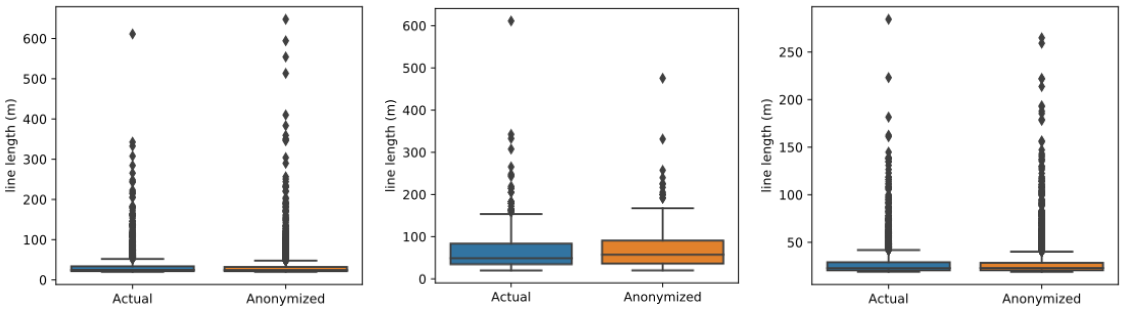
	Rank	**Distribution**	**KS_Test**
Dataset:1	1	lognorm	(‘lognorm’, $\mathbf{D_n = 0.02386}$, **P_value=0.05671)**
	2	pearson3	(‘pearson3’, $\mathrm{D}_n = 0.0846, \mathrm{P\_}$ value $= 0.0)$
	3	beta	(‘beta’, $\mathrm{D}_n = 0.07885, \mathrm{P\_}$ value $= 0.0)$
	4	expon	(‘expon’, $\mathrm{D}_n = 0.20439, \mathrm{P\_}$ value $= 0.0)$
	5	exponnorm	(‘exponnorm’, $\mathrm{D}_n = 0.20422, \mathrm{P\_}$ value $= 0.0)$
	6	norm	(‘norm’, $\mathrm{D}_n = 0.31315$, P_ value $= 0.0)$
Dataset:2	1	exponnorm	(‘exponnorm’, $\mathbf{D_n = 0.0710}$, **P_value =0.1079)**
	2	expon	(‘expon’, $\mathrm{D}_n = 0.07231, \mathrm{P\_}$ value $= 0.09665)$
	3	pearson3	(‘pearson3’, $\mathrm{D}_n = 0.06559, \mathrm{P\_}$ value $= 0.1648)$
	4	beta	(‘beta’, $\mathrm{D}_n = 0.0703, \mathrm{P\_}$ value $= 0.11394)$
	5	norm	(‘norm’, $\mathrm{D}_n = 0.20577, \mathrm{P\_}$ value $= 0.0)$
	6	lognorm	(‘lognorm’, $\mathrm{D}_n = 0.35233, \mathrm{P\_}$ value $= 0.0)$
Dataset:3	1	lognorm	(‘lognorm’, $\mathbf{D_n = 0.01685}$, **P_value=0.2916**
	2	beta	(’beta’, $\mathrm{D}_n = 0.06224, \mathrm{P\_}$ value $= 0.0)$
	3	pearson3	(‘pearson3’, $\mathrm{D}_n = 0.0775, \mathrm{P\_}$ value $= 0.0)$
	4	exponnorm	(‘exponnorm’, $\mathrm{D}_n = 0.15285, \mathrm{P\_}$ value $= 0.0)$
	5	expon	(‘expon’, $\mathrm{D}_n = 0.15354, \mathrm{P\_}$ value $= 0.0)$
	6	norm	(‘norm’, $\mathrm{D}_n = 0.28027, \mathrm{P\_}$ value $= 0.0)$

Table 3.8: Testing of anonymized datasets using two standard criteria

Testing	Dataset 1	Dataset 2	Dataset 3
Criteria:1	Accept Hypothesis: P_value (0.0567) is $> \alpha(0.05)$	Accept Hypothesis: P_value (0.1079) is $> \alpha(0.05)$	Accept Hypothesis: P_value (0.2916) is $> \alpha(0.05)$
Criteria:2	Accept Hypothesis: D_n (0.0238) is $<$ D_{crit} (0.0243)	Accept Hypothesis: $D_n(0.0710)$ is $<$ D_{crit} (0.0806)	Accept Hypothesis: $D_n(0.0168)$ is $<$ D_{crit} (0.0234)

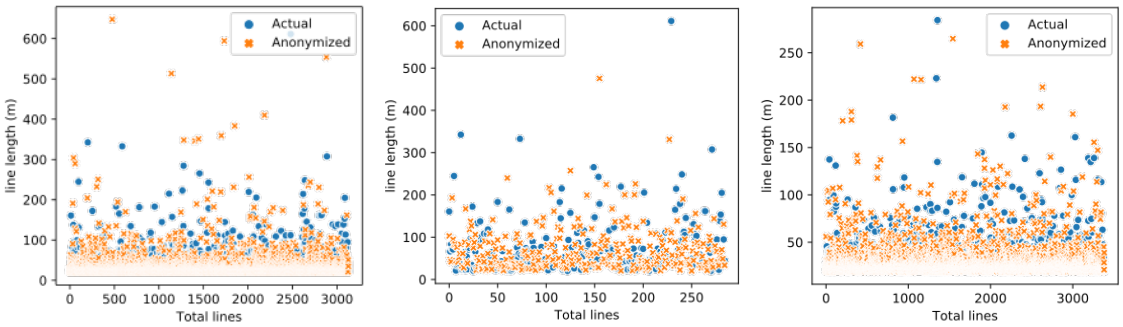
3.3.5 Data quality assessment

The quality of anonymized datasets is investigated using box-and-whisker plots, as suggested in [156]. The aim is to determine how the data is spread in the actual and anonymized datasets and the presence of unusual data points (outliers). The data quality is assessed by dividing the data into five key components as shown in Figure 3.6. This includes the minimum value (lower line), the first quartile (Q1), the sample median (Q2), the third quartile (Q3), the maximum value (upper line), and outliers which are indicated by dot points. Any point above the upper or below the minimum value is considered an outlier or unusual data point. Q1 shows the first 25% of the data, Q2 50%, and Q3 75%. The procedure for computing these five components is briefly discussed in [156]. In Figure 3.7(a),3.7(b),3.7(c), the box-and-whisker plots for the three experimental datasets are shown, demonstrating the dispersion of data samples in actual and anonymized datasets, as well as outliers in the datasets. Figure 3.7(a) shows that the data samples of actual and anonymized datasets fall within the upper and lower lines on the plot. Within the upper and lower lines, the Q1 of an actual dataset is 23.83 and for the anonymized dataset is 22.24. It shows that 25 percent of the feeder line segments are below 23.84 m (meters) in length in the actual dataset and 22.34 m in the case of the anonymized dataset. The Q2 of actual is 24.03 m and the anonymized one is 22.32 m. The Q3 of the actual dataset is 24.20 m and for anonymized dataset s 23.27 m. Any data sample above the upper or below line is considered an outlier or unusual data point in the dataset. Table 3.9 shows the summary of the total numbers of outliers in the datasets and their percentages. The highest number of outliers are found in dataset 3 which contains 341 data points. In addition to the unusual data points (outliers), the positions of individual data points in actual and anonymized datasets are shown using scatter plots, as presented in Figure 3.8. Each data sample in the actual and anonymized datasets is presented in blue and orange, respec-



(a) Feeder line length (b) HV feeder lines (c) LV feeder lines

Figure 3.7: Data variations and outliers in actual and anonymized datasets



(a) Feeder line length (b) HV feeder lines (c) LV feeder lines

Figure 3.8: Individual data points from actual and anonymized datasets are shown in scatter plots

tively. The horizontal position (x-axis) indicates the total number of line segments, and the vertical position (y-axis) indicates the length of lines (in meters). From the plot, it has been found that the number of lines and lengths in actual datasets are closely positioned with the anonymized datasets.

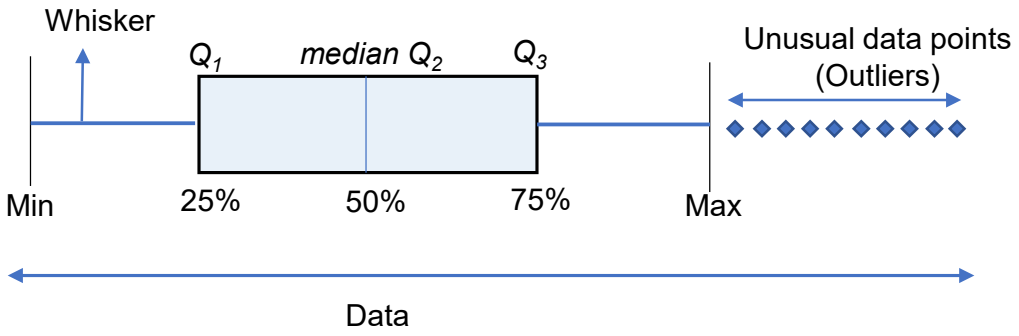


Figure 3.6: Data quality check based on five components of box-and-whisker plots

Table 3.9: Numbers of outliers in actual and anonymized datasets

	Dataset: 1		Dataset: 2		Dataset:3	
	Actual	Anonymized	Actual	Anonymized	Actual	Anonymized
No.of unusual data points	294	365	24	15	291	341
(%) Unusual data points	9.39	10.80	8.36	5.22	8.60	10.07

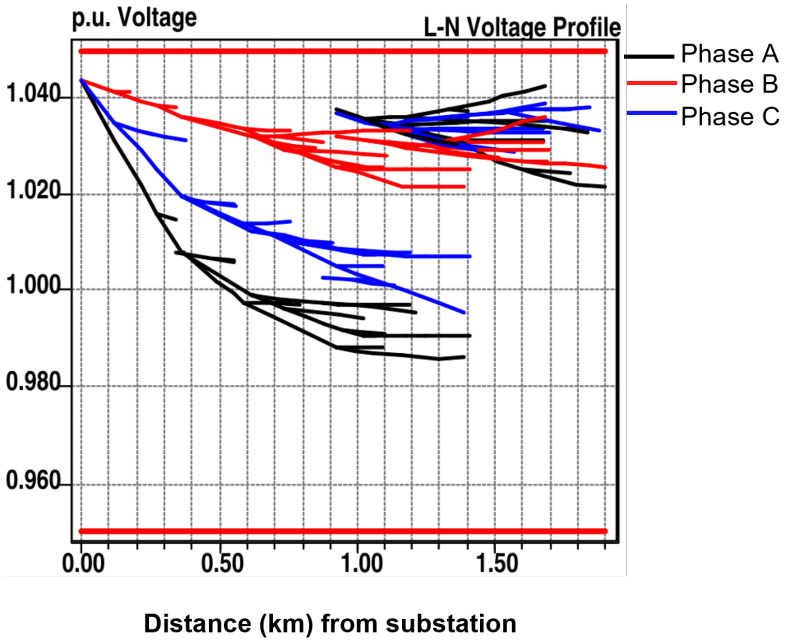
3.4 Implementation on IEEE 123-node test feeder

To verify the practical feasibility of the proposed technique, the methodology is tested on IEEE 123-node system by anonymizing the test system distribution grid data. At first, an anonymized dataset is created for the IEEE 123-node test feeder. The load flow results for the anonymized IEEE 123-node feeder are analyzed and compared with the original IEEE 123-node test feeder from the EPRI website [165]. In this work, we analyze two metrics, the voltage profiles and power flow through the lines, to compare the anonymized and the actual IEEE 123 test feeders.

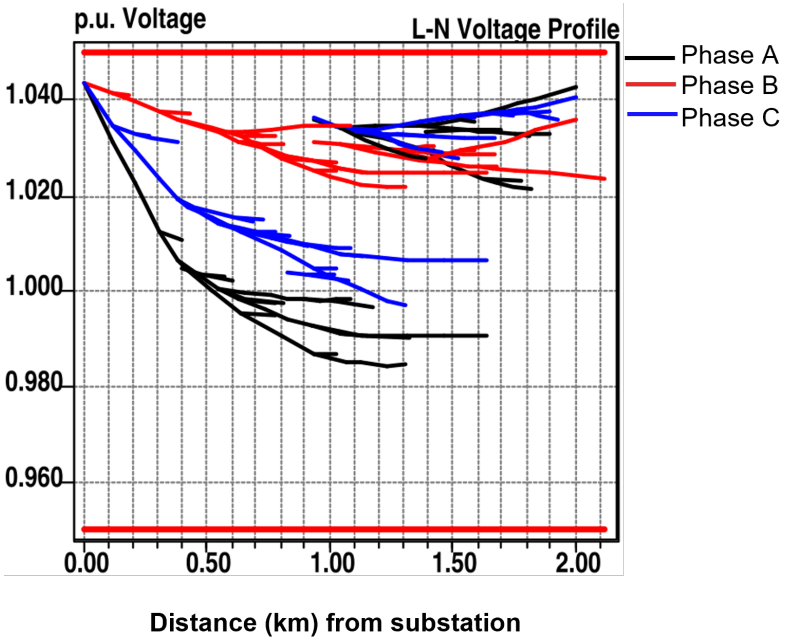
3.4.1 Voltage profiles

The voltage profiles of the anonymized and the actual IEEE 123-node feeder are analysed and compared, as shown in Figure 3.9(a) and Figure 3.9(b). The phase voltages of the anonymized and actual IEEE 123-node feeder are presented using three distinct colors. The black represents phase A, the red phase B, and the blue phase C. From the obtained graphical results, it has been found that the anonymized IEEE 123-node feeder can provide similar load flow results to the actual IEEE 123-test system. Table 3.10 provides the load flow results. Using the results from Table 3.10 and equation (3.14), the percentage mismatch is calculated at randomly selected nodes to quantify the mismatch between the anonymized and the actual results. It was found that the maximum variation between the actual and the anonymized dataset occurs at Bus 57, with 0.420 percent for phase A, 0.256 for phase B, and 0.303 for phase C. To ensure this percentage variation is acceptable, we compared the obtained results with relevant work published in the literature [166]. For clarification, the per-unit voltages are also converted to the actual operating voltages by multiplying them with the base operating voltage of 2.4 kV. As shown in Table 3.10, the maximum voltage difference between the actual and the anonymized

dataset occurs at Bus 57, with a total voltage difference of 10.046 V (0.42 percent) at phase A, 6.336 V (0.256 percent) at phase B, and 7.368 V (0.303 percent) at phase C.



(a) Voltage profile from IEEE 123-node feeder



(b) Voltage profile from anonymized test feeder

Figure 3.9: Comparison of results on IEEE 123-node test feeder

$$\text{Percentage Mismatch } = |\text{ (Actual-Anonymized) }| \text{ / (Actual) } \times 100 \tag{3.14}$$

Table 3.10: Percentage mismatch and voltage difference calculations for actual and anonymized data

		Actual			**Anonymized**			**Percentage mismatch**			**Voltage difference (V)**		
Sr.#	Bus	Phase A (p.u.)	Phase B (p.u.)	Phase C (p.u.)	Phase A (p.u.)	Phase B (p.u.)	Phase C (p.u.)	Phase A (%)	Phase B (%)	Phase C (%)	Phase A (V)	Phase B (V)	Phase C (V)
1	51	0.990	1.024	1.006	0.990	1.024	1.006	0.023	0.004	0.023	0.5568	0.1200	0.5760
2	95	1.033	1.026	1.037	1.032	1.028	1.037	0.091	0.231	0.057	2.2800	5.7120	1.4400
3	**57**	0.994	1.030	1.011	0.990	1.028	1.00	**0.420**	**0.256**	**0.303**	**10.046**	**6.3360**	**7.3680**
4	151	0.990	1.024	1.006	0.990	1.024	1.006	0.023	0.004	0.023	0.5568	0.1200	0.5760
5	8	1.015	1.038	1.025	1.012	1.037	1.023	0.329	0.069	0.203	8.0400	1.7280	5.0160
6	44	0.991	1.026	1.008	0.992	1.027	1.009	0.066	0.073	0.081	1.5720	1.8000	1.9680
7	8	1.015	1.038	1.025	1.012	1.037	1.023	0.321	0.069	0.203	7.8432	1.7280	5.0160
8	52	1.002	1.034	1.016	1.004	1.035	1.017	0.218	0.054	0.121	5.2560	1.3440	2.9520
9	7	1.022	1.039	1.029	1.022	1.039	1.029	0.053	0.007	0.033	1.3200	0.1920	0.8160
10	50	0.990	1.024	1.006	0.990	1.024	1.006	0.005	0.002	0.022	0.1224	0.0720	0.5520

3.4.2 Power flow through the lines

The power flow through the lines of the anonymized and the actual networks are compared and the percentage mismatch between the two load flow results are shown in Figure 3.10. The maximum difference in active power flow through the lines of the actual and anonymized networks is 0.307 percent on line 19 (phase A), whereas the maximum difference in the reactive power is 0.382 percent at line 36 (phase C). It has been found from the obtained results that the difference in active and reactive power flow through the lines of the actual and the anonymized networks is mostly between 0 to 0.4 percent, which is relatively small and acceptable, as discussed in [167]. It can be concluded that the anonymized network can provide reasonably accurate load flow results in terms of power flow in lines compared to the actual IEEE 123 bus network.

3.5 Comparison with the state of the art

The proposed approach is compared with recently published noise addition methods (NAM) [27, 168–170] and the machine learning based data perturbation (MLDP) method [85]. In the case of NAM, anonymization is performed by adding noise to data. However, including noise causes uncertainty in data, reducing its utility [171]. Due to additive noise, the correlations between the attributes are distorted or lead to trends that do not actually exist. A classic example is presented in [32] to demonstrate why it is crucial to

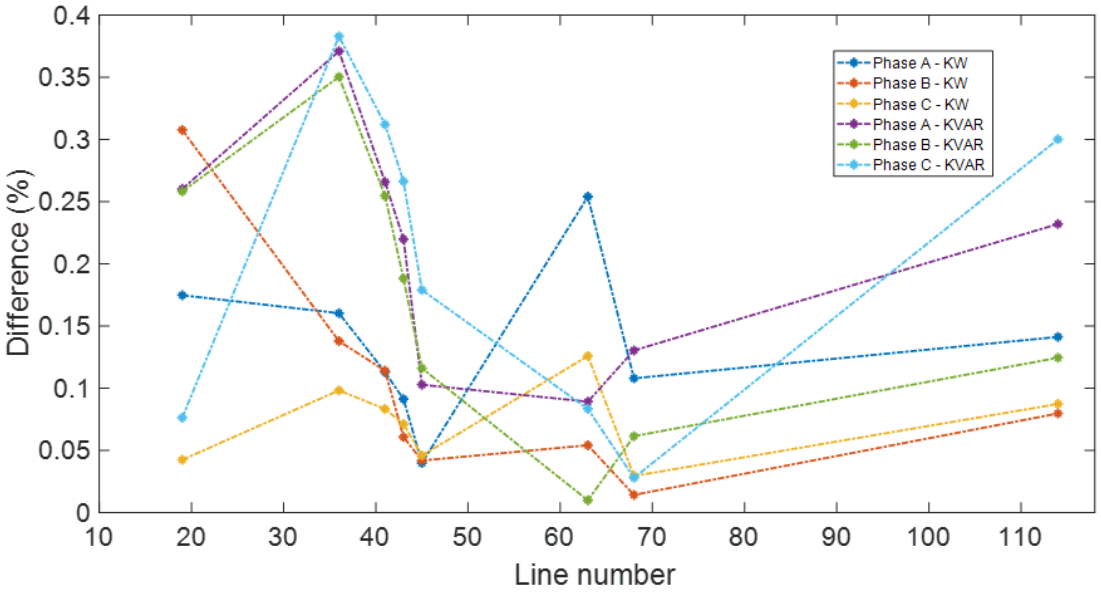


Figure 3.10: Power flow through the lines of IEEE 123-node test feeder

preserve correlations between data attributes. In other words, a trade-off between privacy preservation and utility of data is essential as the loss of statistical properties makes the data meaningless. In the MLDP approach, a machine learning model is trained from the actual data and similar kind of anonymized datasets are created to ensure data privacy.

For validation, data anonymization is applied on the three datasets as described in Table 3.1. A standard metric, root-mean-square error (RMSE) [172, 173], is used as a performance measure to compare the proposed method with previous works. The RMSE is calculated by equation (3.15) for each data sample in the actual and anonymized datasets, and the results are shown in Table 3.11. The smaller RMSE indicates a better performance with less information loss. For fair comparisons, the results are assessed on different percentages of the datasets. For the first 25 percent of data, the proposed scheme obtained a RMSE value of 48.37 percent compared to 50.6 and 52.68 of MLDP and NAM methods. The performance is also evaluated on 50 percent of the data and 49.86 percent RMSE value is obtained by the proposed method compared to 51.95 and 52.88 from two other methods. The process is repeated for 75 percent and 100 percent of data and results are shown. The RMSE values of the proposed scheme are lower than the traditional methods on different percentages of datasets, indicating a comparable performance from published works. There are two reasons for the performance improvements compared to existing approaches: the presented method ensures that the anonymized datasets are not dubious as

Table 3.11: Comparison of performances for different percentages of datasets

Experimental Datasets	**Method:1** Proposed (RMSE)				**Method:2** MLDP (RMSE)				**Method:3** NAM (RMSE)			
	Data percentage											
	25%	50%	75%	100%	25%	50%	75%	100%	25%	50%	75%	100%
Anonymized Dataset:1	48.37	49.86	48.22	50.94	50.6	51.95	49.64	51.62	52.68	52.88	50.45	52.28
Anonymized Dataset:2	70.63	77.02	75.13	76.8	72.59	77.96	78.44	77.71	73.19	78.32	79.33	79.01
Anonymized Dataset:3	25.52	26.26	24.89	26.6	27.35	28.55	28.32	27.26	28.71	29.06	28.69	28.82

it correctly identifies the best parameters from distributed data. Also, it successfully captures the underlying correlations (trends) of real data, which is important in generating synthetic records and maintaining a balance between privacy preservation and data utility.

$$\text{RMSE} = \sqrt{(\frac{1}{N})\sum_{i=1}^{N}(Actual_i - Anonymized_i)^2} \tag{3.15}$$

3.5.1 Percentage improvements

The performance improvements obtained by the proposed method over recent methods were computed as [174]

$$\text{Improvement } = \frac{R_t - R_p}{R_t} \times 100 \tag{3.16}$$

where R_t and R_p represent the RMSE values of the traditional and proposed method. Figure 3.11 shows the performance improvements obtained by the proposed method over two recent methods. The x-axis shows the different percentages of dataset (25 percent, 50 percent, 75 percent, and 100 percent) and y-axis presents the percent improvements. The improvements over dataset 1 is presented in blue, whereas on datasets 2 and 3 are shown in orange and grey. An average improvement of 1 to 13 percent is observed on three separate datasets. It demonstrates that the results are comparable to those of previously published works. The success of the proposed method is due to three important factors. The statistical trends are exploited in the data synthesis process to generate similar characteristics to the real samples. The K-S test is used to evaluate analysis and synthesis results, ensuring that anonymized datasets adhere to the statistical compliance of actual

datasets. The proposed algorithms search for the statistical distribution parameters of actual datasets to reproduce similar anonymized datasets.

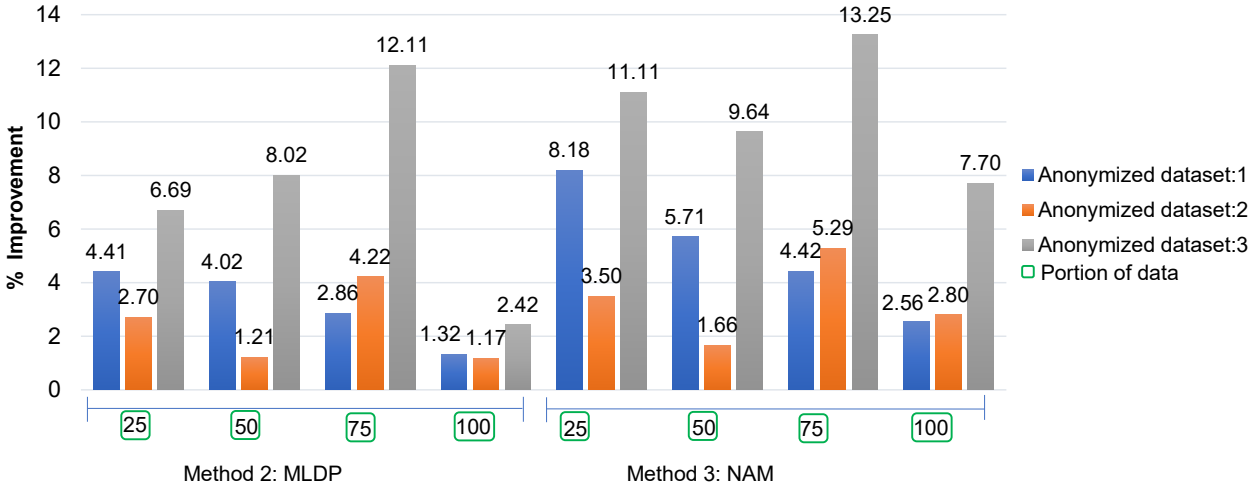


Figure 3.11: Performance improvements obtained by the proposed method over recent methods on three datasets

3.6 Concluding remarks

In this chapter, a privacy-preserving data anonymization scheme for obfuscating the sensitive information in distribution networks is presented. The scheme accommodates the statistical distribution with the parameters estimated from the data provided. It involves two algorithms, a MLE for estimating the parameters from the data and a data anonymization procedure for generating anonymized datasets that are sufficiently realistic. The statistical patterns of real utility data are studied, and representations of anonymized datasets are analyzed using the K–S hypothesis test. The method is validated on the IEEE 123-node test feeder by simulating the anonymized datasets on OpenDSS. The experimental results show that the presented approach offers competitive performance in terms of voltage profiles and power flow through the lines, with a maximum difference of 0.420 percent and 0.383 percent between actual and anonymized datasets. The anonymized datasets show good statistical compliance with trends identified in real distribution feeders. The method was then experimentally proven by comparing it to the benchmark data anonymization methods. Validation is conducted on three different scenarios (datasets) for a fair assessment, and findings are shown using a standard error metric, RMSE. The comparison results indicate a performance improvement of 1 to 13

percent over traditional approaches. This is due to two factors: its ability to capture the key characteristics and correlations from the data and the correct estimation of parameters that reflect the actual data.

As a benefit, the data anonymization contribution facilitates open-data sharing to remove the data access barriers between academics and industry users. The proposed techniques provide competitive performances and a practical solution for anonymizing distributed datasets. The anonymized datasets can be used for research purposes without the need to access confidential data. The research in this chapter can be regarded as a step towards the implementation of information security in future distribution networks to maintain data integrity in real-world applications.

Chapter 4

Synthetic power distribution networks and datasets from open-data and data synthesis algorithms

The work presented in this chapter is published or submitted in the listed articles:

1. **[Journal]** M. Ali, K. Prakash, C. Macana, MQ. Raza, AK. Bashir and HR. Pota, "Modelling synthetic power distribution network and datasets with industrial validation," *Elsevier Journal of Industrial Information Integration*, Dec. 2021. [under-review] **IF: 10.063**

2. **[Journal]** M. Ali, K. Prakash, C. Macana, AK. Bashir, A. Jolfaei, A. Bokhari, JJ. Klemes, and HR. Pota, "Modeling residential electricity consumption from public demographic data for sustainable cities," ***Energies***, vol. 15, no. 6, Art. no. 6, Jan. 2022. **IF: 3.004**

3. **[Conference]** M. Ali, C. A. Macana, K. Prakash, R. Islam, I. Colak, and H. Pota, "Generating open-source datasets for power distribution network using OpenStreetMaps," Sep. 2020, pp. 301–308. *IEEE International Conference on Renewable Energy Research and Application (ICRERA)*

4. **[Conference]** M. Ali, C. A. Macana, K. Prakash, B. Tarlinton, R. Islam, and H. Pota, "A novel transfer learning approach to detect the location of transformers in distribution Network," Jun. 2020, pp. 56–60. Sep. *IEEE International Conference on Smart Grid (icSmartGrid)*

Summary: *The methods for anonymization established in chapter 3 of the thesis are based on data given by electricity companies. This chapter presents a practical approach for generating synthetic distribution networks and datasets by combining public databases and data synthesis algorithms. A synthetic network is developed in an open-source QGIS platform by leveraging the open-data from local government databases, OpenStreetMaps, and mapping engines such as Google Street View. New data synthesis algorithms are proposed to create synthetic networks and datasets. The practical feasibility of the proposed solutions is demonstrated by an illustrative case study of the Colac region in Australia. Synthetic networks and datasets are created for the distribution feeder, and then evaluated on industry servers. The results are compared using a two-step validation procedure: comparing the synthetic and actual network datasets using geo-based visualizations and by incorporating feedback from industry experts familiar with the analysis. The comparison results demonstrate the efficacy of developed networks and datasets as they show resemblance to real network and datasets while providing the geographical validation of distribution network models. The procedure for creating synthetic test networks is illustrated graphically in the Figure. 4.1.*

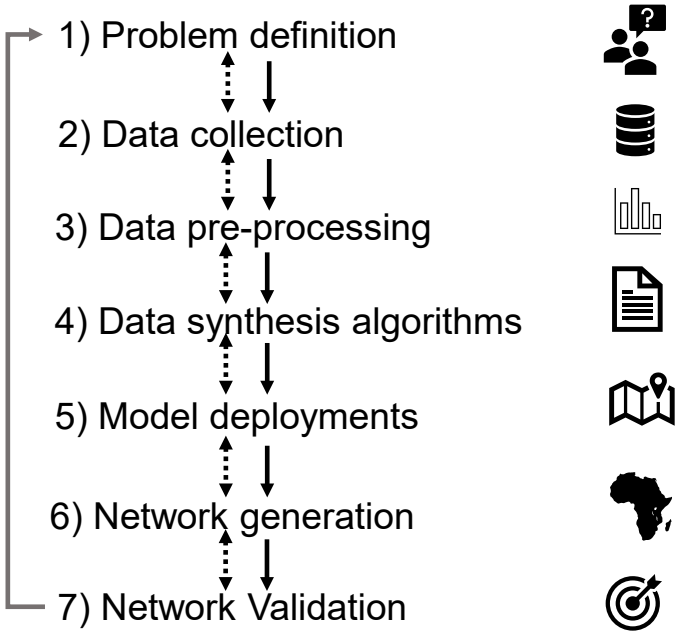


Figure 4.1: Process for generating synthetic networks

4.1 Research gaps and contributions:

The motivation for the work presented in this chapter is based on the research gaps presented in Chapter 2. The contributions made in this context are summarized below:

1. A methodology is proposed to design the topology of power distribution lines using public road infrastructure. The proposed method simplifies the design of power lines by using the concept of nodes and edges. This concept is supported in the power distribution planning book [175] and power system planners can leverage from this approach to select suitable routes for new power lines.

2. The geospatial locations of network elements are added to the generated datasets to fill the gaps in existing datasets. The importance of geographic information is indicated in [52, 53]. It is crucial for the planning, modeling and management of the assets of a distribution network. For instance, the geographical locations assist planners in installing new assets in existing networks.

3. A new way to create network datasets from publicly available platforms including local government energy databases, OpenStreetMaps (OSM) and mapping engines such as Google Street View is proposed. The opportunity to access these trustworthy public sources means that there is now a greater level of transparency than ever before, especially when it relates to government information.

4. An algorithm based on the virtual layer approach (FromNode and ToNode) concept is proposed to establish electrical connectivity between different components of distribution networks.

5. A new method for identifying the number of energy consumers supplied by a transformer in a distribution network is developed. A hub-line algorithm is demonstrated to connect energy consumers based on their nearest spatial distance to a transformer. A standard cut-off distance from the transformer to households is maintained by adhering to the guidelines in CIGRE publications [54].

6. Most existing solutions are designed for European and North American systems. Our research efforts concentrate on Australian distribution networks. As the proposed concept is applicable to both small- and large-scale networks, an attempt is made to address the problem of scalability

7. The practical validation of the proposed algorithms is demonstrated by an illustrative case study of the Colac region in Australia. A synthetic dataset is created for the distribution feeder, and datasets are deployed and visualized in industry servers. The results are then compared to the original feeder datasets to verify the applicability of the proposed techniques.

4.2 Methodology for creating synthetic distribution networks and datasets

The proposed framework consists of four main modules: the collection of information from public resources, the development of a synthetic distribution network, the formulation of new data synthesis algorithms for synthetic data generation, and its deployment in industrial servers for methodology validation. The four phases are shown in Figure 4.2, and the descriptions of each module are as follows:

- **Module 1:** In the first stage, the raw data from publicly available platforms is collected and processed for extracting the maximum level of information. This study

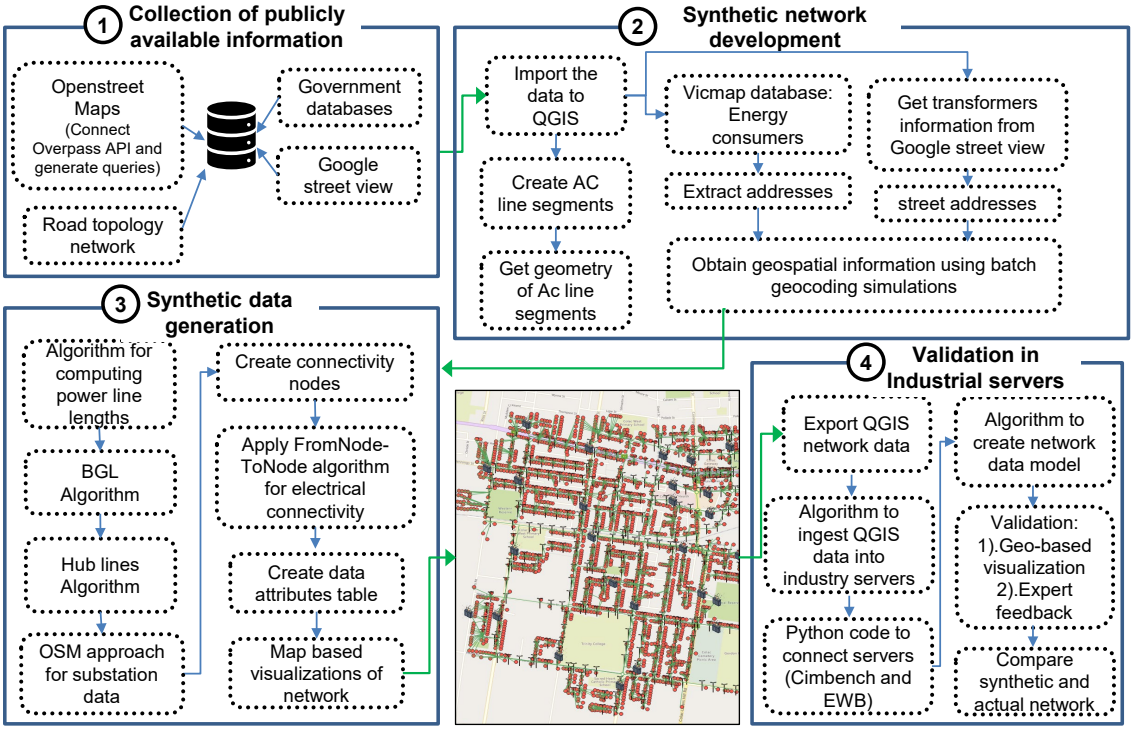


Figure 4.2: Framework of the proposed methodology

used the land and parcel data from the local government databases to design the topology of power distribution lines. The databases include spatial datamart [176], that provides topologically structured datasets of road networks. To minimize missing data, the data of OSM are also used simultaneously by generating overpass API queries for the relevant region. The critical information of energy consumers and building footprints are retrieved from a geocoded database of property address points. The data includes real life locational property addresses provided by local government. In the case of power transformers, the information is collected by leveraging the platform of Google Street View.

- **Module 2:** In the second stage, the acquired data is imported into an open-source application such as QGIS [177] for creating the synthetic network. The objective is to develop the structure of the distribution grid, and then create the required system components. The system components such as power lines (ac line segments) are created from the road network information. The lines are linked using the concept of nodes and edges, where nodes represent the intersection points of the lines and edges represent a link between two nodes. For energy consumers, the addresses of individual consumers are first retrieved. The address points are then translated into geometry coordinates (latitude, longitude) using batch geocoding simulations. The simulation generates geographical data for each energy consumer. The transformer information is obtained from Google Street View by searching for transformers on particular streets in a region.

- **Module 3:** This stage includes the data synthesis algorithms to complete the missing data of the entire network. A solution for estimating the lengths of power lines is proposed. A batch-geolocation algorithm (BGL) to identify the geographical locations of the energy consumers is developed. A hub-lines algorithm is designed to estimate the number of energy users connected to each transformer. The algorithm is based on the nearest neighbor concept, and more information is provided in Section 4.3. The substation information is retrieved from OSM approach by creating overpass queries in overpass-turbo portal [178]. The connectivity nodes data is generated with the concept of utility poles and lines associated with them. A FromNode and toNode technique is developed to establish electrical connectivity between different electrical components. The proposed techniques are then applied to create a

synthetic distribution network and datasets for a Colac region in Australia. The synthetic datasets describing system components are then stored as tables for further analysis. A map-based visualization is created to validate the geographical layout of the network.

- **Module 4:** The final step is to test the synthetic datasets in industrial servers and compare the results to the original feeder datasets. The synthetic data is translated into GeoJSON format, which is supported by industrial servers [179]. An algorithm is developed to automatically ingest data from QGIS into industry servers. The experiments are carried out on two servers, Cimcap and Energy workbench (EWB). The Cimcap converts synthetic data into CIM-based format by taking QGIS datasets as input, which is a widely used data standard in the power industry [55]. The EWB server translates the generated data into geo-based visualizations and enables the geo-validation of created datasets. The developed synthetic network and datasets are then compared with real distribution networks. The comparison is made by geo-based visualization in industry servers and feedback from industry experts who are familiar with the analysis.

4.3 Algorithms for synthetic network creation and data generation

This section describes the approach for developing synthetic networks and generating data. The details of the data synthesis methods are described in detail, with practical demonstrations.

4.3.1 Building power line topology

Since power lines generally follow road paths [180], the topology of power distribution lines is designed using public road infrastructure. Figure 4.3 illustrates the concept. On the left, the road network obtained from the local government databases is shown. On the right, power line paths are shown representing the real distribution network. The road transport network information is obtained from a spatial datamart from the local government, that offers topologically organized datasets of road networks. By integrating this

(a)

(b)

Figure 4.3: Illustration of (a) road network and (b) paths of power lines in real network

information, the suggested approach enables a realistic representation of the lines in a network.

The network is topologically structured by N nodes and E Edges as shown in Figure 4.4. The nodes (blue) indicate the places at which the lines intersect, while the edges (green) represent the link between two nodes. Given the importance of line length in distribution networks, an algorithm for calculating the length of each line segment is proposed. This method is based on the Euclidean distance, with the steps outlined in Algorithm 4.1. The input of the network is node points Np taken from the road network RN, while N represents the total number of node points in the network. The goal is to

compute the Euclidean distance $d_{eucl}[x, y]$ for each line segment which is the output of the network.

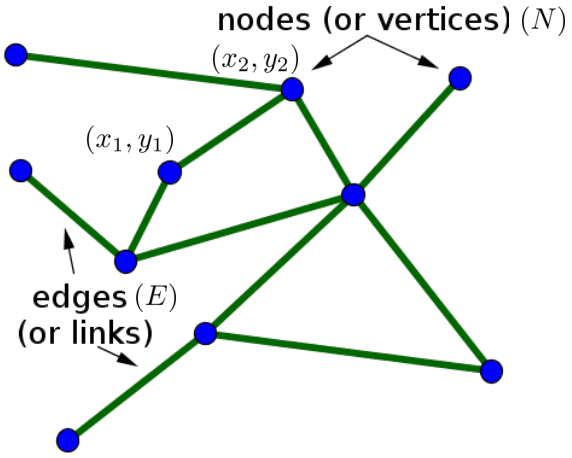


Figure 4.4: A concept of nodes and edges that forms power lines in the network.

Algorithm 4.1 Algorithm for computing lengths of line segments

for Computing L_p **do** ▷ L_p: Length of power lines
  Get Node points N_p
  for $N_p(x, y = 1, 2, \ldots n)$ **do**
    Get Road network R_N
    for R_N **do**
      Access land parcel data from G_d ▷ G_d: Government databases
      Filter the area based on postcode
      Processing on raw data ▷ Filter relevant columns
      Get the line features (nodes and edges)
    end for
    Origin $N_p(x_i)$, destination $N_p(y_i)$,
    compute: $d_{eucl}[x, y] \leftarrow dist(\mathrm{N}_p[x_i], N_p[y_i])$
  end for
  Output: L_p, estimated length of power lines
end for

4.3.2 Energy consumers data generation

The information of energy consumers is generated from databases of property address points obtained from local government land and planning departments. The database comprises real-world property addresses, household identifiers, and street addresses for broader application scenarios. The benefit of this approach is that the raw data is freely

Algorithm 4.2 BGL process for geo-locations of energy consumers

```
for Batch-geolocation do
    Get property address points A_p
    for A_p(i = 1, 2, ... n) do
        Connect geocoded government database
        Select the area based on postcode
        Get the locational property address identifiers
        Request the order and download property A_p
    end for
    for Bulk geolocation do
        Pass the shapefile of A_p to QGIS
        Add geometry attributes to each point feature.
        Compute geometric properties (x-y coordinate of the address point).
        Get the geographical coordinates (Lat, Long) of individual house
    end for
    Store the results in the output variable
    Map the geographical coordinates into online maps
end for
```

available without the requirement to register details or any licenses that limit how it may be used. As it contains only household addresses, their geographical locations are obtained by implementing a batch-geolocation (BGL) process in Algorithm 4.2. The first stage involves retrieving household addresses or property addresses Ap from the spatial datamart database. The second phase then transforms residential addresses into geographic coordinates (latitude and longitude). A bulk geolocation method is carried out to execute large numbers of addresses at their geographical locations. Each address is parsed to return geocoded locations. The output includes the entire address, location, and features like the postcodes. As geolocation services are usually not free for bulk addresses, the 'add geometry' function [181] in the QGIS tool [177] is used which provides free-of-cost bulk geocoding services. The results obtained from the BGL algorithm are then visualized using QGIS maps for geographical validation of energy consumers and the data are stored in the form of CSV files. The geographical locations may not be explicitly linked to the energy usage of individual buildings, their spatial position specifically defines energy usage among groups of buildings and may be used to identify high-consumption regions.

4.3.3 Power transformers data generation

The dataset of transformers is created by following a two-step procedure. In step one, the number of transformers in a specific region is identified using Google Street View. Then, those for each street are searched using the appropriate postcodes and their locations saved in data tables. In the second step, a hub-line algorithm is developed to identify the number of consumers connected to a single transformer. This establishes a hub (transformers) and the nearest feature in a destination layer (energy consumers). Figure 4.5 shows an illustration where hub (transformer) is denoted with p and the nearest objects (energy consumers) are presented with e_1, e_2, e_3, and e_4. The closest energy consumers from the hub are determined based on k-nearest neighboring concept [182], and it estimates the closest consumer based on spatial distance. The difficulty is maintaining a standard distance between a transformer and the households. The transformer should not connect energy users that are more than 2 kilometers away to avoid losses and voltage drops. To address this issue, we adhere to the guidelines outlined in CIGRE publications [54]. According to the CIGRE C6.24 document [16], if the rated capacity of a LV transformer is 400kVA, the average distance from it to households is 415.5m. Based on this definition, a cut-off distance whereby an energy consumer can be no more than 415 meters away from its nearest hub (transformer) is defined. Shorter distances can also be selected based on transformers' ratings and other requirements for the sizes of conductors and lines. The steps in this hub-line technique are given in Algorithm 4.3.

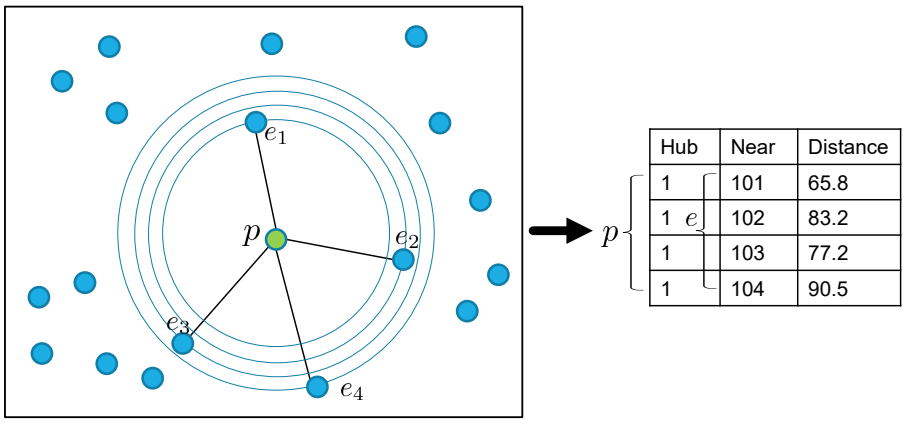


Figure 4.5: Illustration of closest energy consumers search from the main hub based on k-nearest neighbor

Algorithm 4.3 Hub lines Algorithm

procedure
  Step:1 Define the hub layer p and spoke layer e in the network.
  The hub is transformer and spoke layer is energy consumers
  Step:2 Define allocation criteria: Nearest feature N_f
  Step:3 Search for the N_f in e from p
  for N_f **do**
    Determine the distances d from p to each nearest features e
    $d(p,e) = \sqrt{\left(p_x - e_x^i\right)\right)^2 + \left(p_y - e_y^i\right)^2}$
    Find N_f from p within a radius, $r < 415$ m, ▷ CIGRE C6.24 standards [16]
    Record the d in the output table
  end for
  Step:5 Iterate the steps for each feature in the p and e
  Step:4 Based on N_f and d, connect p with e
  Step:6 Add connection lines on the map.
  Step:7 Count number of e connected to each hub p
end procedure

4.3.4 Substations and connectivity nodes

An up-to-date repository of substation information is retrieved from the OSM database which is accessible via the overpass-turbo portal [178]. Multiple queries based on query language are created in an overpass-turbo to identify the locations of the substations. The existence of substations in a given region is detected by adding a tag `power=substation` to the overpass-turbo wizard. Figure 4.6 demonstrates the approach for obtaining substation information. On the left, queries generated in overpass-turbo wizard are shown. The area id used in the query is given on the right. For instance, the area ID of the Colac region in OSM is 360314511. Based on the given ID, OSM gives the required information of substations in the area.

Figure 4.7 demonstrates the concept of connectivity nodes in a network where the end-users (loads) are connected using these nodes (poles). The connectivity nodes data is generated with the concept of utility poles and lines associated with them [183]. According to IEC 61970 CIM standards [184], a connectivity node is a point where the conducting equipment's terminals are all connected. For example, connectivity nodes are created by specifying the standard distance between two electric poles. For distribution networks (35 kV and less), typical spacings between two poles range from about 40–100m [185]. Based on this distance, connectivity nodes are created along the lines. The steps taken to

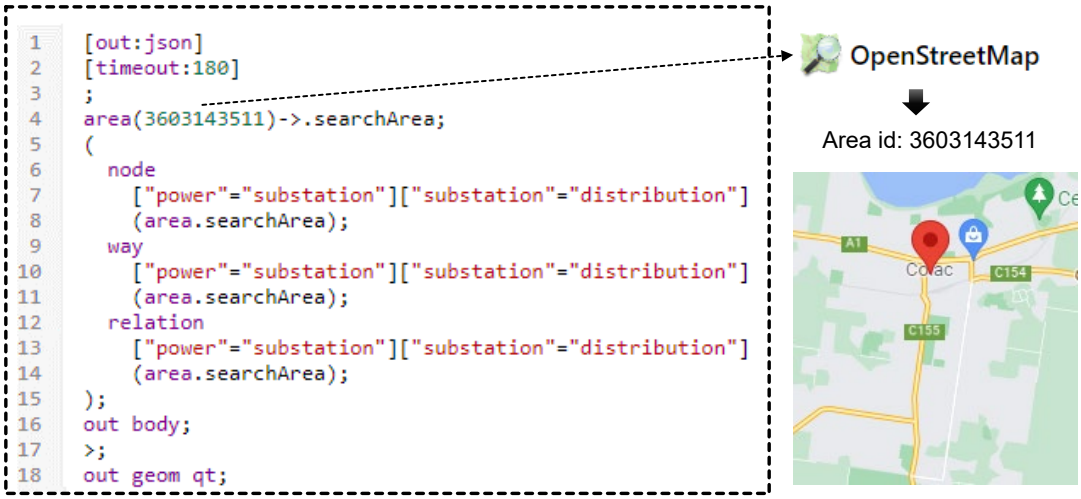


Figure 4.6: Approach to obtain substation data (a) Overpass queries in open street maps (left) and (b) Area id (right)

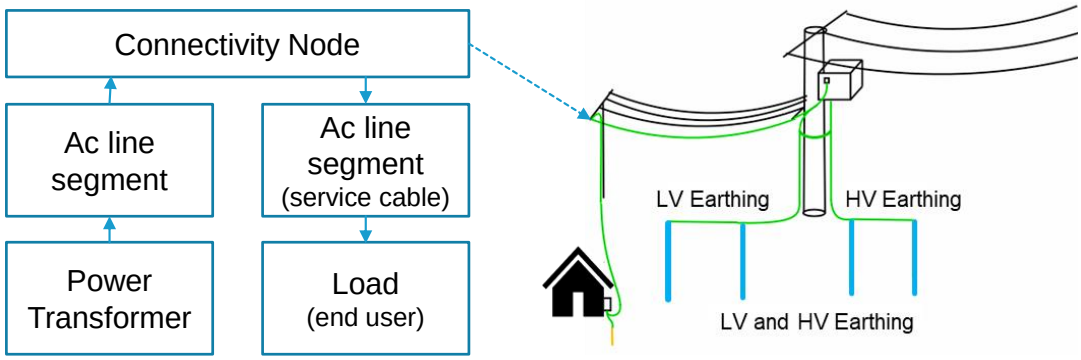


Figure 4.7: A concept of connectivity nodes in a network

achieve this are given in Algorithm 4.4. Initially, the length of each segment of a power line is found. Then, based on the estimated distance, connectivity nodes are created along with it at specified intervals. This approach is implemented in PyQGIS (a Python environment inside QGIS). Then, the interpolate technique from QgsGeometry [186] in QGIS is performed to assign connectivity nodes to lines. After creating connectivity nodes, the topology checker operation in QGIS is undertaken to identify disconnected nodes. Two rules are added, one for the connected nodes and one for the associated lines (links). Under rule 1, each link's end point must be covered by a node. Under rule 2, each node must be covered by the end point of a link. The resultant output is the proportion of connectivity nodes remaining in the network.

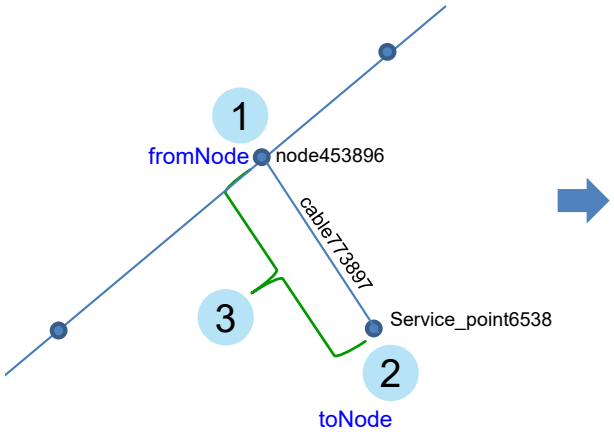
Algorithm 4.4 Connectivity nodes in the network

for Connectivity nodes generation, C_n **do**
  Get L_i^p, ▷ the length of each line segment,
  if L_i^p is <50 m **then**
    Create one C_n along L_i^p
  else if L_i^p is >50 m **then**
    Create two or more C_n along L_i^p
  end if
  Perform interpolate method from QgsGeometry [186]
  Identify disconnected nodes D_n with topology checker [187]
  for D_n **do**
    set rule 1: the end point of each line must be covered by a node
    set rule 2: each nodes must be covered by the endpoint of a line
  end for
  Remove duplicate connectivity nodes
  return C_n
end for

4.3.5 Algorithm for establishing electrical connectivity

This section provides a standard way of representing electrical connections between two or more equipment in the network. An algorithm based on the virtual layer approach (FromNode and ToNode) is proposed to establish electrical connectivity between different components of networks. Figure 4.8 illustrates the concept where a nearby connectivity node connects the service point (energy consumer). The connectivity is established by defining the starting point and ending point of network elements. The start and endpoints are defined with 'from' and 'to' nodes respectively. Each starting and ending point is assigned with a unique object id to prevent duplication or false connections.

Algorithm 4.5 explains the proposed methodology in detail. The 'from' and 'to' information of line layer L_{layer} is obtained using `st_startpoint(geometry)` and `st_endpoint(geometry)` functions [188]. For doing so, a virtual layer concept in QGIS is used. SQL queries are generated in the virtual layer to extract the required information. The query saves the start and endpoints in two columns. An algorithm is created in PyQGIS platform to create 'from' and 'to' information of node points n_p. Unique id,s are assigned to the data attribute and the process is enumerated for all features in a layer. The connectivity is then established by joining the FomNode and ToNode information. The connectivity is established based on the nearest feature criteria given in



Feature	Value
id	cable773897
class	Ac line segment
length	71.83
name	acls1
location	Queen Street
Creation date	4 Oct 2021
Network level	Distribution
fromNode	node453896
toNode	Service_point6538

Figure 4.8: Concept of FromNode and ToNode for establishing electrical connectivity

Algorithm 4.5 Electrical connectivity based on FromNode and ToNode concept

procedure
  for electrical connectivity between electrical components **do**
    Get *FromNode* f_n and *ToNode* t_n information
    Stage:1, line layer L_{layer}
    for f_n and t_n in L_{layer} **do** virtual layer query
      Select FromNode, *st_startpoint(geometry)* as geometryFrom,
      ToNode, *st_endpoint(geometry)* as geometryTO
      From L_{layer}
    end for
    Stage:2, node points n_p
    for f_n and t_n in n_p **do** PyQGIS
      set layer l= n_p
      Add data attributes *(FromNode, ToNode)* in l
      Update attributes in main l
      Assign object id's in data fields = l.fields()
      Enumerate for all layer feature, (l.getFeatures():
      Update the fields with n_p(count+1)
    end for
    Select the start point S and end point E of network elements.
    Define connectivity criteria: Nearest feature N_f
    Obtain N_f from Algorithm 4.3
    join S with E to build connectivity C_l
    Iterate the steps for each feature in S and E in the network.
    Add C_l on the map.
  end for
end procedure

Algorithm 4.3. The process is iterated for each feature in the network components and connectivity lines are added on the maps for further validation.

4.4 Case studies: Generation of a test network

In this section, the proposed techniques are applied to create a synthetic distribution system in the Colac region of Australia. A repository of network elements including power lines, transformers, energy consumers, substations, and connectivity nodes is created. This approach establishes grid connectivity and generates geographical locations of grid components to obtain a comprehensive test case. Additionally, datasets of network elements are constructed and visualized, and interactive maps for validation are created.

4.4.1 Synthetic test system: Colac area, Australia

The first step is to design power distribution lines in the network. For this purpose, public road infrastructure is used based on the concept presented in section 4.3.1. This generic design approach significantly increases their re-use potential and the ease they can be for any national context. Figure 4.9 shows the power lines for the Colac region with line segments shown on the street map. The network is topologically structured by nodes and edges. In total, the dataset consists of 311 nodes and 559 edges. Geospatial information of each node is derived and power-line characteristics, such as lengths, are computed

Figure 4.9: Topology of power lines created from public road infrastructure

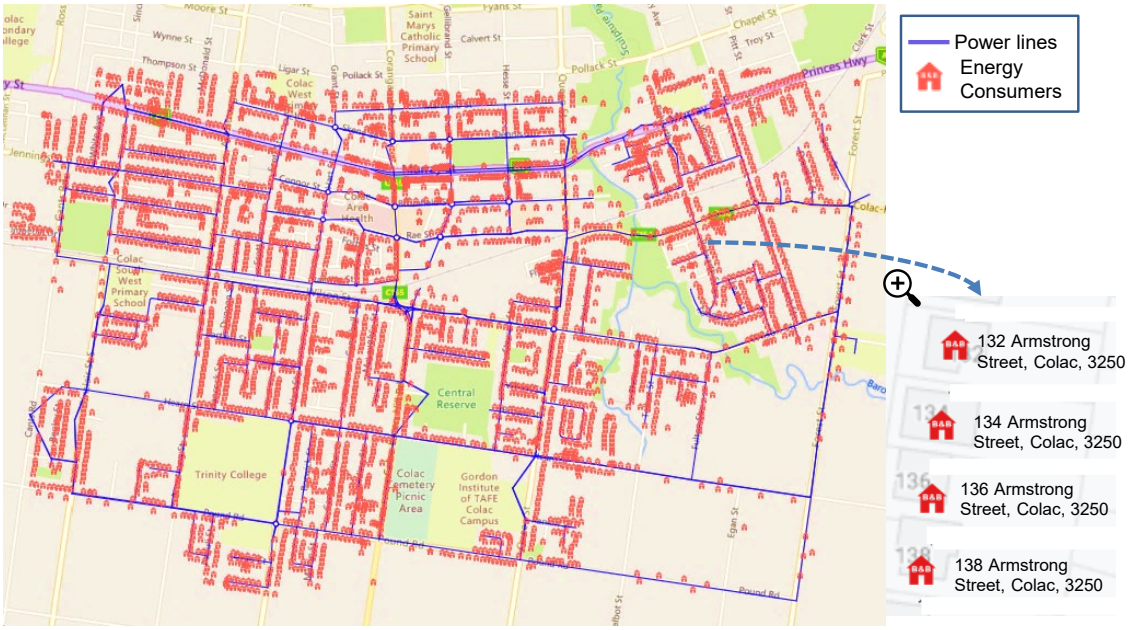


Figure 4.10: Energy consumers in Colac region, Australia.

using Algorithm 4.1. The data records are saved in .CSV files for further analysis and validation.

The second step is to create energy consumers in the network. The necessary information is obtained by querying a public local government energy database comprising real-world metadata of property addresses. The consumers in the test area are filtered based on the postcode of the Colac region which is 3250. In Figure 4.10, 4155 identified energy consumers, which are in a densely populated residential area, are shown. The points of the buildings closest to the streets represent consumers and are indicated by red dots. As this public database contains only the addresses of consumers, their spatial information (latitude, longitude) are obtained by implementing the BGL technique, as previously explained in Algorithm 5.2. A bulk geolocation methodology is used to convert large numbers of addresses in their geographical locations. The geospatial locations obtained are then examined with interactive network maps to assess the completeness and functionalities of the datasets derived. These geospatial locations are essential because they provide an overview of the networks topology and the distribution of its assets [189].

The third step is generating data for the distribution transformers. For this purpose, a Google Street View method is used because it supports the real-time locations of transformers that connect energy consumers. Figure 4.11 shows the location of the distribution transformers on the street map. In total, 48 distribution transformers are identified

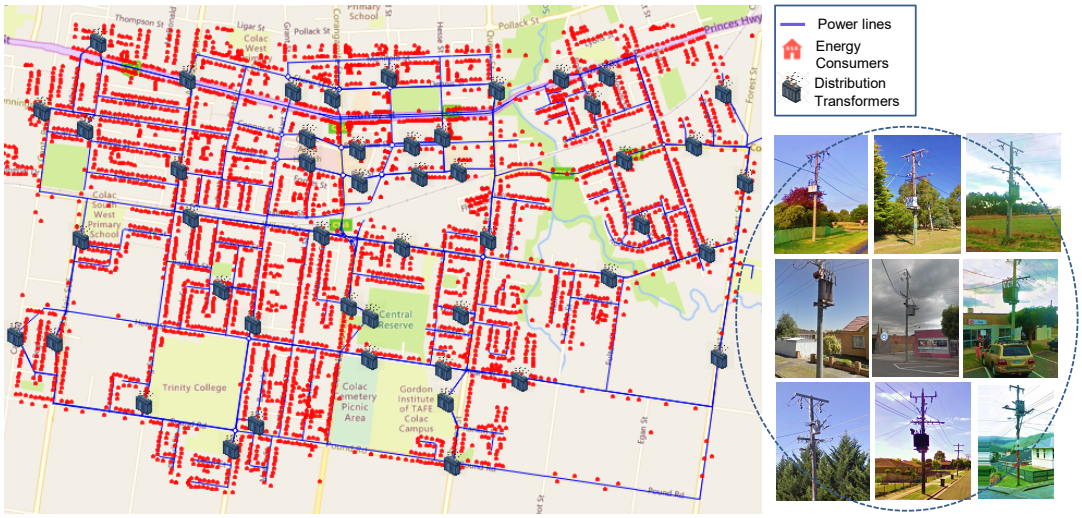


Figure 4.11: Power distribution transformers in the Colac region, Australia.

in the region. These are public transformers with a typical rating of 11 kV with primary and secondary voltages of 220V and 415V, respectively. These are public transformers with typical ratings of 11 kV and primary and secondary voltages of 220V and 415V, respectively. To determine how many customers are linked to a single one, the hub-line algorithm (Algorithm 4.3) discussed in section 4.3.3 is implemented. In Figure 4.12, its outputs, where each transformer is connected to its nearest energy consumers, are shown. The bar chart on the left side shows the location of each transformer with the number of connected energy consumers. On the right-hand side, a simple illustration of the energy consumers connected to the transformer located at 65 Sinclair St, Colac, Victoria, 3250 is provided. In total, 95 energy consumers are connected to this transformer which are shown with red dots on the right side of Figure 4.12.

The fourth step is to add substation information to the network. The substation data is created by querying the area from the OSM databases. A unique area id (3603143511) is used to identify the substation of a certain region. A filter tag of `"power=substation"` is applied to locate the substations in the area. The query `["power"="substation"]["substation"="distribution"]` returns all the distribution transformers in the region (3603143511) which represent Colac, Australia. In Figure 4.13, the results of the substation information are presented. On the turbo left generated in the overpass wizard. On the right, the query results with additional substation information such as voltages, names, indoor-outdoor locations, and operators are shown. The graphical results reveal that there is a 66/22kV substation in the Colac area operated by a distribution company

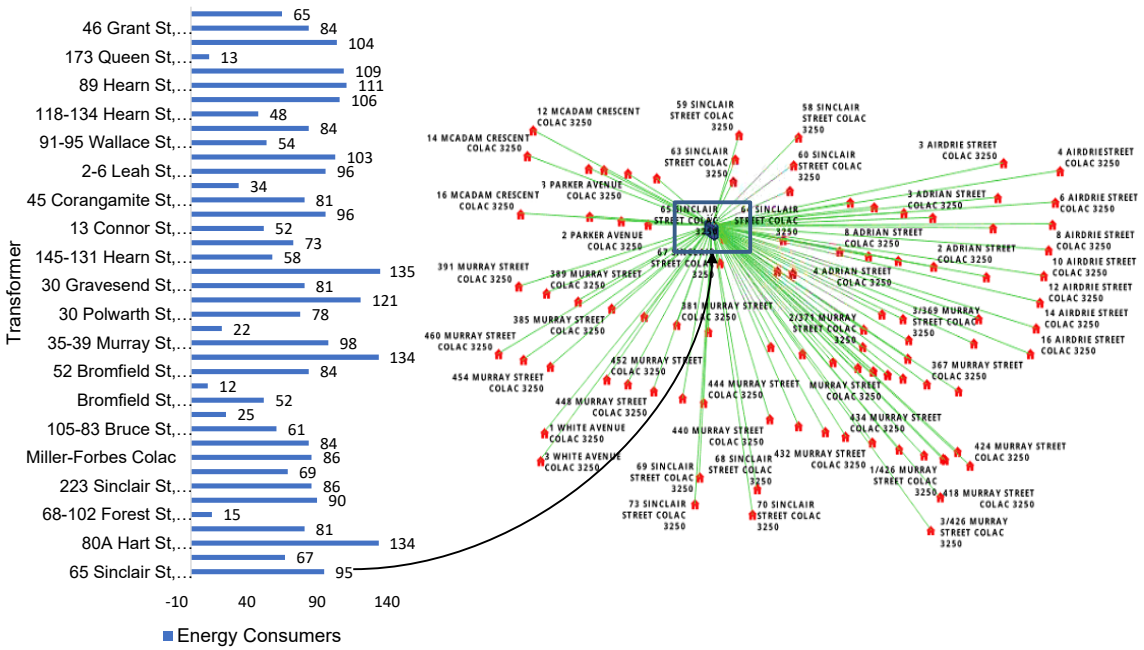


Figure 4.12: Hub-line algorithm for identifying the number of energy consumers linked to distribution transformers

Figure 4.13: Substation data: (left) overpass queries in open street maps; and (right) query results

called Powercor.

The last stage is to establish connectivity (links) between different network elements in the datasets. For example, details on how buildings are connected to a connectivity node (pole) on the street and how that pole is connected to the distribution transformer are provided. Connectivity is established by the 'FomNode' and 'ToNode' concept, as discussed in section 4.3.5. Figure 4.14 shows a simple illustration of the structure for connecting grid components. For instance, to connect a utility pole with an end-consumer, the 'FromNode' and 'ToNode' information of the pole and energy user, respectively, are used to connect the two grid components. Each consumer is connected to its nearest pole which is connected to the distribution transformer via low-voltage lines.

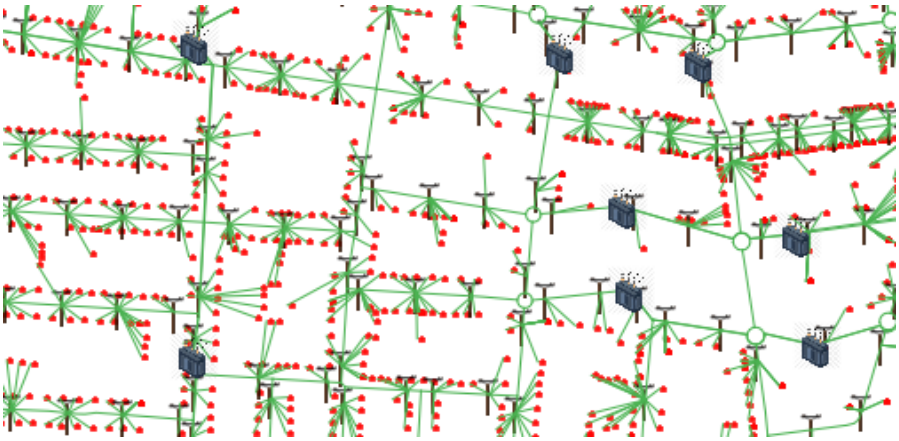


Figure 4.14: Demonstration of electrical connectivity between different network elements

Once all network elements are created, the complete network is visualized in QGIS software, as shown in Figure 4.15. The catalog of network elements includes 4714 power lines, 48 distribution transformers, 4155 energy consumers, 609 electrical nodes, and one substation. The satellite view of the synthetic network is shown in Figure 4.16. A network data repository is then established by saving the files in a `.geojson` format. This format is used because most industry servers can read and write geojson files, and it is simple to validate and test [179]. The dataset files are also structured and prepared as `.CSV`, with the file name indicating what it contains. To clarify the data, its attributes in the dataset files are explained in Table 4.1.

Figure 4.15: Complete visualization of synthetic distribution network in QGIS

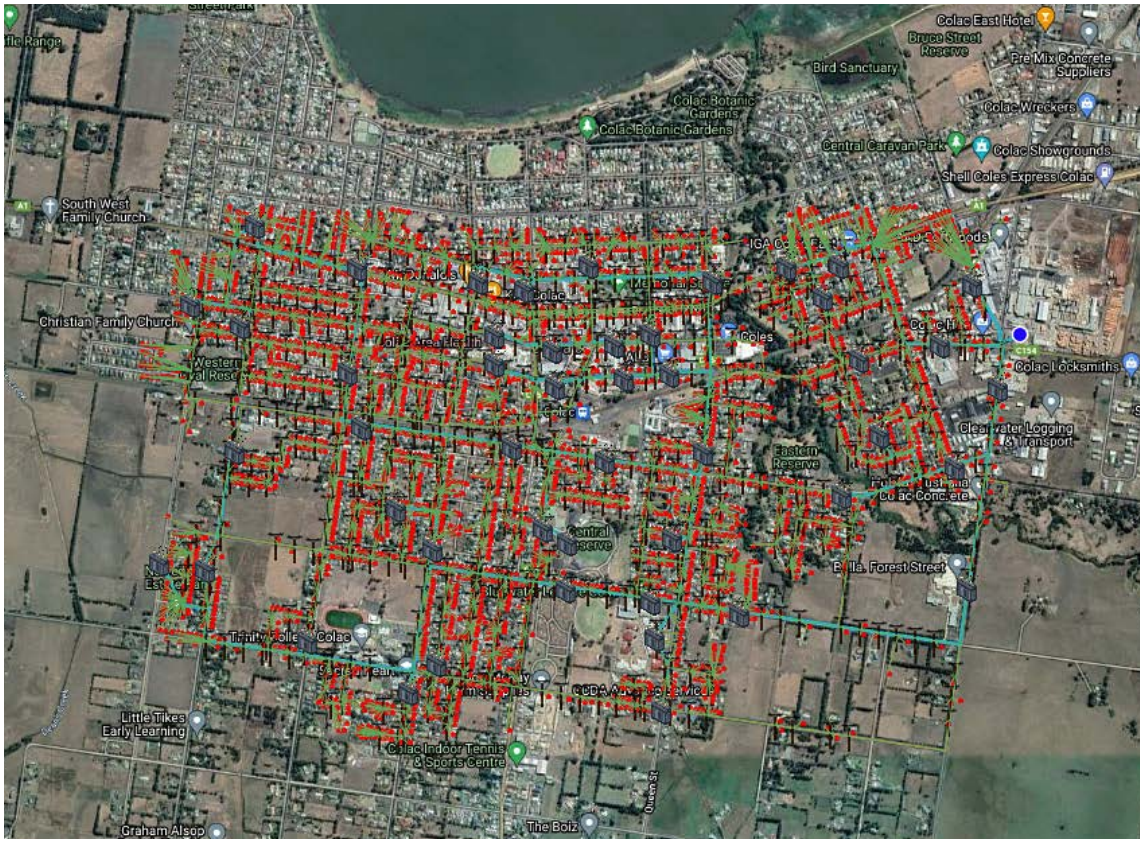


Figure 4.16: Satellite view of synthetic network developed for Colac region in Victoria

Table 4.1: Data attributes in datasets

Attribute	**Description**	**Example**
• Object ID	identification index of nodes and edges	'4571'
• Name	label of the equipment	'81 Murray Street, Colac, 3250'
• Lat	latitude of the node	'-38.342835390'
• Lon	longitude of the node	'143.605306759'
• Global ID	unique identifier of equipment	'Line4572'
• Class	equipment type	'transformer', 'Energy consumer'
• FromEq	starting point for connection 'Source'	'node1854088615'
• ToEq	end point for connection 'Target'	'node1854088612'
• BaseVoltage	voltage levels of lines and equipment	'415V, 22kV'
• headTerminal	power source of network	'Substation'
• Length	length of line segments (m)	'78.13'

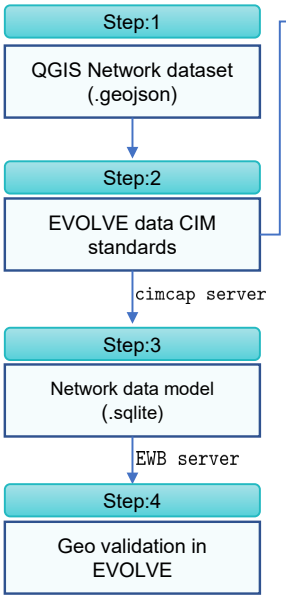
4.5 Comparisons and validation in industry servers

A two-step procedure is used to validate the synthetic network and datasets: 1) energy utility industry servers; and 2) expert feedback and evaluation. The validation process compares the synthetic and real distribution systems. They also aim to test the generated datasets for their completeness, consistency, and usefulness. While validating the synthetic network and datasets, geographical network maps are constructed to assist in verifying the logical consistency of the relationships among elements.

4.5.1 Validation 1: Industry servers

This procedure involves the four steps shown in Figure 4.17. Firstly, a QGIS network dataset is exported into `.geojson` format. Secondly, it is transformed into evolve data server standards [190] by implementing Algorithm 4.6. This step is important because industry servers, such as EWB, require a certain data file structure in order to run. The developed method converts synthetic data into a Common Information Model (CIM)-based format, which is a widely used data standard in the power industry [55]. A group of equipment is converted into evolve CIM classes by importing the utility libraries such as `import zepben.evolve as ev`. These libraries map the GIS standard data into CIM based schema. The `ev.location` class converts the geographical locations into evolve CIM data standard which is `ev.PositionPoint(coord[0],coord[1]))`. The position points represent the geographical data such as latitude and longitude. Thirdly, a Cimcap server is executed in parallel to map the dataset into SQLite network data model (`.sqlite`). Fourthly, this model is then loaded into an EWB server that converts the synthetic data into geo-based visualizations and enables geographical validation of created network and datasets. For more details about the evolve platform and servers, please refer to [190].

The experimental setup used for dataset validation is shown in Figure 4.18. The procedure begins with the execution of the Cimcap server, which then communicates with the data ingestion process in stage 2. The second stage is a data ingestion procedure to EVOLVE server as shown in Algorithm 4.6. This step translates the QGIS data to an EWB compatible database by interacting with the Cimcap server at stage 1. The created SQLite network data model is then simulated in the EWB server to validate the synthetic datasets using geographical visualizations.



Algorithm 4.6 Data ingestion from QGIS (.geojson) to Evolve data server

procedure
  Read synthetic network `.geojson file`
  Create SQLite database $SQL_{database}$ according to CIM standards
  for $SQL_{database}$ **do**
    Define Evolve *(ev)* cim classes
    Add location to network elements, *(ev.PositionPoint(coord[0],[1])*
    Assign base voltages, *(ev.BaseVoltage)*
    Create unique mRID of network equipments, *(eq.mrid)*
    Create energy source,*(ev.EnergySource)*
    Add power transformer *(ev.PowerTransformer)*
    Detect head equipment, *field_name("headTerminal") == 1*
    Establish equipment connectivity, *Connecting:("fromEq"), ("toEq")*
    Configure local host and server port, *default="50051"*
    Parse security certificate to connect servers, *(pkey=key, cert=cert)*
  end for
  Output: `network-model.sqlite`
end procedure

Figure 4.17: Procedure for technical validation of synthetic distribution network in utility servers

4.5.1.1 Demonstration using a small network

In Figure 4.19, the synthetic network for a small region in Colac, Australia, is shown. The synthetic network is created on a QGIS platform and datasets of network elements are extracted in different formats such as `.csv` and `.geojson` files. This synthetic data is then deployed in evolve data server platform and the results are compared with the real utility network as shown in Figures 4.19(a), 4.19(b). The network begins from the energy source (substation), then high voltage (11kV) to low (415V) power lines to feed the energy consumers. The junction represents the node (pole) that connects the end-users. The comparison results demonstrate that the dataset created for the synthetic network has similar features to those of the real utility one, as verified by both geo-visualizations and the data attributes of both datasets.

4.5.1.2 Demonstration using large-scale network

The proposed methodologies are then applied to create a large-scale synthetic system. The full synthetic system is illustrated in Figure 4.20. This network is designed for the distribution feeder in the Colac region in Victoria, Australia, with datasets simulated in industry servers. The results obtained are then compared with those of the original feeder datasets to verify the applicability of the proposed techniques. The objective is to

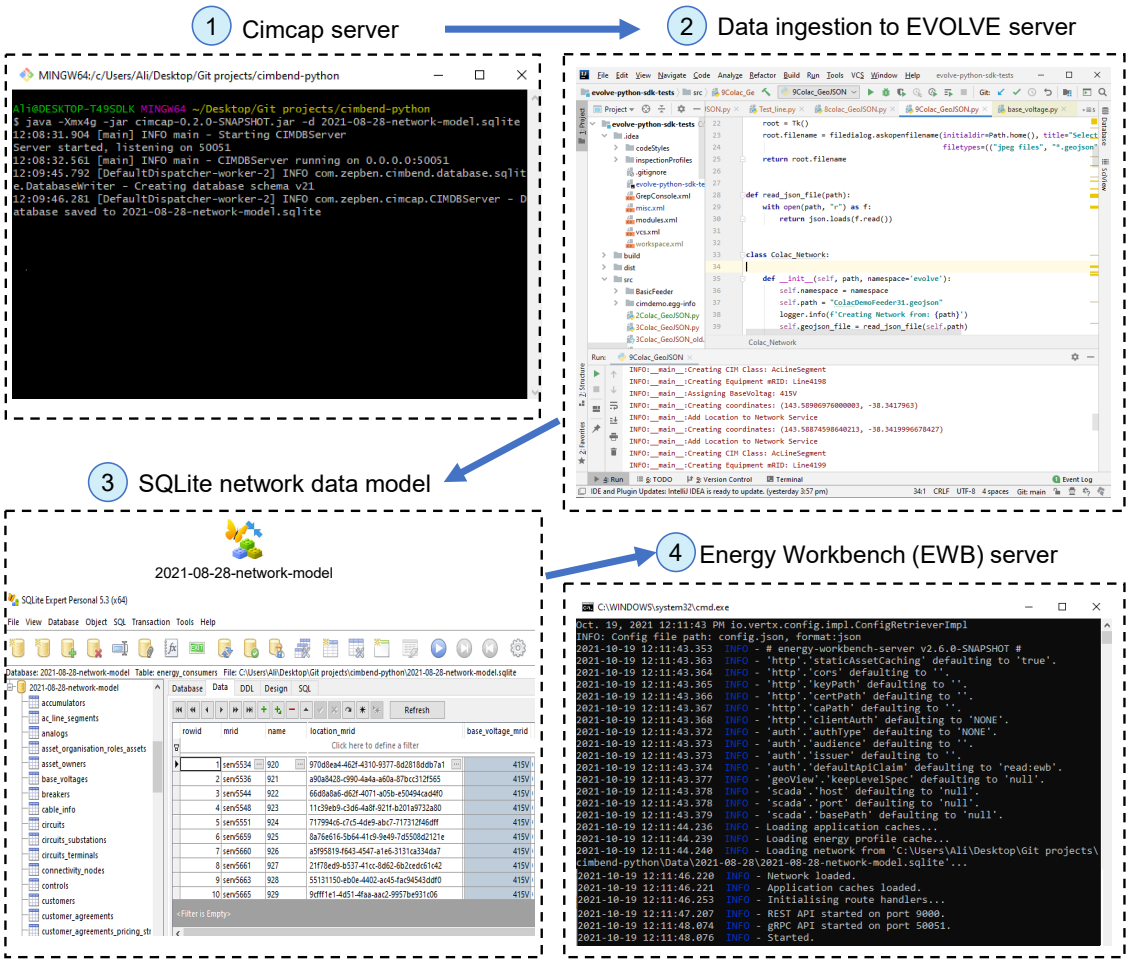
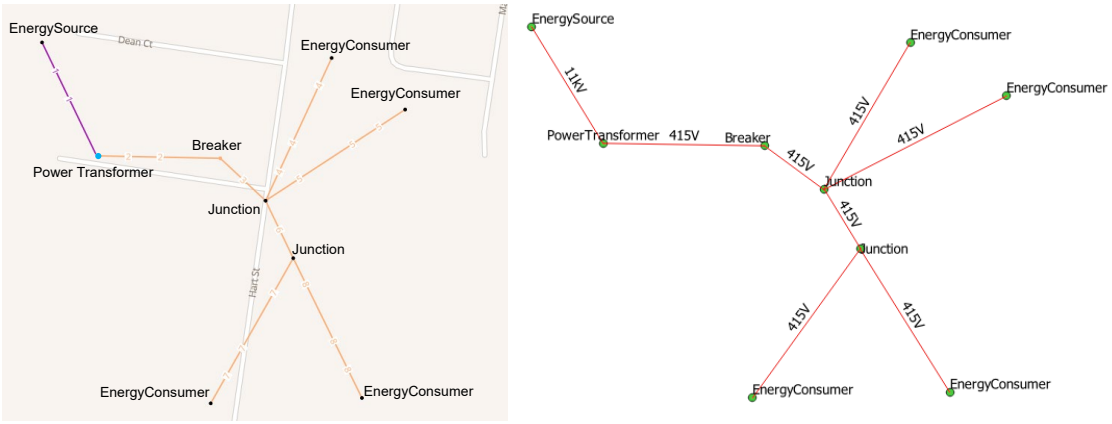


Figure 4.18: Experimental setup used in the dataset validation procedure

demonstrate how the topology of synthetic network and dataset resemble with the data of distribution network operator (industry). For validation, the synthetic network and datasets are supported with map-based visualizations that consider real street maps as shown in Figures 4.20(a), 4.20(b), 4.20(c) and 4.20(d). The visualization shows the entire network as well as zoomed-in details of individual components. In Figure 4.20(c), the individual element of real network in industrial servers such as transformer with the asset id “20527638”, supply point (energy consumer) with the id“32817388”, an energy source with the id “15400706”, the HV line with the id “62137567” and other parts are presented in detail. In Figure 4.20(d), the individual component of the synthetic network is illustrated. For example, the energy source is shown with asset name “Energy Source1” that represents the substation in the network, energy consumers with service points “servi9242”, the HV line “Line4675” and LV lines “Line5576”, power transformers with “PowerTransformer16”, and the connectivity nodes are presented with asset id



(a) Real utility network in evolve platform (industry)

(b) Synthetic network for a small region in Colac area

Figure 4.19: Validation on a small network

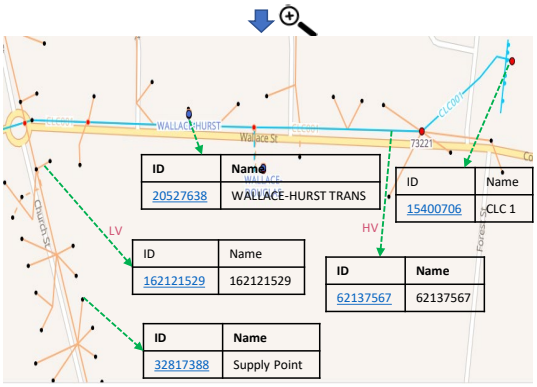
"node498". The comparison leads to two conclusions. Firstly, the synthetic datasets accurately reflect the data obtained from the distribution network service provider (DNSP), as evidenced by geo-based visualizations. Secondly, the synthetic datasets appear complete and dense, whereas the industry datasets are usually outdated. As the population grows, more energy consumers are added every day, which is missing in actual networks. The use of street maps and geo-visualization of networks and data synthesis algorithms in industry platforms enables the realism of synthetic networks that are comparable to actual distribution networks.

4.5.2 Validation 2: Expert feedback and evaluation

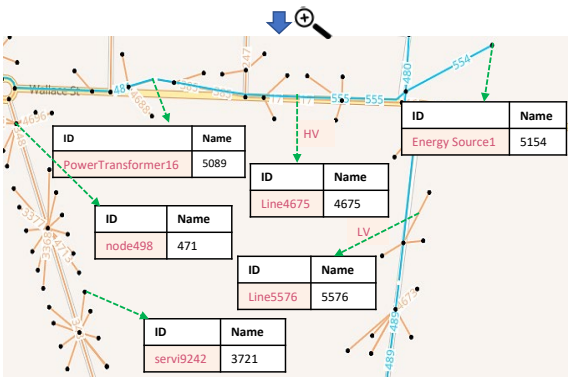
This validation is performed by industry and academic experts providing their feedback based on the methods adopted, datasets created and visualizations of results in industry servers. The suggestions include the number of customers connected per distribution transformer, synthetic datasets ingestion to proven industry servers, the configuration and validation of industry servers, industry data standard practices such as CIM, network elements at relevant locations, a connection topology in the network, standard voltage levels of the equipment, lengths of distribution lines and geographic representations. These improvements suggested throughout the expert validation procedure greatly improved the realism of the synthetic systems. Such details in the system creation process are difficult to find in real utility data. The advantage of this technique is that it eliminates the com-

(a) Real network (industry) for Colac area

(b) Synthetic network for larger region in Colac



(c) Zoom-in information of real network



(d) Zoom-in information of synthetic network

Figure 4.20: Validation using a larger network

putationally expensive necessity of simulating large number of real-world networks by including input from industry specialists who are familiar with the analysis.

4.6 Discussion and findings

Based on the results obtained in this chapter, the following comments are made.

- An undressed problem of synthetic network and open-source datasets that represent the real distribution system is addressed. The aim is to remove the barriers that exist between academic researchers and industry users when it comes to sharing open data for the evaluation and testing of newly developed algorithms. The use of open-data from government databases, OpenStreetMaps, geo-visualization of networks and data synthesis algorithms in industry platforms enable the establishment of realistic synthetic networks that are comparable to actual distribution ones.

- New ideas and solutions are integrated for the development of power distribution networks. For instance, power lines are designed by using public road infrastructure. Also, information of the end-users (energy consumers) on the distribution side, which was lacking in previous studies, is added.

- A distinctive characteristic of distribution networks that was not properly addressed in existing studies is the geographical structure of the system. Using prior works as a foundation, this study includes the geospatial locations of a network's elements, with each component having a specific geographical position. Although geospatial information is not always included in typical power system models, geographic coordinates allow market models to be benchmarked against realistic outcomes [189].

- A new method for identifying the number of energy consumers supplied by a transformer in a distribution network is developed. For example, the hub-line algorithm demonstrated in section 4.3.3 connects energy consumers based on their nearest spatial distance to a transformer. A standard cut-off distance from the transformer to households is maintained by adhering to the guidelines in CIGRE publications [54] to avoid losses and voltage drops.

- A standard way of representing electrical connectivity between two or more equipment's in the network is provided by proposing fromNode and ToNode concept. Connectivity is established by defining the start and end points of the network's elements.

- The methodology is applied to the Australian case because most test cases in the literature typically focused on European and American grids, with no test cases or data representations developed for Australian networks.

- The practical feasibility of the proposed algorithms is demonstrated by an illustrative case study of the Colac region in Australia. A synthetic dataset is created for the distribution feeder, and the datasets are deployed in the industry servers. The results are then compared to the original feeder datasets to verify the applicability of the proposed techniques. Case studies demonstrate that the synthetic distribution system has similar data characteristics to actual networks as validated by geo-based visualizations.

- Unlike previous works that validated synthetic networks in only a statistical manner [1], our solutions are tested using a two-step validation process. In the first stage, solutions are validated by replicating them in real-world industrial servers, and in the second stage, solutions are verified using expert comments and validation. This method contributes to expanding the utility of synthetic network and datasets beyond university researchers to industry users.

- The datasets are created with interactive maps in both QGIS and evolve data server platforms allowing users to manage and visualize the key assets in an existing energy infrastructure.

- The synthetic network creation and visualizations are developed entirely using open-source software which enables the wider research community to reproduce or improve the presented results. The created synthetic networks exhibit the critical electrical characteristics of real-world networks. However, they are entirely fictitious, and users cannot extract any actual network information from synthetic networks by reverse engineering.

4.7 Chapter conclusions

This chapter develops a comprehensive framework for creating synthetic power distribution networks and datasets by integrating public databases and data synthesis algorithms. Firstly, raw data from public databases is retrieved and processed for extracting the maximum level of information. Using this information, a synthetic network is developed in open-source software such as QGIS. Data synthesis algorithms are proposed to complete the entire network and then validated in industry severs. During network creation, a concept of road network topology is proposed to create the power distribution lines. A BGL algorithm is proposed to identify the geographical locations of energy consumers. A hub-line method is developed to identify the number of energy consumers linked to a distribution transformer in a network. The information of substations is retrieved from the OSM approach. A standard way of establishing electrical connectivity between two or more equipments in the network is provided by proposing fromNode and ToNode concept. The chapter demonstrates the practical feasibility of the proposed solutions by evaluating the synthetic network and datasets on industrial servers. The validation is demonstrated

by creating synthetic datasets for the distribution feeder in the Colac region of Australia. The results of the synthetic datasets are then compared to the actual feeder datasets. The comparison results indicate the effectiveness of created synthetic network and datasets, as they are validated by geo-based visualizations on industrial servers and including the expert feedback analysis.

The findings of this research will contribute to the growing field of industrial information integration and informatization in current distribution networks. The proposed approaches are generic in the sense that they are not limited to a specific region. It is possible to create synthetic test cases for any geographical region as the developed methods are based on OSM databases that facilitate public data sharing for the entire planet. This generic approach permits overcoming the issues related to the dimensions and diversity of distribution systems in different national contexts. Using synthetic network and datasets, utilities will no longer be concerned about making data publicly available at the request of industry and academia.

Network creation and all dataset availability

Network creation and all datasets are available on `https://github.com/casemsee/Synthetic-Network-Creation-and-Datasets.git` for improvements and reproducibility. We used freely available open access tools such as QGIS and PyQGIS to develop solutions. During data creation, we mainly used Python 3.5 along with Numpy 1.9.1, Fiona 1.8.18, Dataclassy 0.6.2, Libgdal 2.66.4, Osmnx, and Sharply 1.7.1. For industry servers and simulations, we used DNSP libraries such as Zepben-evolve 0.21.0, Zepben-cimbend 0.16.0, and Zepben-protobuf 0.10. For industrial database integration and SQL queries, we used SQLite express software. To visualize the network, we used geopandas 0.19.2, QGIS interactive maps, Mapbox studio, and Matplotlib 1.4.2. For geopspatial datasets and government databases, we used Overpass API queries in overpass-turbo wizard of openstreetmaps.

Chapter 5

Buildings in synthetic network: A multi-stage transfer learning approach for classification of building load profiles

The work presented in this chapter is submitted or published in the listed articles:

1. **[Journal]** M. Ali, K. Prakash, MQ. Raza, AK Bashir, and HR. Pota, "Load profile classification of buildings using AI-based multi-stage transfer learning approach for sustainable energy future," (Under-review) ***Elsevier Energy***, May. 2022. **IF: 7.147**.

2. **[Conference]** M. Ali, K. Prakash, and H. Pota, "M. Ali, K. Prakash, and H. Pota, "A Bayesian approach based on acquisition function for optimal selection of deep learning hyperparameters: a case study with energy management data ," Apr. 2020, vol. 2, no. 1. *Science Proceedings Series*

Summary: *The synthetic network developed in chapter 4 contains end-users (buildings). The load profiles of buildings have a substantial influence on distribution network operations, especially in the current situation of COVID-19 where most people work from home and spend most of their time in buildings. To effectively manage and optimize the energy consumption of buildings, a careful analysis of the building load profiles is essential. This chapter presents a deep learning solution to address the problem of electric load profile classification in the context of buildings. An unsupervised feature extraction*

process based on sparse autoencoders (SAE) is developed to automatically learn useful features from the data. A layer wise multi-stage transfer learning (MSTL) approach is proposed by combining unsupervised and supervised learning to improve the classification accuracy. To address the problem of missing data and class imbalance, a minority over sampling algorithm is presented, that effectively balances missing or unbalanced data by equalizing minority and majority samples for fair comparisons. The practical feasibility of the proposed approach is demonstrated by presenting two case studies. In case study 1, the techniques are evaluated on public benchmark datasets of buildings. In case study 2, the results are validated using real-world datasets of 105 buildings (35 residential, 35 commercial, and 35 industrial). The results indicate an overall accuracy of 93 percent on benchmark and 85 percent on real buildings datasets despite the high variations in building load profiles. An empirical comparison with the proposed method is carried out with popular machine learning methods, including support vector machine, random forest, k-nearest neighbors, and naive Bayes. The comparison results indicate superior performance over traditional machine learning methods, with a performance improvement that ranges from 1 to 10 percent.

5.1 Research gaps and contributions:

This chapter has made the following contributions to address the research gaps presented in Chapter 2.

- Two new methods based on sparse autoencoders (SAEs) and the multi-stage transfer learning (MSTL) are proposed. Different from conventional hand-crafted feature representation, SAEs can learn useful features from a large number of buildings data in an unsupervised automatic way. This is important since each building has unique electrical load patterns, and manually extracting the key features of every building is not possible in practical situations. A MSTL approach is applied to enhance the classification accuracy by combining sequential unsupervised and supervised learning.

- A minority oversampling (MOS) method is proposed to handle the incomplete or unbalanced real-world data. The data inconsistencies can produce error-prone results in developed models [149]. The MOS algorithm successfully balances missing or unbalanced data by equalizing minority and majority samples for fair comparisons.

- Three groups of buildings (residential, commercial, and industrial) are investigated for load classification. In the existing literature, only one type of reference building was considered for classification analysis. Using only one or a few typical profiles might not be sufficient to characterize a building's operational patterns [128].

- The load profiles of buildings vary from region to region. The practical feasibility of the proposed approach is demonstrated by presenting two case studies. Case study 1 involves testing the algorithms on the public benchmark dataset of buildings [191]. Case study 2 validates the results using real-world datasets of 105 buildings (residential, commercial, and industrial).

- An empirical comparison is made with the most widely used machine learning methods including SVM, RF, kNN, and naive Bayes (NB). Standard performance metrics such as a confusion matrix, receiver operating characteristic (ROC) curves, precision, recall, specificity, and F1-score are used to compare the findings. For a fair assessment, an average percentage performance improvement obtained by the proposed scheme over traditional machine learning approaches is computed.

5.2 Load profiling classification solution for buildings in the distribution network

In Figure 5.1, an overview of the proposed framework is illustrated. It consists of four parts: (1) data analytics on raw data; (2) a feature learning framework; (3) a multi-stage transfer learning approach; and (4) a case study of two datasets that evaluates the model's performances and visualizes its results. A short description of each mechanism is provided in the following paragraphs.

- *Step 1:* The first step is to clean the dataset from invalid and missing values. This is important because real-world data is often incomplete or inconsistent, and it is likely to generate error-prone results. In this study, two cases (datasets) are used to validate the solutions. In the first scenario, a publicly available benchmark dataset [191] of electric load profiles of buildings is used to test the proposed algorithms. This dataset contains the hourly load profiles over a year (8760 hours data) for a set of 24 representative facilities from various end-use sectors, including

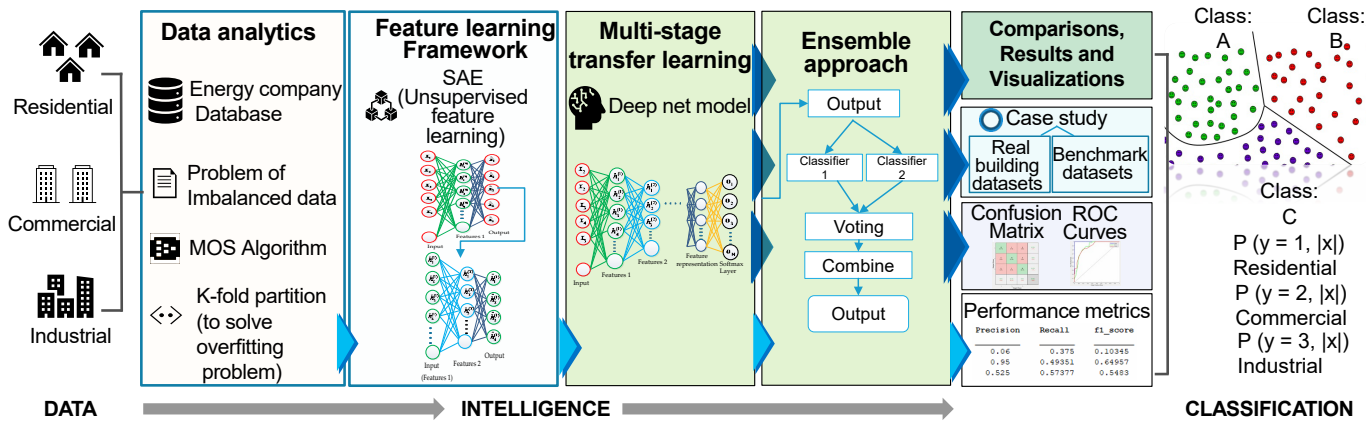


Figure 5.1: Proposed framework.

industrial, commercial, and residential consumers. During data-processing, the null data and missing values are removed to ensure the high quality of data. In the second scenario, realistic load profiles of 105 buildings are obtained from a distribution system operator in Australia. The data includes the energy consumption values of the buildings monitored in five-minute time steps.

- *Step 1.1:* In raw data, class imbalance and missing data are common issues that have been actively discussed in recent literature [192, 193]. A dataset is unbalanced if the classification classes are not equally represented. In practice, commercial and industrial buildings are fewer in numbers than residential buildings. In other words, residential buildings are more dominant (majority class) compared to commercial and industrial buildings (minority class). If the machine learning model is tested on an unbalance dataset, this leads the classifier to be biased in favor of the majority classes and consequently suffers from higher misclassification on the minority classes [193]. To address this issue, a MOS technique is applied to the raw data of buildings to balance the minority and majority samples. Firstly, k-nearest neighbors close to the minority class are identified. Then, new synthetic samples S are created along with the line segments by taking the difference D between the actual minority sample O_s and its nearest neighbor K. This difference is then multiplied by a random value ranging from 0 to 1 and added to the original minority sample O_s. The output from this method is a balanced dataset with an equal distribution of samples in each class. The intuition behind the MOS is summarized in Algorithm 5.1 and visually demonstrated in Figure 5.2. This approach is inspired from the oversampling strategy suggested in [194] and is effective because it creates new

synthetic instances from the minority class that are realistic to existing minority class samples.

- *Step 2:* This stage involves the development of an unsupervised feature learning framework based on SAEs that automatically discovers useful features in the data for classification. The motivation is to discover the comprehensive features in electric load profiles of building as every building has unique load characteristics. These features are then employed to classify the electric load profiles of buildings. The overall procedure is briefly explained in Section 5.3.

- *Step 3:* Following the feature learning process, a multi-stage transfer learning strategy is developed. It incorporates both unsupervised and supervised learning process to fine-tune the results and improve overall performance. The output of MSTL is then utilized to integrate the ensemble classification. It determines the best accuracy for each class by using the majority voting mechanism. Details of these techniques are provided in Section 5.4.

- *Step 4:* The practicality of the proposed approach is demonstrated using two case

Algorithm 5.1 Proposed MOS framework

MOS (M, J, N, K)
Input:
M: minority samples;
J: majority samples;
N: number of new samples;
K: number of nearest neighbours
Output: Synthetic samples for minority classes ($S = len(J)$)
for $i \leftarrow 1$ to M **do**
    Compute k nearest neighbors for i^{th} minority samples,
    Save the indices in the NA. ▷ NA: new array
    Populate (N, i, NA)
end for
for Creating synthetic samples for minority class **do**
    Compute: D = O_s - K ▷ D: difference, O_s: actual minority samples
    Compute: G = Generate random number between [0,1] ▷ G: gap
    Synthetic samples = $O_s + D * G$
    Repeat step $10 - 12$ until desired proportion of minority class reached.
end for
return (Balanced dataset)

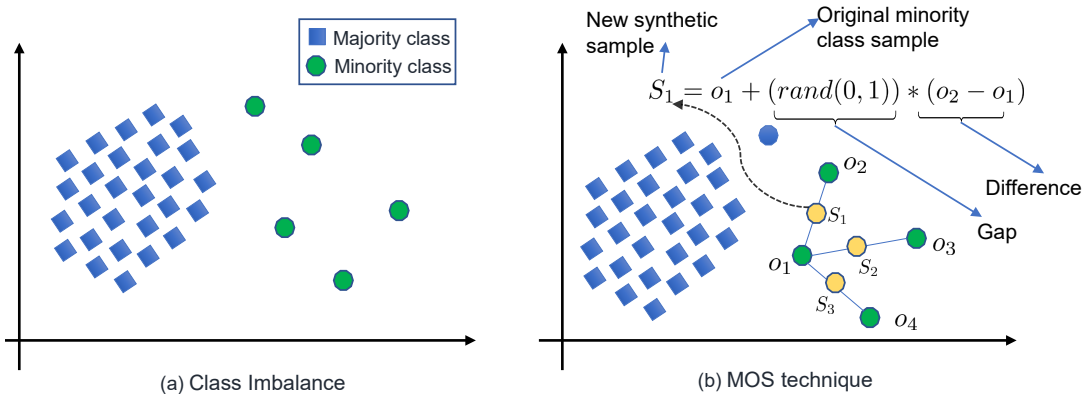


Figure 5.2: Results of MOS technique applied to handle missing and class imbalance problem. Synthetic data S_1, S_2 and S_3 is created from o_1 considering three nearest neighbors (o_2, o_3 and o_4)

scenarios. The techniques are evaluated on benchmark and real building datasets and the results are graphically illustrated by a confusion matrix and ROC curves. While evaluating the performance, the classification scores of each class are recorded.

- *Step 5:* In the final stage, an empirical comparison is performed using typical machine learning methods such as SVM, RF, kNN and NB, and results are compared using standard performance measures such as confusion matrix, ROC curves, and F1-score. A 4-fold cross-validation approach is used to evaluate the performance of the models.

5.3 SAE based feature learning for electric load profile classification

A SAE is a type of artificial neural network that works on the principle of unsupervised feature learning. Its unique feature is its capability to detect key features in a dataset using the concepts of encoding and decoding. Due to this behavior, it is widely utilized for feature extraction, dimensionality reduction, and compression [195]. This is of particular significance for building load profile classification as every building has unique load characteristics.

5.3.1 Why SAE for electric load feature learning?

Classical feature learning methods includes principal component analysis (PCA) [196], independent component analysis (ICA) [197], linear discriminant analysis (LDA) [198], and t-distributed stochastic neighbor embedding (t-SNE) [199]. Using SAE for feature learning in this chapter has four main motivations. These include, accuracy improvements [200], overfitting risk reduction using L2 regularization [201], speed up in training [202], and improved data visualization and model implementation [203]. PCA is one of the most extensively used algorithm; yet recent research shows that PCA considers correlations in data to be linearly mapped [204]. Building datasets are not always linearly solved and possess high complexity in the load profiles. Also, PCA searches for features by using principal components. The manual selection of principal components using the variance in data results in the loss of critical information [205]. A comparative analysis of feature selection algorithms showing the strengths and weaknesses of each technique is explored in [204, 206]. The SAE used in this chapter is designed to automatically discover patterns and dependencies in data by learning compact and broad representations that make it easier to extract useful information from electric load profiles of buildings.

Figure 5.3 shows the three layers of an autoencoder: an input layer, a hidden layer, and an output layer. The hidden layer shows the learned features. The two main components are an encoder and a decoder. The encoder $f(x)$ receives the input x and transforms it into hidden representations h_d. The decoder $g(h)$ receives these hidden representations from the encoder and transforms them into a reconstruction of the original input $\hat{x}$.

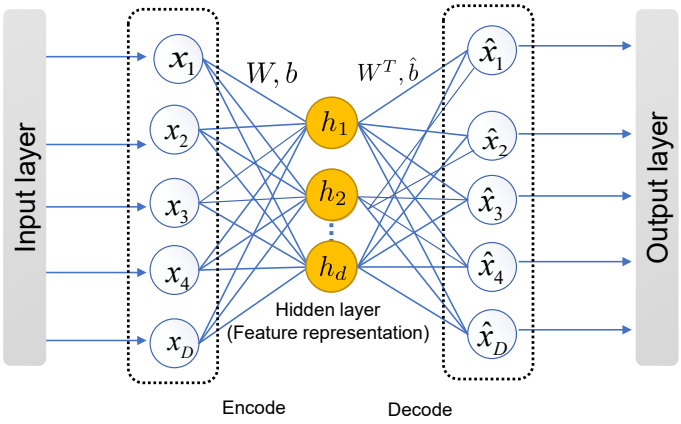


Figure 5.3: Feature learning mechanism in an autoencoder

For a dataset having input samples $x^{(1)}, x^{(2)}, x^{(3)}, \ldots, x^{(D)}$ where $x^{(i)} \in \mathbb{R}^n$, the

feature representation vector h_d and reconstruction vector $\hat{x}$ are calculated by equations (5.1) and (5.2). The weight of the encoder is W and it is equal to the transpose of the decoder weight W^T. The terms b is the bias from the input layer, and $\hat{b}$ represents the bias from the hidden layer to the output layer. The expression $\sigma(\cdot)$ indicates the activation function in the network. The reconstruction error is given with mean square error cost function J in equation (5.3). For a given set of n training samples, the overall cost function $E(W, b)$ is defined in equation (5.4). The goal is to minimize $E(W, b)$ as a function of W and b to obtain the optimal results.

$$h = \sigma(Wx + b), \tag{5.1}$$

$$\hat{x} = \sigma(W^T h + \hat{b}), \tag{5.2}$$

$$J = \frac{1}{2} \left\| \hat{x} - x \right\|^2, \tag{5.3}$$

$$E\left(W, b\right) = \left[\frac{1}{n} \sum_{i=1}^{n} \left(J\left(W, b; x^{(i)}, \hat{x}^{(i)}\right) \right) \right] + \frac{\lambda}{2} \sum_{l=1}^{n_l - 1} \sum_{i=1}^{s_l} \sum_{j=1}^{s_{l+1}} \left(W_{ji}^{(l)} \right)^2 \tag{5.4}$$

Substituting equation (5.3) in equation (5.4)

$$E\left(W, b\right) = \left[\frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{2} \left\| \hat{x}^{(i)} - x^{(i)} \right\|^2 \right) \right] + \frac{\lambda}{2} \sum_{l=1}^{n_l - 1} \sum_{i=1}^{s_l} \sum_{j=1}^{s_{l+1}} \left(W_{ji}^{(l)} \right)^2 \tag{5.5}$$

where in equation (5.5), n_l is the number of layers in the network, λ is a weight decay term, and $W_{ji}^{(l)}$ is the connection of weights between unit j and unit i of layer l. The term s_i indicates the number of units in layer l. The reconstruction error is shown by the first part of equation (5.5), and the second part is a weight decay term, commonly known as the regularization term, which aims to address the over-fitting problem.

In general, an autoencoder simply reconstructs the input data at the output layer, this technique is less effective for extracting essential features from data. A sparse constraint is typically added to the autoencoder cost function in equation (5.5) to assure improved feature representation. In this study, a sparsity constraint is placed in the autoencoder hidden units to obtain appropriate features from the input data. The sparsity parameter γ is presented in equation (5.6). The term $h_j(x^{(i)})$ denote the activation of hidden unit j on

the i-th sample of input x. The sparsity parameter is denoted with γ and has a small value close to 0 (e.g., 0.1). To keep the activation of each hidden neuron j to be close to 0.1, penalty factor KL $\left(\gamma \| \hat{\gamma}_j\right)$ is added to equation (5.4). The penalty term KL $\left(\gamma \| \hat{\gamma}_j\right)$ takes a form that penalizes $\hat{\gamma}_j$ for deviating significantly from γ, exploiting the KL divergence. The term KL $\left(\gamma \| \hat{\gamma}_j\right)$ is the Kullback-Leibler (KL) divergence [207], and it is shown in equation (5.7). After adding the sparsity parameter and KL divergence factor, the overall cost function of (5.5) is updated with equation (5.8).

$$\hat{\gamma}_j = \frac{1}{n} \sum_{i=1}^{n} \left[h_j \left(x^{(i)} \right) \right] \tag{5.6}$$

$$\mathrm{KL}\left(\gamma \| \hat{\gamma}_j\right) = \sum_{j=1}^{s_2} \gamma \log \frac{\gamma}{\hat{\gamma}_j} + (1-\gamma) \log \frac{1-\gamma}{1-\hat{\gamma}_j} \tag{5.7}$$

$$E_{\mathrm{sc}}(W, b) = E(W, b) + \beta \sum_{j=1}^{s_2} \mathrm{KL}\left(\gamma \| \hat{\gamma}_j\right) \tag{5.8}$$

where, β in equation (5.8) is the sparsity term, and s_2 denotes the hidden units. In essence, SAE improves the performance of classical autoencoder and identifies sparse feature representation. Instead of simply converting the inputs, it extracts more information from them by adding a sparsity constraint that forces the model to respond to the unique statistical features of the data. The penalty induces the model to activate (i.e., produce an output value close to 1) in specific parts of the network based on the input data, while deactivating all other neurons (i.e., to have an output value close to 0). The steps of the feature learning process are summarized in Algorithm 5.2.

5.4 Multi-stage transfer learning approach

A layer-wise multi-stage transfer learning approach is implemented to improve classification accuracy. The method is named as a multi-stage or two-phase protocol because it combines the pretraining (unsupervised phase) and a supervised learning phase. A SAE model is initially trained in an unsupervised fashion while saving the key data features as described in Section 5.3. The knowledge (features) learned by an SAE is then re-trained with DNN in a supervised manner to fine-tune the results.

In the first step, a SAE is trained on the raw inputs x^k to obtain primary features in

Algorithm 5.2 Feature learning for building load profiles based on SAE

Input: Data samples $x^{(i)} = x^{(1)}, x^{(2)}, \ldots, x^{(D)}$ where $x^{(i)} \in \mathbb{R}^n$
for Data samples $x^{(i)}, i = 1, 2, \ldots n$ **do**
    Transform the $x^{(i)} \in \mathbb{R}^n$ into $h_d = \mathcal{F}$ ▷ $\mathcal{F}$:feature space
    Encoder stage: $h = (Wx + b)$
    Add element-wise activation function $\sigma(x) = 1/(1 + e^{-x})$
    Initialize W and b in network ▷ W:weight, b: bias
    Map the h_d to the reconstruction vector $\hat{x}$
    [t] Decoder stage: $\hat{x} = \sigma(W^T h + \hat{b})$
    Return the reconstruction vector $\hat{x}$
    Compute reconstruction error using eq (3)
    Apply the weight decay term λ
    Minimize the error using eq (4)
end for
for Improve feature learning **do**
    Add sparsity term γ
    Add penalty factor KL $\left(\gamma \| \hat{\gamma}_j\right)$ to eq (4)
    Penalizes $\hat{\gamma}_j$ for deviating significantly from γ
    Update the eq (4) with γ and KL $\left(\gamma \| \hat{\gamma}_j\right)$
    Minimize eq (7) with optimizer, i.e., scaled conjugate gradient
end for

hidden layer $h_k^{(1)}$ as shown in first stage of Figure 5.4. In the second step, the features (data representations) learned by SAE is transferred as input to the second SAE to extract secondary features $h_k^{(2)}$ as shown in second stage of Figure 5.4. The key difference between the two SAEs is that the features extracted from the first autoencoder are used as the training input to the second autoencoder. This layer-wise representation is also known as 'pre-training in deep learning [208]. In the third step, a softmax regression classifier is added on top of the features learned in the pretraining phase for classification task. In other words, it has been added to the main network to classify the various levels of load profiles of buildings. This stage involves supervised fine-tuning of the entire network learned in the pretraining. A softmax layer and all the layers of SAE are stacked to form a deep network and the whole network is retrained in a supervised manner to fine tune the classification results. The steps of the approach are illustrated in Figure 5.4. Using this approach has two benefits. First, it reduces the generalization error. Second, it facilitates the development of deeper networks by stacking multiple SAEs, resulting in improved classification performance.

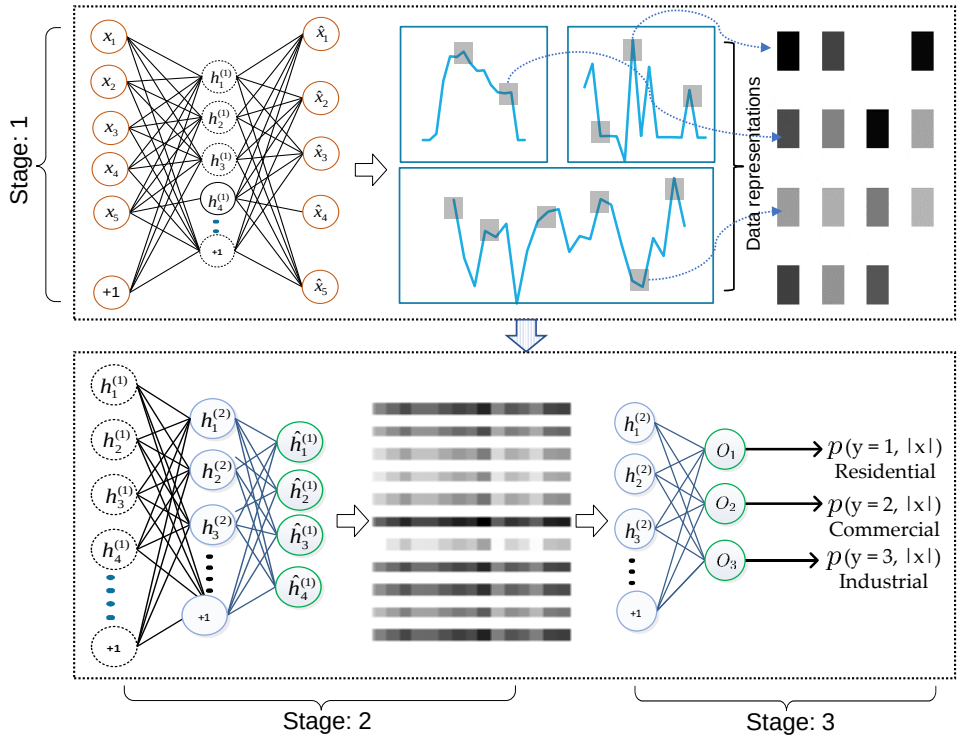


Figure 5.4: An illustration of multi-stage transfer learning concept

As this study deals with multi-class problem (residential, commercial, industrial), a softmax regression classifier is added to the main network. It means that the output class labels are multi-class classification instead of binary classification. For a given input of x, softmax layer estimates the probability of training sample $x^{(i)}$ belongs to class c given the weight W and net input $z^{(i)}$. The probabilities of each class $c = 1, \ldots, k$ are determined as follows.

$$p\left(O = c \mid x^{(i)}; Wc\right) = \frac{e^{z^{(i)}}}{\sum_{c=1}^{k} e^{z(i)}} \tag{5.9}$$

where O is the output class that corresponds to the input vector $x(i)$ and Wc is the weight parameter for each class c, where $c = 1, 2, 3 \ldots, k$. The maximum probability of each class of building is calculated as follows

$$\text{Class } \left(x^{(i)}\right) = \max p\left(O = c \mid x^{(i)}; Wc\right) \tag{5.10}$$

where $x(i)$ denotes the class with the maximum probability. As softmax layer returns probabilities in the range [0-1], the target values are binarized using one-hot encoding technique [209] as presented in Table 5.1.

Table 5.1: Binarized target values for multi-class using one-hot encoding

Class	Labels	Binary form
1	Residential class	100
2	Commercial class	010
3	Industrial class	001

After obtaining the softmax probabilities and class labels, the best accuracy on each class is determined using the majority vote ensemble learning technique. It totals the votes for all of the predicted labels from the ensemble learners, and generates a final prediction based on the label with the most votes. Alternatively, it averages the labels from basic learners and selects the label with the highest value. For majority voting, three ensemble learners including Bagging [210], AdaBoost [211] and RUSBoost [212] are used in this study. The steps of MSTL process are summarized in Algorithm 5.3.

Algorithm 5.3 Multi-stage transfer learning approach (MSTL)

procedure
  for multi-stage learning strategy **do**
    First stage, get $f \leftarrow h_k^{(1)}$
    Second stage: forward $h_k^{(1)}$ to $h_k^{(2)}$
    Add softmax regression classifier S_c on top of $h_k^{(2)}$
    Stack layers L of network with S_c
    Re-train in supervised manner $f \leftarrow (f, \boldsymbol{X}, \boldsymbol{Y})$
    Fine-tune the results ▷ lower generalization error
    Binarize the output values ▷ one-hot encoding
    Estimate $p\left(O = c \mid x_i; Wc\right) = e^{z^{(i)}} / \sum_{c=1}^{k} e^{z(i)}$
    Find Class $(x_i) = \max p\left(O = c \mid x_i; Wc\right)$
    Decide predicted class based on highest probability
    Create ensemble classification for each class label
    Decide best accuracy based on majority voting
    Make final prediction using label with most votes
  end for
end procedure

5.5 Results and experimental verification

Figure 5.5 shows the model validation procedure. A k-fold cross-validation method is used to assess the classification performance. While testing a model, k smaller sets

of data are used to train the model and the remaining $(k-1)$ sets of data are used to evaluate it. This method is repeated for all k distinct sections of the data, resulting in each component being used once as validation data and $(k-1)$ times as training data. The average classification accuracy of k experiments and error are computed to create a single performance metric.

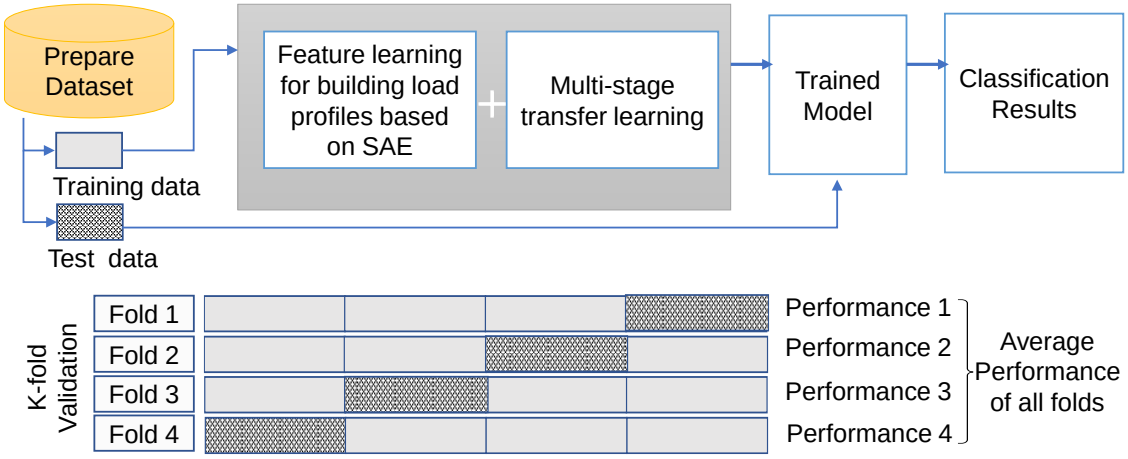


Figure 5.5: Evaluation of model performance and k-fold validation

5.5.1 Performance criteria

Standard performance metrics are used to evaluate the model performance, as explained in the following sections.

1) Confusion Matrix

A confusion matrix is used to visualize the classification performance of each class. It is a simple yet efficient way of evaluating classification performance and visualising actual versus expected class accuracies [213]. The predicted class is represented by each column in the confusion matrix, whereas the actual class is shown by each row. It shows the total number of true negatives (TN), false negatives (FN), true positives (TP), and false positives (FP). The TP represents the labels that belong to the class and is accurately predicted. The FP denotes labels that do not belong to the class are predicted as positive by the classifier. Similarly, TN indicates that the labels do not belong to the class, and it is rightly predicted. The FN indicates that the labels belong to the class are predicted as negative. Figure 5.6 shows the layout of a confusion matrix with three classes.

Confusion Matrix		Predicted			False Negative (FN)
		Class 1	Class 2	Class 3	
Actual	Class 1	A	B	C	B+C
	Class 2	D	E	F	D+F
	Class 3	G	H	I	G+H
False Positive (FP)		D+G	B+H	C+F	Overall Accuracy = A+E+I / (Sum of orange and green squares)

Figure 5.6: Structure of confusion matrix having three classes

2) Recall, precision, specificity, and f1-score

Based on the information of TN, TP, FN and FN in the confusion matrix, other metrics including accuracy, recall, precision, specificity, and F1-score are calculated. The ratio of correct predictions TP and TN to the total number of occurrences is defined as accuracy. The recall is a model ability to correctly classify the TP. For example, recall indicates how many residential profiles were correctly identified among all the load profiles that truly have the residential class. Precision is defined as the ratio of the TP to all positives. The F1-score is a combination of accuracy and recall. Table 5.2 shows the details for each metric.

Table 5.2: Performance metrics to assess the classification model

Performance Metric	Formula	Calculations
Accuracy	TP + TN / $\sum(TP + TN + FP + FN)$	(A + E+I) / (A + E+I + D+G+B+C)
Precision	TP / $\sum(TP + FP)$	(A / A + D+G)
Recall	TP / $\sum(TP + FN)$	(A / A + B+C)
Specificity	TN / $\sum(FP + TN)$	(E+I / D+G + E+I)
F-score	2 TP / 2 ($\sum TP + FP + FN$)	(2 * A) / (2*A + D+G + B+C)
TP: True positives; TN: True negatives; FP: False positives; FN: False negatives.		

3) Performance analysis using ROC curves

This technique is used to determine how well a method performs when it comes to classification accuracy, particularly sensitivity and specificity. In comparison to precision, recall, and F1-score, ROC curves mainly show the relationship between the false positive rate (FPR) and the true positive rate (TPR).

5.5.2 Scenario:1 Validation on the benchmark dataset

An open-access benchmark dataset of buildings is used [191] for validation. The dataset is available for a broad range of analysis, and it contains hourly load profiles of 24 buildings from various end-use sectors, including industrial, commercial, and residential consumers. For performance analysis, the entire year data of 2020 is used. The data set contains a total of 8,760 observations. Using 4-fold cross-validation, the entire dataset is divided into training and testing datasets. The training dataset is used for feature engineering and classification model building, while the testing data set is used to calculate and report classification performance.

The classification performance on 4-fold cross-validation is shown in Table 5.3. The model is validated on various parts (folds) of the dataset and the accuracy of each fold is provided, along with its error rate. The results show that the classification accuracy is above 90 percent for each subset of the dataset. The highest accuracy is recorded in iteration 1 ($k = 1$), with a classification rate of 93.28 percent. The maximum number of misclassifications are observed in iteration 3 ($k = 3$) with a misclassification rate of 7.28 percent. The results of four iterations (folds) are averaged to yield a single classification metric. The average classification accuracy is 92.97 percent without overfitting the data.

Figure 5.7 shows an illustration of the classification of electric load profiles. The input of the model is the electric load profiles of buildings, and the output is the classification of load profiles. The green mark shows that the load profiles are classified correctly, whereas the red cross indicates the misclassification. For example, if the actual load profile is industrial and model classified it is as residential, then it is counted as misclassification. For the sake of clarity, 15 load profiles are shown. Out of the 15 profiles, 11 are correctly classified and 3 profiles are misclassified by classification model.

To visualize the classification accuracy of each class, a confusion matrix is used, and results are shown in Figure 5.8(a)(b). The matrix compares the actual target values (rows)

Table 5.3: Classification accuracy on test dataset

K-fold	Classification Accuracy (%)	Error rate (%)
K = 1	93.28	6.72
K = 2	92.77	7.23
K = 3	92.72	7.28
K = 4	93.10	6.90
	Average: 92.97	7.02

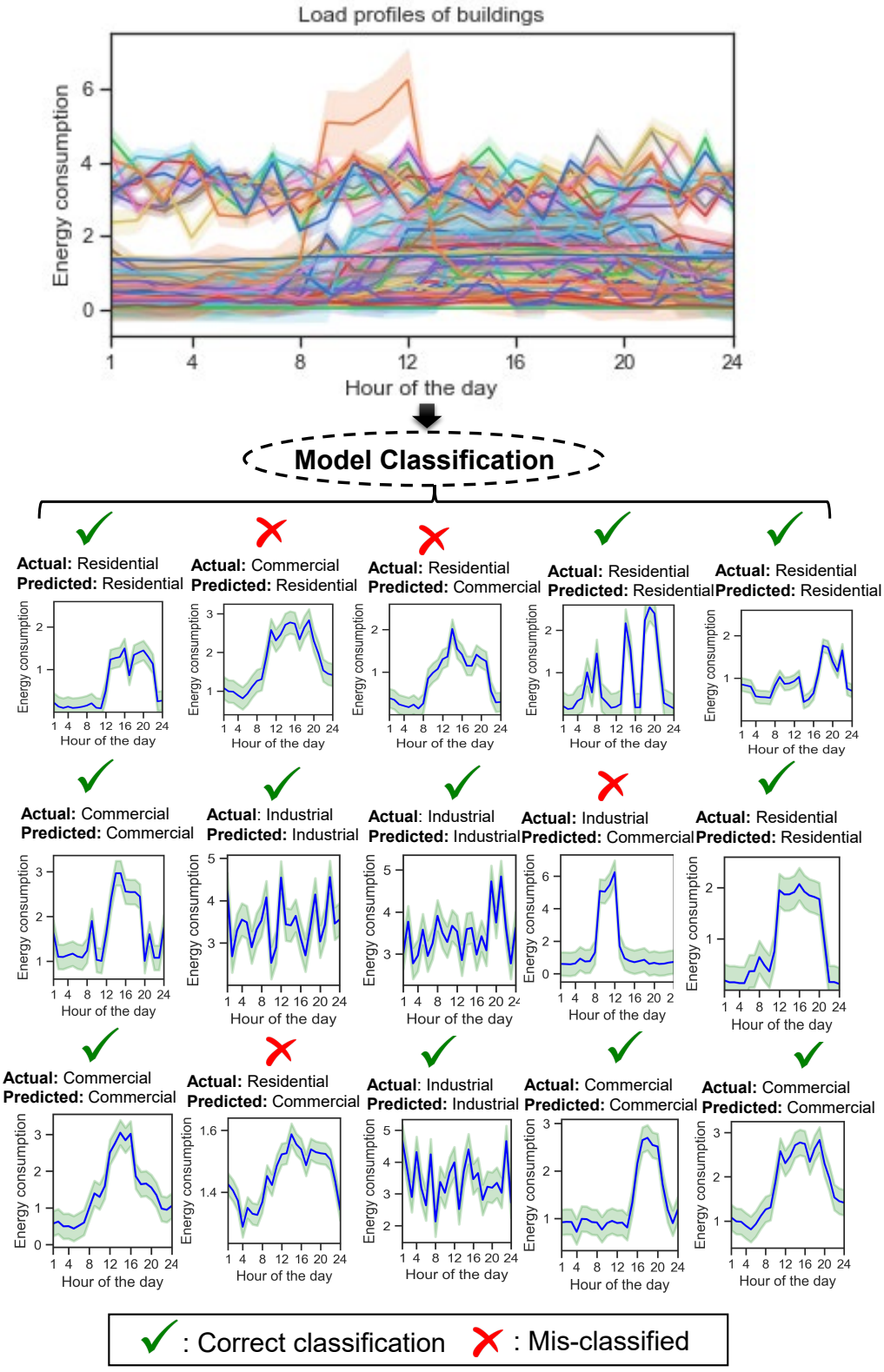


Figure 5.7: An illustration of classification of load profiles of buildings.

with those predicted (columns) by the classification model. It gives a holistic view of how well the classification model is performing and what kinds of errors it is making. Figure

5.8(a) shows the classification accuracy of each class on training data, whereas Figure 5.8(b) displays the classification results for each class on test dataset. The diagonal cells (blue) correspond to observations that are correctly classified. The off-diagonal cells (orange) correspond to incorrectly classified observations. Figure 5.8(b) indicates that all residential observations (2190) are correctly classified. However, there are 260 and 193 misclassifications in the commercial and industrial classes, respectively. A row summary (right side) displays the percentages of correctly and incorrectly classified observations for each true class. A column summary (bottom) displays the percentages of correctly and incorrectly classified observations for each predicted class. The overall accuracy is shown at the top of the confusion matrix. Figure 5.8(b) shows that the proposed approach successfully classified the load profiles of buildings, i.e., residential, commercial and industrial classes with an overall accuracy of 93.11 percent on the test dataset.

Figure 5.8(c)(d) shows the ROC curve for both the training and test sets to illustrate the classification between the three classes. The FPR on the x-axis shows the positive cases that were wrongly classified as negative during the classification process, whereas the TPR on the y-axis describes the positive cases that were correctly classified during the test. Classification models with ROC curves that touch the top left corner of the curve indicates good classification performance. The ideal point on the curve would be (0,1) in the upper left corner, where all positive instances are correctly classified and no negative cases are incorrectly classified. The training and test curves in 5.8(c)(d) are near to the top left corner (0,1), indicating a higher TPR. As a result, it indicates a good classification on both the training and test sets.

Classification accuracy alone is usually not enough to determine whether the developed model is sufficient to provide accurate classifications. The issue with accuracy is that it cannot distinguish between various types of misclassifications and is dependent on the distribution of classes in the dataset [214, 215]. Suppose a dataset contains 1000 samples, 995 of which are from the residential class and five from the commercial class. If the model correctly classifies all of them as residential, the accuracy is 99.5 percent, even though the classifier missed all commercial samples. In such situations, classification accuracy as a measure of model quality is not an adequate measure. This chapter constructs various performance metrics, including precision, recall, specificity, and F1-score. The results are shown in Table 5.4.

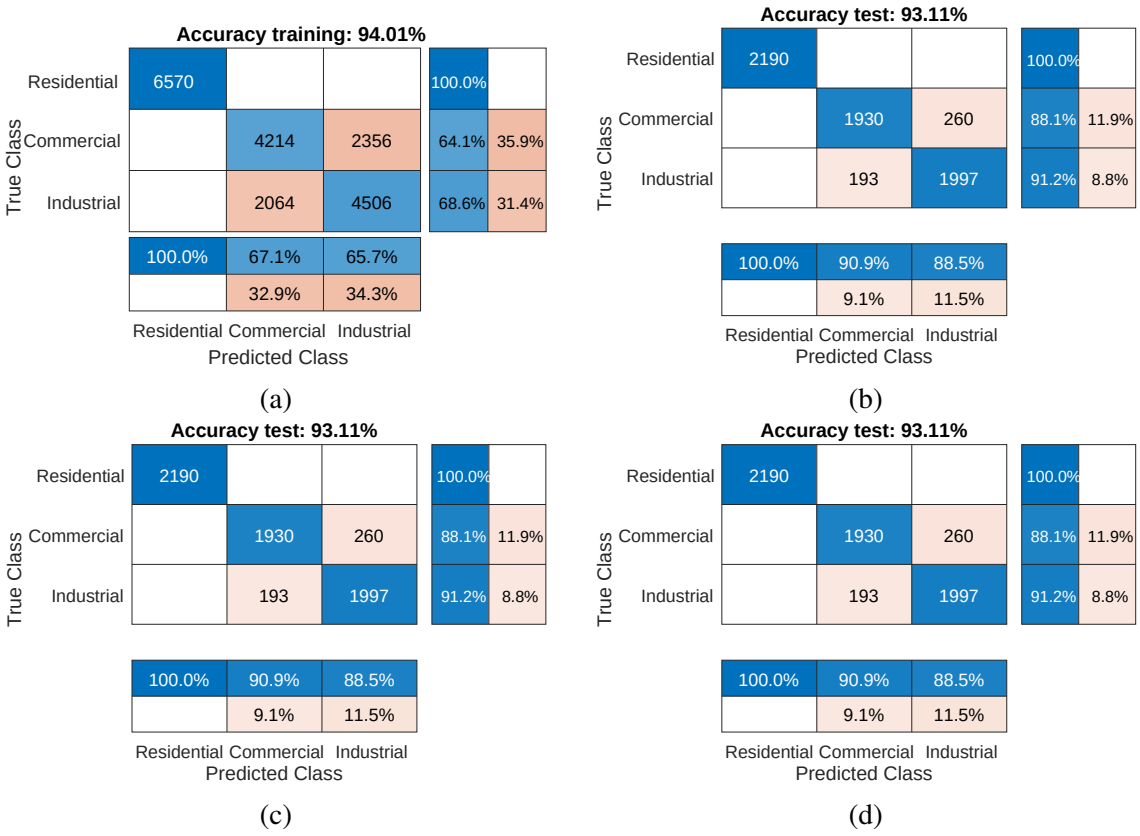


Figure 5.8: Classification results on benchmark dataset (a) Confusion matrix results on training data (b) results on test data (c) ROC curve on training data (d) ROC curve on the test dataset

Table 5.4: Classification performance on commonly accepted performance metrics

Performance Metric	**Class 1 (Residential)**	**Class 2 (Commercial)**	**Class 3 (Industrial)**	**Average**
Precision (%)	100	90.82	88.50	93.11
Recall (%)	100	88.17	91.09	93.09
Specificity (%)	100	95.55	94.08	96.54
F1-score (%)	100	89.48	89.78	93.08

5.5.3 Scenario:2 Validation on real buildings dataset

After first validation with benchmark data of buildings and discussion of obtained results, the results are expanded on real building datasets collected from local distribution network operator. The goal is to evaluate classification performance on diverse datasets and under different regional prevailing conditions. A five-minute time-step is used to evaluate the classification performance. Figure 5.9 shows the load profiles of buildings at five-minute time-step with 288 values for each building for a single day. The graphic

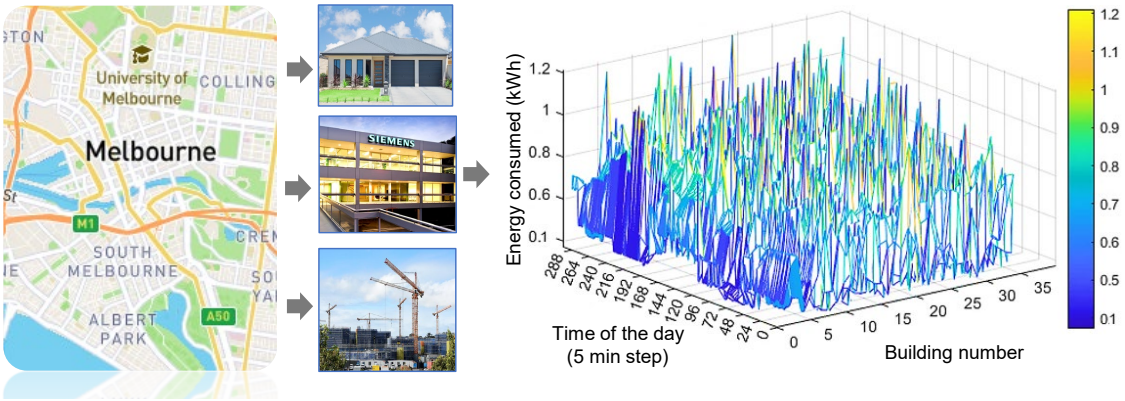


Figure 5.9: Load profiles of buildings showing the demand at different times of the day.

shows that each building has diverse load profiles, and this is probably due to different energy consumption behaviors of occupants, and activity schedules in individual buildings.

The accuracy of classification is calculated using 4-fold cross-validation, and results are shown in Table 5.5. The results are validated on four different folds (sets) of the dataset and accuracy is noyed for the individual fold. The tabular results show that the model achieved classification accuracy of above 84 percent on each fold. The maximum accuracy is noticed in iteration 4 ($k = 4$) with a classification rate of 85.90 percent. The highest misclassifications are observed in iteration 2 ($k = 2$) with a misclassification rate of 15.81 percent. The four iterations (folds) are averaged to yield a single classification metric. The average classification accuracy is 85.26 percent without overfitting the data.

Figure 5.10 shows the performance in terms of the confusion matrix and ROC curves. The accuracy achieved on the training dataset is shown in Figure 5.10(a), with an overall accuracy of 85.95 percent. Figure 5.10(b) shows the model performance on the test dataset, with an overall accuracy of 85.90 percent. For residential class, 411 data samples are misclassified. In the case of commercial and industrial classes, 260 and 193 data samples are misclassified by model. As accuracy metric only is not enough to evaluate the classification performance [216], ROC curves are used to test the model performance. Figure 5.10 (c)(d) shows a plot of three ROC curves, each representing one of the three classes. ROC curves are shown on the training and test dataset. The curve closer to the top left corner indicates a high level of accuracy. Similarly, the higher AUC values indicate better performance. For instance, the ROC curve and AUC value of the industrial class is higher in both training and test dataset, indicating better accuracy from the other two classes. The performance of the model on each class is shown in Table 5.6.

Table 5.5: Classification accuracy on test dataset of buildings

K-fold	Classification Accuracy (%)	Error rate (%)
K = 1	84.78	15.22
K = 2	84.19	15.81
K = 3	85.16	14.84
K = 4	85.90	14.10
Average	85.26	14.74

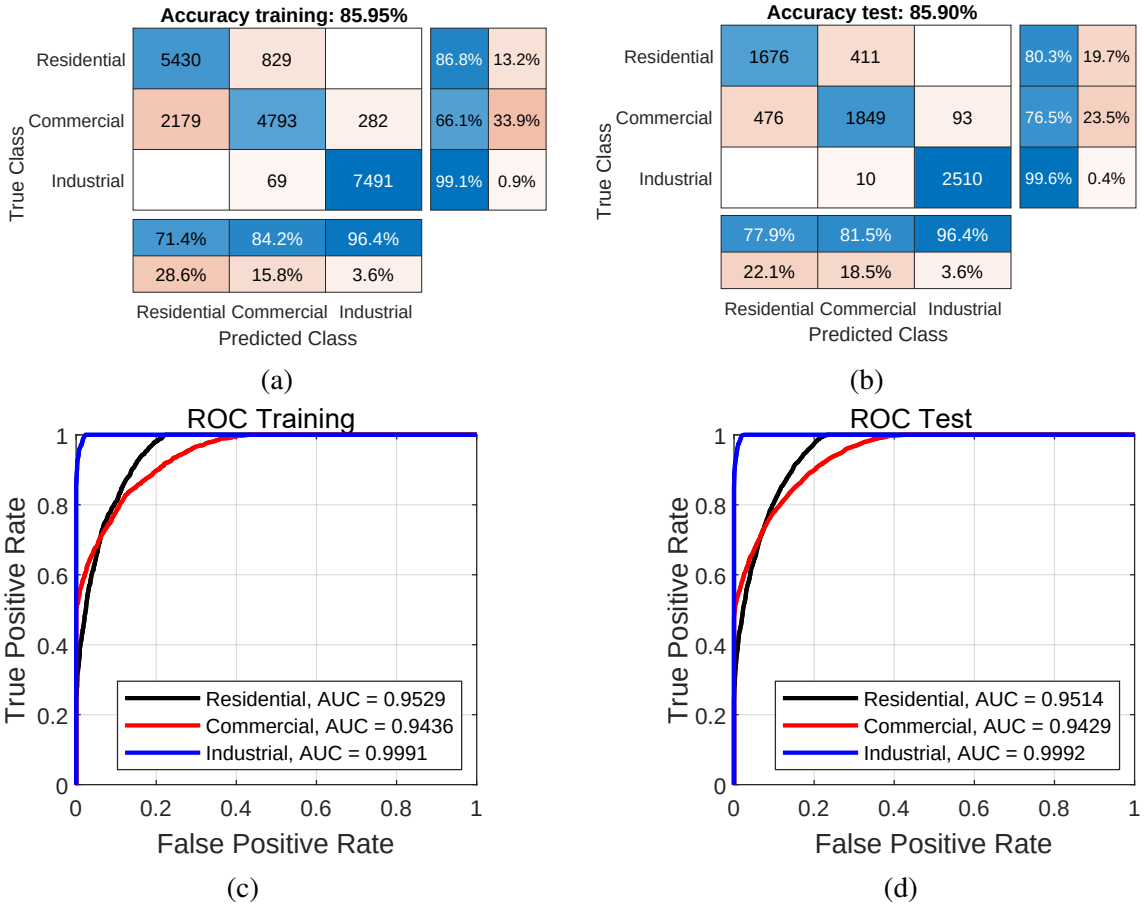


Figure 5.10: Classification results on real buildings dataset (a) Confusion matrix results on training data (b) Results on test data (c) ROC on training data (d) ROC on test set

Table 5.6: Performance of proposed method on classification metrics.

Performance Metric	**Class 1 (Residential)**	**Class 2 (Commercial)**	**Class 3 (Industrial)**	**Average**
Precision (%)	77.88	82.54	96.42	85.61
Recall (%)	80.80	76.66	99.60	85.68
Specificity (%)	90.36	90.86	97.93	93.05
F1-score (%)	79.07	78.88	97.98	85.31

5.6 Comparison with the state of the art

To compare the performance of the proposed method, four state-of-the-art methods including SVM, RF, kNN and NB are considered. For a fair comparison, same datasets are used for all methods and results are reported with standard performance metrics such as precision, recall (sensitivity), specificity, F1-score, and ROC-AUC values. As parameters greatly impact the performance of the models, the best combination of hyperparameters are selected by adopting the random search technique [217].

5.6.1 Experiment 1: Comparison on benchmark dataset

In the first phase, optimal parameters of each model are obtained by applying a random search technique [217]. A hyper-parameter optimization based on random search is performed across all models and the best combination of parameters are selected based on their minimum error. The optimal parameters for the models are shown in Table 5.7 and the graphical demonstrations are shown in Figure 5.11. After finding the optimal parameters, the comparisons are drawn on widely accepted classification metrics and results are

Table 5.7: Hyperparameter tuning to select the optimal parameters for each model

Model	Hyperparameter search range	Optimal hyperparameters
Proposed	Ensemble method: Bag, AdaBoost, RUSBoost Number of learners: 10-500 Maximum number of splits: 1-21072 Number of predictors to sample: 1-2	Ensemble method: Bag Number of learners: 252 Maximum number of splits: 19 Number of predictors to sample: 2
SVM	Multiclass method: One-vs-All, One-vs-One Box constraint level: 0.001-1000 Kernel scale: 0.001-1000 Kernel function: Gaussian, Linear, Quadratic, Cubic Standardize data: true, false Training time limit: false, true	Multiclass method: One-vs-All Box constraint level: 0.94414 Kernel scale: 0.0014509 Kernel function: Gaussian Standardize data: true Training time limit: false
RF	Maximum number of splits: 1-21072, Split criterion: Gini's diversity index, Twoing rule, Maximum deviance reduction, Training time limit: false, true	Maximum number of splits: 45, Split criterion: Maximum deviance reduction, Training time limit: false,
kNN	Number of neighbors: 1-10537 Distance metric: City block, Chebyshev, Correlation, Cosine, Euclidean, Hamming, Jaccard, Mahalanobis, Minkowski (cubic), Spearman Distance weight: Equal, Inverse, Squared inverse Standardize data: true, false	Number of neighbors: 25 Distance metric: Mahalanobis Distance weight: Squared inverse Standardize data: true Training time limit: false
NB	Distribution names: Gaussian, Kernel Kernel type: Gaussian, Box, Epanechnikov, Triangle Width: 0 - 0.99349	Distribution names: Kernel Kernel type: box Width: 0.0091471

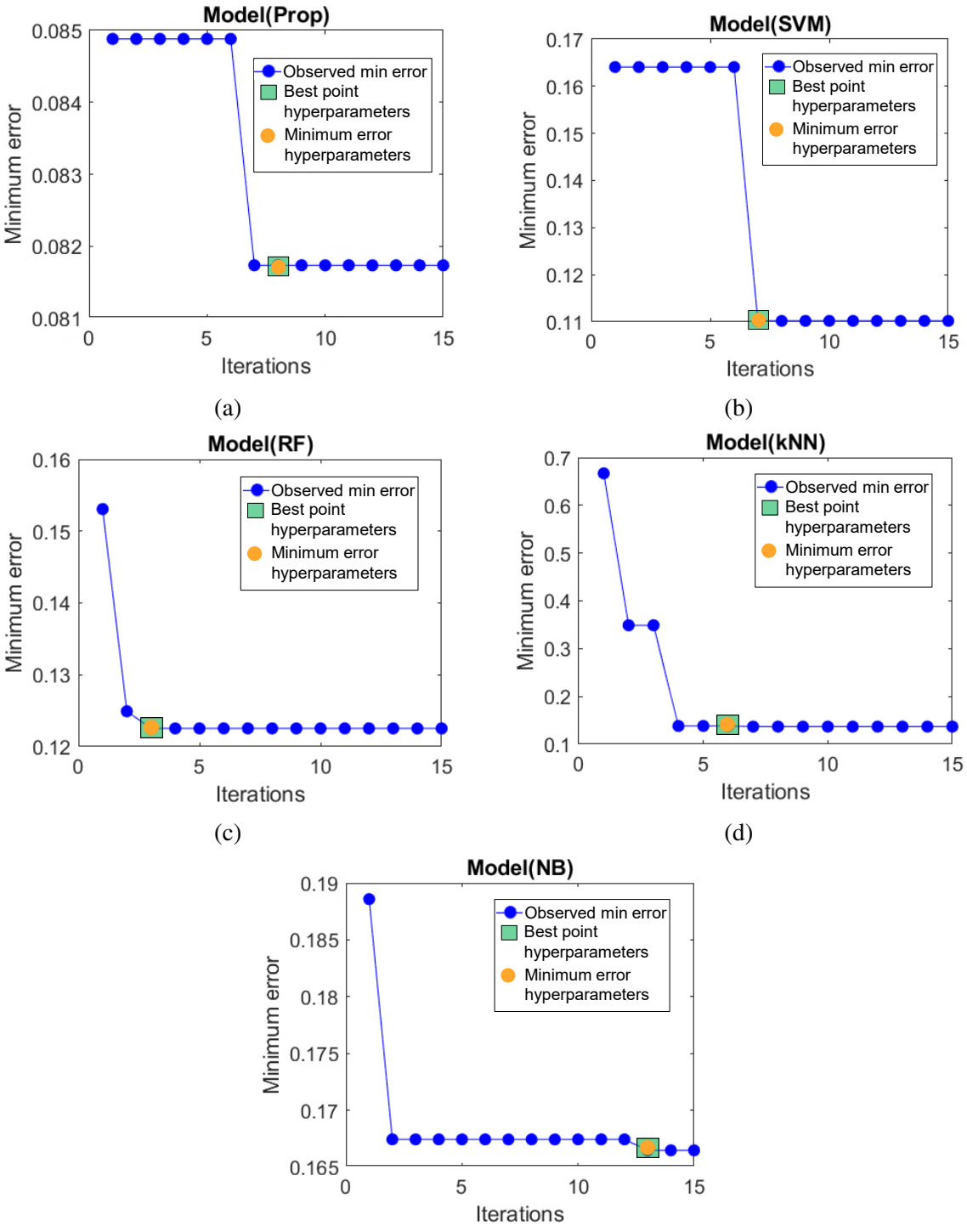


Figure 5.11: Optimal parameters selection for models evaluation. (a) Minimum error hyperparameters for proposed method (b) for SVM (c) for RF (d) for kNN, and (e) for NB model

shown Table in 5.8. This table presents the classification performance of proposed methods on the other four methods. For each class of the dataset, classification performance

is compared in terms of sensitivity (recall), specificity, precision, F1-score and AUC. The results show that proposed method has on average the best classification accuracy from all other methods. The proposed method has an average sensitivity of 93 percent, compared to 89 percent for the SVM, 87 percent for the RF, 86 percent for the kNN, and 83 percent for the NB model. In terms of specificity, precision, and F1-score, the proposed method produces the best classification results, ranging from 93 to 97 percent. The other approaches achieved classification performance ranging from 83 to 95 percent.

Figure 5.12 demonstrates the performance of different models on ROC-AUC curves. The results show that the proposed model obtains good results for ROC-AUC, i.e., 1.0 for class 1, and 0.98 for both class 2 and class 3. SVM achieves a ROC-AUC of 1.0 for class 1, and 0.95 for other two classes respectively. RF also perform well with ROC-AUC values of 1.0 on class 1, and 0.94 for the other two classes. Similarly, kNN and NB achieve good performance with ROC-AUC score of 1 for class 1, and 0.94 and 0.90 for other two classes. These findings demonstrate that the proposed solution efficiently classifies the load profiles samples from the dataset. For the sake of clarity, the average of AUC values are presented in Table 5.9. SVM obtained a reasonable ROC-AUC scores with an average of 0.97 percent, 0.96 percent for RF and KNN and 0.94 percent for NB. The proposed scheme achieves the highest ROC-AUC of 0.99, proving its superiority among all other conventional techniques.

5.6.2 Experiment 2: Comparison on real buildings dataset

In this section, extensive simulations are conducted using real-world building measurements to demonstrate the efficacy of the proposed approach in comparison to existing methods. Table 5.10 provides the summarized performance comparison of results. The results reveal that the proposed scheme outperformed existing methods in terms of sensitivity (recall), specificity, precision, and F1-score. Over the given dataset, the proposed model achieves average values of 0.86, 0.93, 0.86, and 0.85 for sensitivity (recall), specificity, precision, and F1-score. The SVM, on the other hand, yields average values of 0.84, 0.92, 0.85, and 0.83 for sensitivity (recall), specificity, precision, and F1-score. Looking at the RF and kNN, the average results are quite similar (0.83 for RF vs 0.82 for kNN, 0.92 for RF vs 0.92 for kNN, 0.85 for RF vs 0.82 for kNN and 0.82 for RF vs 0.82 for kNN). It is worth noting that the average results for RF and kNN are quite identical; nevertheless, classification performance varies between individual classes in the dataset. NB

Table 5.8: Comparison of performance with benchmark classification algorithms.

Case study:1 (Validation on benchmark datasets)

Method	Class	Sensitivity	Specificity	Precision	F1-score	AUC
Proposed	Residential	1.00	1.00	1.00	1.00	1.00
	Commercial	0.88	0.96	0.91	0.89	0.98
	Industrial	0.91	0.94	0.88	0.90	0.98
	Average:	**0.93**	**0.97**	**0.93**	**0.93**	**0.99**
SVM	Residential	1.00	1.00	1.00	1.00	1.00
	Commercial	0.76	0.96	0.90	0.83	0.95
	Industrial	0.92	0.88	0.80	0.85	0.96
	Average:	**0.89**	**0.95**	**0.90**	**0.89**	**0.97**
RF	Residential	1.00	1.00	1.00	1.00	1.00
	Commercial	0.77	0.92	0.82	0.80	0.95
	Industrial	0.84	0.89	0.79	0.81	0.95
	Average:	**0.87**	**0.94**	**0.87**	**0.87**	**0.96**
kNN	Residential	1.00	1.00	1.00	1.00	1.00
	Commercial	0.78	0.90	0.80	0.79	0.94
	Industrial	0.80	0.89	0.79	0.79	0.94
	Average:	**0.86**	**0.93**	**0.86**	**0.86**	**0.96**
NB	Residential	1.00	1.00	1.00	1.00	1.00
	Commercial	0.72	0.89	0.76	0.74	0.90
	Industrial	0.77	0.86	0.73	0.75	0.90
	Average:	**0.83**	**0.91**	**0.83**	**0.83**	**0.94**

Table 5.9: Comparison of performance on average values of ROC-AUC

Performance Metrics	Classes	Methods				
		SVM	RF	kNN	NB	Proposed
AUC (Case study:1)	Residential	1.00	1.00	1.00	1.00	1.00
	Commercial	0.95	0.95	0.94	0.90	0.98
	Industrial	0.96	0.95	0.949	0.90	0.98
	Average:	**0.97**	**0.96**	**0.96**	**0.94**	**0.99**

has the lowest values with an average sensitivity of 0.81, specificity of 0.91, precision of 0.83, and F1-score of 0.93.

Figure 5.13 displays the performance comparison in terms of ROC-AUC curves. AUC describes how much the curve is stretched towards the upper left corner from the diagonal. It exhibits how the number of correctly predicted positive cases varies with the number of incorrectly predicted negative cases. The AUC numbers clearly show that the proposed strategy produces superior outcomes by attaining the highest AUC values for each class (class 1, AUC = 0.92, class 2, AUC= 0.94, class 3, AUC= 0.99). SVM obtain the second highest AUC values with AUC=0.93 for class 1, 0.92 for class 2 and 0.99 for class 3. The AUC values for the other models, which included RF, kNN, and

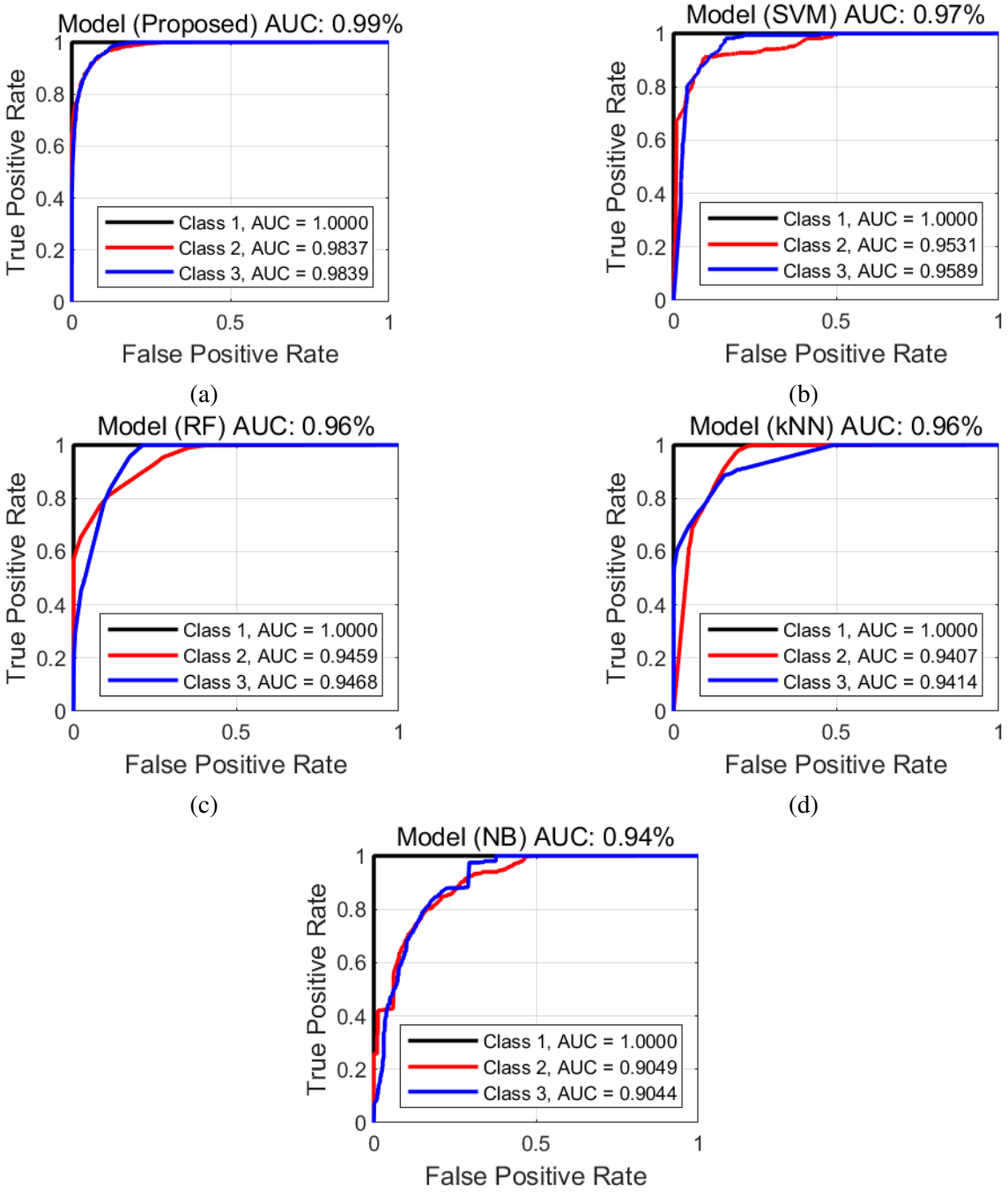


Figure 5.12: ROC-AUC curves for models evaluation. (a) Proposed (b) SVM (c) RF (d) kNN (e) NB

NB, ranged from 0.81 to 0.98 depending on the class type in the dataset. AUC values are given alongside three curves (blue, black, and red) for each class type to demonstrate classification performance. The ROC curve closest to the top left corner or the point (0, 1) on the plane shows the most accurate classification. For example, the proposed

Table 5.10: Comparison of performance with popular classification algorithms

Case study:2 (Validation on actual data of buildings)

Method	Class	Sensitivity	Specificity	Precision	F1-score	AUC
Proposed	Residential	0.81	0.90	0.78	0.79	0.95
	Commercial	0.76	0.91	0.82	0.79	0.94
	Industrial	1.00	0.98	0.96	0.98	1.00
	Average:	**0.86**	**0.93**	**0.86**	**0.85**	**0.97**
SVM	Residential	0.90	0.84	0.70	0.79	0.94
	Commercial	0.62	0.96	0.88	0.73	0.93
	Industrial	1.00	0.98	0.96	0.98	0.99
	Average:	**0.84**	**0.92**	**0.85**	**0.83**	**0.95**
RF	Residential	0.94	0.81	0.69	0.79	0.90
	Commercial	0.56	0.91	0.91	0.70	0.86
	Industrial	1.00	0.95	0.95	0.97	0.99
	Average:	**0.83**	**0.92**	**0.85**	**0.82**	**0.91**
kNN	Residential	0.72	0.89	0.74	0.73	0.89
	Commercial	0.77	0.86	0.75	0.76	0.81
	Industrial	0.98	0.99	0.99	0.98	1.00
	Average:	**0.82**	**0.92**	**0.82**	**0.82**	**0.90**
NB	Residential	0.96	0.80	0.66	0.78	0.91
	Commercial	0.46	0.98	0.93	0.62	0.89
	Industrial	0.93	0.93	0.90	0.94	1.00
	Average:	**0.81**	**0.90**	**0.83**	**0.78**	**0.93**

Table 5.11: A comparison of performance based on average ROC-AUC values

Performance Metrics	Classes	Methods				
		SVM	RF	kNN	NB	Proposed
AUC (Case study:2)	Residential	0.94	0.90	0.89	0.91	0.95
	Commercial	0.93	0.86	0.81	0.89	0.94
	Industrial	0.99	0.99	0.99	1.00	1.00
	Average:	**0.95**	**0.91**	**0.90**	**0.93**	**0.97**

approach obtains the best ROC curves among the four methods. For the sake of clarity, the mean values of AUC for each model are shown in Table 5.11. The findings show that the proposed strategy obtained 0.97 AUC on average, 0.95 for SVM, 0.91 for RF, 0.90 for kNN, and 0.93 for NB.

5.6.3 Analysis of percentage improvement in accuracy

The percentage improvement in performance obtained by the proposed method over traditional schemes is computed using equation (5.11), and the results are shown in Figure 5.14. The term A_P represents the accuracy obtained by the proposed method on

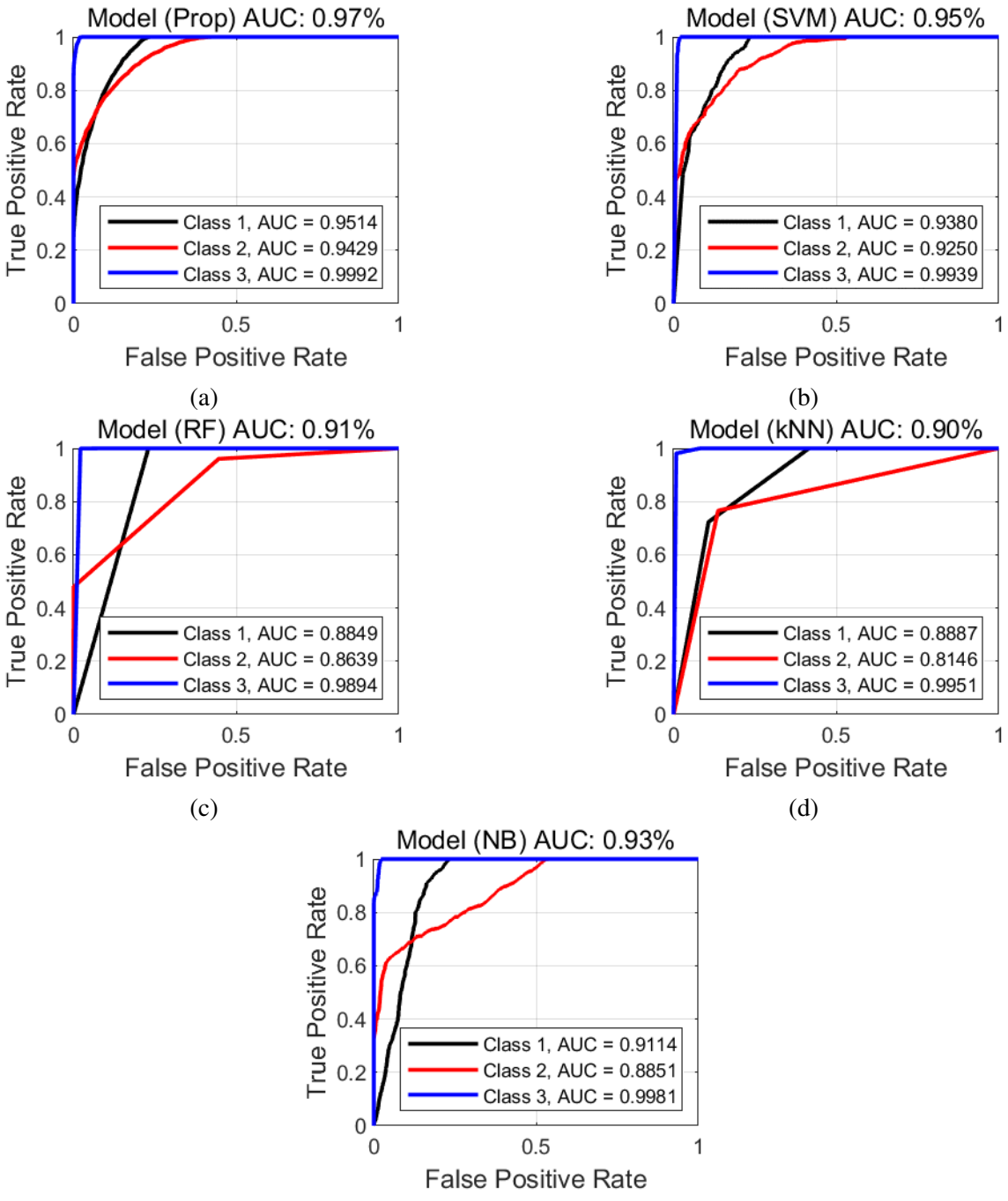
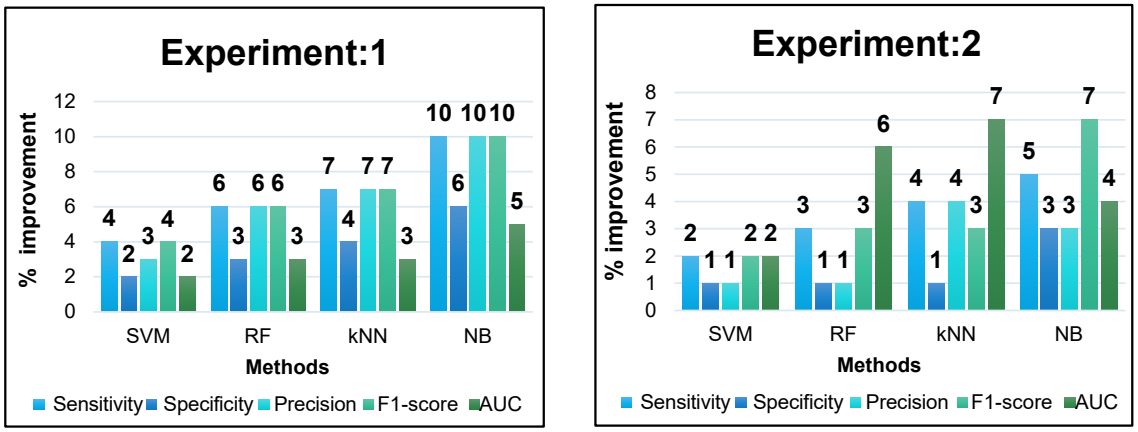


Figure 5.13: Confusion matrix and ROC curves for model evaluation. (a) Classification performance on training data (b) Results on test data (c) ROC curve on training data (d) ROC curve on test set

certain performance metrics such as sensitivity, specificity, precision, and F1-score. A_O represents the accuracy of other reference models on the same performance metrics. For example, in experiment 1 (dataset 1), the proposed method achieve a sensitivity of 0.93, while SVM achieves 0.89. The percent improvement is calculated by subtracting the

value of the proposed method from the value of the reference method, such as SVM, i.e., $0.93 - 0.89 \times 100 = 4$ percent. This implies that the suggested approach outperforms SVM on sensitivity measures by percent. When compared to the RF, KNN, and NB models for the same metrics, improvements of 6, 7, and 10 percent are noted. In the case of specificity, performance gains of 2, 3, 4, and 6 percent are obtained when compared to the RF, KNN, and NB models. Similarly, the performance of the four models for accuracy, F1 score, and AUC ranges from 3 percent to 10 percent. For experiment 2 (dataset 2), the percentage improvements over four reference methods range from 1 percent to 7 percent.

$$\% \text{ improvement } = (\mathrm{A_P} - \mathrm{A_O}) \times 100 \tag{5.11}$$



(a) (b)

Figure 5.14: Comparison of percentage improvements over four methods (a) Improvements on benchmark dataset (experiment 1) (b) Improvements on real data of buildings (experiment 2)

5.7 Discussion and reasons

Based on the obtained results, the following comments are cited from this chapter.

- **Reason for improved accuracy:** The performance of the proposed method is better than traditional machine learning methods in three situations. Firstly, it automatically handles the missing data and class imbalance problem in the given datasets. Second, it captures efficient high-level feature representations of buildings load profile data, resulting in better performance on classification task. Third, it incorpo-

rates an MSTL approach that follows a two-phase protocol combining pretraining (unsupervised phase) and supervised learning phase to improve the accuracy.

- **Unsupervised feature learning:** The benefit of the proposed model is that it incorporates the automatic (unsupervised) feature learning with minimum human intervention. This is important in the context of building load profiles since in practical situations, each building has unique electric load patterns and the manual extraction of key features for each building is not possible. The current baseline approaches rely on handcrafted features, which leads to poor classification performance because they are unable to generalize due to a lack of ability to capture useful features. This demonstrates the importance and necessity of utilizing essential information and features for building profile classification.

- **Impact of different regions and datasets:** This study shows that diverse datasets and prevailing conditions of regions considerably affect the accuracy of models. A thorough analysis of classification systems require appropriate data. Unbalanced and incomplete data is hard to analyse [218]. To address this issue, a minority oversampling technique is proposed to fill in the missing data and balance the minority and majority samples.

- **Classification groups and implications:** The experiments are performed for diverse group of buildings including residential, industrial, and commercial sectors. The accurate and reliable classification of building load profiles is helpful for the development of building energy conservation measures. The results could contribute to the establishment of an energy saving policies, thereby possibly aiding energy managers to accurately predict electricity demand. The managers could conduct energy simulation by considering these load classifications. Thus, the energy-saving design of buildings can be optimized.

- **Effect of time interval:** In terms of time interval, the results showed that the classification accuracy improves with the increase in time interval. This has been verified by experimenting the results on 5 min and 1 hour interval datasets. It is observed that the classification accuracy is better on longer time intervals. This finding meets the fact that the increase of the data time interval can reduce the volatility of building load profiles, thereby making the load variation more stable and thus more predictable. This is also echoed in existing studies [219, 220].

- **Validation using two case studies:** The validation is performed using real measurements of buildings and benchmark datasets. The classification metrics including accuracy, F1-score, recall, precision, and AUC are used for performance evaluation. Case studies demonstrate that the proposed model obtained good results compared to state-of-the-art methods, exhibiting higher accuracy on all the evaluation metrics.

5.8 Concluding remarks

A novel deep learning framework for the classification of building load profiles is proposed. The goal is to classify different groups of buildings and to study the load profile of each group. The approach incorporates an unsupervised feature extraction based on SAEs to automatically learn the useful representations from data. The obtained features are then utilized in the MSTL technique to improve the classification accuracy. The MSTL approach employs a two-phase protocol that combines unsupervised pretraining with a supervised learning strategy. The model's performance is validated using both real-world and benchmark simulated datasets. The results of the 4-fold cross-validation demonstrate that the model exhibits a strong ability in classifying the building load profiles precisely. On benchmark buildings dataset, classification scores exceed 90 percent on the testing set, while accuracy exceeds 85 percent on the real-world building dataset. For fairness, the results are compared to the most commonly used approaches in the literature. It is found that the proposed methodology outperforms state-of-the-art methods, displaying superior accuracy across all evaluation metrics. More importantly, an analysis is performed to determine the percentage improvement in accuracy, and the findings are compared using standard performance metrics such as overall accuracy in confusion matrix, precision, recall, F1-score, and ROC-AUC values. The results indicate an increase in accuracy, with a percentage improvement of about 1 to 10 percent over conventional classification models.

As a benefit, the proposed approach is generic in the sense that it is not restricted to a specific classification task or a specific region. It can be used to create prototypes for developing more advanced tools for building energy classification. The findings of the study contribute to automate and improve the predictive modeling process while bridging the knowledge gaps between deep learning and building professionals. The created

algorithms performed well and considerably increased the viability of a building profile classification procedure in real-world building datasets.

Chapter 6

Conclusion and future research directions

This thesis began with three aims:

1. providing a data protection framework for anonymizing a distribution network's data and validating the solutions on the IEEE 123-node test system;

2. developing novel solutions for building a synthetic network and datasets, and validating them in practical environments such as industrial servers; and

3. addressing the issue of classifying the load profiles of buildings to effectively manage energy sources across power distribution networks.

These aims are pursued in the three core chapters of this thesis. In this chapter, the findings of this study are summarized, its contributions to the existing literature identified and future research directions and final remarks provided.

6.1 Research significance and outcomes

This research led to the development of algorithms that address the following issues:

- Firstly, a new way of conducting a literature review is presented by performing VOSviewer experiments. The scattered literature is transformed into visual presentations and key research gaps are identified in the form of visual clusters.

- Based on the findings of the review, a novel method for anonymizing distribution networks data using a statistical distribution and parameter estimation approach is proposed.
- An algorithm based on the MLE is proposed for finding the statistical distribution parameters that represent the actual data.
- A K-S test is conducted to make anonymized network datasets realistic. Two standard criteria are used to test the statistical compliance of anonymized datasets with real ones.
- A data anonymization process is developed to create representative anonymous datasets that can be used for research purposes without accessing the confidential data. The comparison results of actual and anonymized datasets are evaluated using three different datasets.
- In Chapter 2, new data synthesis algorithms are created for synthetic power distribution networks. The topology of power distribution lines is developed from public road infrastructure. The proposed method simplifies the design of power lines by using the concept of nodes and edges. This concept is supported in the power distribution planning book [175] and power system planners can leverage from this approach to select suitable routes for new power lines.
- An algorithm for computing the lengths of power line segments is proposed.
- A batch-geolocation algorithm is proposed for identifying the geographical location of households in a network.
- A hub-line algorithm is developed to identify the number of consumers connected to a single transformer. A standard distance between a transformer and the households is maintained by following the CIGRE C6.24 standards [16].
- An up-to-date repository of substation information is retrieved from the OSM approach. Multiple queries based on a query language are created in an overpass-turbo [178] to identify the locations of substations.
- A new algorithm for creating connectivity nodes data is generated using the concept of utility poles and their associated lines.

- A standard way of representing electrical connectivity between two or more equipment's in the network is provided by proposing fromNode and ToNode concept. Connectivity is established by defining the start and endpoints of the network elements.

- The datasets are created with interactive map-based visualizations. The maps display the topological structure of the developed network, allowing different sections of the network to be examined in detail. The interactive geographic maps show the names of the nodes and IDs of the network elements (substations and power transformers) assigned to them.

- In Chapter 3, two novel approaches based on SAEs and the MSTL are proposed. Different from standard hand-crafted feature representations, SAEs can learn meaningful features from vast amount buildings data in an unsupervised automated manner. This is significant as each building has unique electrical load patterns, and manually extracting the key features of every building is not possible in actual scenarios. An MSTL approach is developed to improve the classification accuracy by combining sequential unsupervised and supervised learning.

- A MOS method is proposed for dealing with incomplete and unbalanced real-world data. It effectively balances missing or unbalanced data by equalizing minority and majority samples for fair comparisons.

6.2 Comparisons with benchmarks

In this dissertation, several case studies that validate the proposed solutions are presented.

- In chapter 1, the methodology is tested on the IEEE 123-node system by anonymizing the test system distribution grid parameters. Firstly, an anonymized dataset is created for IEEE 123 node test feeder. Then, the anonymized IEEE 123 node feeder was simulated in OpenDSS by reference to the actual IEEE 123-node test feeder

- The results are compared on three different industrial datasets. The validation is performed by collecting the network datasets from a local DNSP in Canberra.

- Comparisons of existing methods and proposed solutions are conducted. The results are compared with recently published noise addition methods. Also, for fairness, the results are evaluated and compared on two metrics, the voltage profiles and power flow through the lines.

- In Chapter 2, the practical feasibility of the proposed algorithms is demonstrated by an illustrative case study of the Colac region in Australia. A synthetic network and dataset is created for the distribution feeder, and validated on industrial data platforms.

- The results of the proposed solutions are compared using a two-step validation process. In the first stage, solutions are validated by replicating them in real-world industrial data platforms such as EVOLVE [221], and in the second stage, solutions are verified using expert comments and validation. This method contributes to expanding the utility of synthetic networks and datasets from university researchers to industry users.

- Geographical validation of distribution network models is assessed using interactive maps in the QGIS platform which enables users to manage and visualize the key assets of an existing energy infrastructure.

- In Chapter 3, empirical comparison is conducted with the most widely used machine learning methods including SVM, RF, kNN, and NB. The standard performance criteria, a confusion matrix, ROC curves, recall, F1-score, specificity and precision, are used. For a fair assessment, an average percentage performance improvement obtained by the proposed method over traditional methods is computed.

- As the load profiles of buildings differ from region to region, the practical feasibility of the proposed solutions is shown through two case studies. Case study 1 involves testing the algorithms on a public benchmark dataset of buildings [191]. Case study 2 validates the results using real-world datasets of 105 buildings (residential, commercial, and industrial).

6.3 Extensions and future research directions

The algorithms developed in this thesis could be extended in the following ways.

1. In Chapter 3, the value and effectiveness of anonymization approaches are discussed. Anonymizing data can lead to the loss of its critical information. In future, more robust and advanced anonymization systems that can maintain a balance between data protection and utility are required to maintain data integrity in real-world applications.

2. In Chapter 4, novel solutions for developing a synthetic network and datasets are provided. The methodology is applied to the Australian case while demonstrating a case study of a distribution feeder. Future work includes the implementation of these methods in large scale networks to enhance the utility of the proposed solutions.

3. The electrical characteristics are not considered during synthetic network development. However, by integrating them with the topology of a synthetic network, their value for testing and developing new algorithms, such as a load-flow analysis, could be increased.

4. It is also noticed that a lack of available data platforms hinders the applications of synthetic networks and datasets. Data platforms that facilitate efficient querying of realistic distribution network models and sharing research results are required. Future work will include implementing cloud-based data platforms for data protection whereby the data will eventually be transmitted to a centralized platform server to provide real-time services to the targeted research community.

5. In Chapter 5, two novel AI-based approaches created to solve the problem of classifying load profiles are demonstrated. The methods perform well and considerably increase the viability of a building profile classification procedure in real-world building datasets. In the future, more intriguing approaches, such as graph neural networks, will be investigated to evaluate their expressive capacities in classification problems.

6.4 Key takeaways

The following are the key takeaways of this thesis.

- The evolving constraint of LV network visibility in power systems is addressed by developing synthetic networks and datasets. The associated challenges, such as privacy and confidentiality concerns in critical network data and the problem of load profiling classification in the distribution networks are addressed with novel solutions.

- The various approaches reported in this thesis have been evaluated through rigorous experimentation. For example, a case study of a synthetic network for the Colac region in Australia is presented as a practical application of the methodology. The methods are also experimentally proven by simulating them in the IEEE 123-node test feeder.

- The methodologies are validated on an industry scale and also by including the feedback from industry practitioners familiar with this field.

- The solutions are also shared with the wider audience in the field by submitting the results to peer-reviewed journals. In total, eleven research papers are produced from this research (eight as the lead author and three as a co-author). This thesis also resulted in an industry-sponsored research grant for solving issues associated with current electrical distribution networks.

It is hoped that the solutions presented in this thesis will kindle the interest of future researchers and practitioners in synthetic networks and datasets, and encourage them to develop more realistic distribution network models. The three developments highlighted in this thesis can be used to improve present network operations while continuing to expand innovations of synthetic networks and datasets in scientific research.

Appendix A

Supplementary Materials

Developed procedure for evaluating synthetic networks and datasets on industry (DNSP) servers

```
# Python environment
# Experimental setup and ingestion of synthetic networks
# Case study of Australian region, Evolve-project

import geopandas as gp
import zepben.evolve as ev
from zepben.evolve import connect_async, ProducerClient
from tkinter import filedialog
from tkinter import *
from pathlib import Path
import logging
import asyncio
import argparse
import pydash
import json
import os

logging.basicConfig(level=logging.DEBUG)
logger = logging.getLogger(__name__)


def get_path():
    root = Tk()
    root.filename = filedialog.askopenfilename(initialdir=Path.home(),
   title="Select file",
```

```
                                                    filetypes=(("jpeg files"
    , "*.geojson"), ("all files", "*.*")))
    return root.filename


def read_json_file(path):
    with open(path, "r") as f:
        return json.loads(f.read())


class Colac_Network:

    def __init__(self, path, namespace='evolve'):
        self.namespace = namespace
        self.path = "ColacDemoFeeder31.geojson"
        logger.info(f'Creating Network from: {path}')
        self.geojson_file = read_json_file(self.path)
        self.mapping = read_json_file('cim-mapping.json')
        self.config_file = read_json_file('geojson-config.json')
        self.feeder_name = os.path.basename(self.path)
        self.fdr = ev.Feeder(name='ColacDemoFeeder31', mrid='
    ColacDemo_mrid')
        self.headEqMrid = None
        self.gdf = gp.read_file(self.path)
        self.ns = ev.NetworkService()
        self.ds = ev.DiagramService()
        self.add_base_voltages()

    def get_cim_class(self, gis_class):
        if self.mapping.get(gis_class):
            return self.mapping[gis_class]["cimClass"]
        else:
            return None
    def get_field_name(self, field):
        if self.config_file.get(field):
            return self.config_file[field][self.namespace]

    def add_diagram(self):
        diagram = ev.Diagram(diagram_style=ev.DiagramStyle.GEOGRAPHIC)
        self.ds.add(diagram)
```

```
        return diagram

    def add_location(self, row):
        loc = ev.Location()
        for coord in row["geometry"].coords:
            logger.info(f'Creating coordinates: {coord}')
            loc.add_point(ev.PositionPoint(coord[0], coord[1]))
            logger.info('Add Location to Network Service')
            self.ns.add(loc)
        return loc

    def add_base_voltages(self):
        self.ns.add(ev.BaseVoltage(mrid='415V', nominal_voltage=415,
    name='415V'))
        self.ns.add(ev.BaseVoltage(mrid='11kV', nominal_voltage=11000,
    name='11kV'))
        self.ns.add(ev.BaseVoltage(mrid='22000', nominal_voltage=22000,
     name='22000'))
        self.ns.add(ev.BaseVoltage(mrid='UNKNOWN', nominal_voltage=0,
    name='UNKNOWN'))

    def create_equipment(self, row, loc):
        class_name = self.get_cim_class(row[self.get_field_name('class'
    )])
        if class_name is not None:
            logger.info(f'Creating CIM Class: {class_name}')
            class_ = getattr(ev, class_name)
            eq = class_()
            if isinstance(eq, ev.EnergySource):
                logger.info(f"Creating EnergySourcePhases for
    EnergySource: {eq}")
                esp_a = ev.EnergySourcePhase(phase=ev.SinglePhaseKind.A
    )
                esp_b = ev.EnergySourcePhase(phase=ev.SinglePhaseKind.B
    )
                esp_c = ev.EnergySourcePhase(phase=ev.SinglePhaseKind.C
    )
                self.ns.add(esp_a)
                self.ns.add(esp_b)
                self.ns.add(esp_c)
```

```
                eq.add_phase(esp_a)
                eq.add_phase(esp_b)
                eq.add_phase(esp_c)

            logger.info(f'Creating Equipment mRID: {row[self.
get_field_name("mrid")]}')
            eq.mrid = str(row[self.get_field_name("mrid")])
            eq.name = str(row[self.get_field_name("name")])
            eq.location = loc
            if type(eq) == ev.PowerTransformer:
                pte1 = ev.PowerTransformerEnd(power_transformer=eq)
                eq.add_end(pte1)
                pte2 = ev.PowerTransformerEnd(power_transformer=eq)
                eq.add_end(pte2)
                self.ns.add(eq)
                self.ns.add(pte1)
                self.ns.add(pte2)
                self.ns.add(eq)
                self.ns.add(pte1)
                self.ns.add(pte2)
                logger.info(f'Creating PowerTranformerEnds for
PowerTransfomer: {eq}')
            if row[self.get_field_name('baseVoltag')] is not None:
                logger.info(f'Assigning BaseVoltag: {row["baseVoltag"]}
')
                eq.base_voltage = self.ns.get(row[self.get_field_name('
baseVoltag')])
            else:
                logger.info(f'baseVoltag = None. Assigning BaseVoltag:
UNKNOWN')
                eq.base_voltage = self.ns.get('UNKNOWN')
        else:
            raise Exception(f'GIS Class: {row[self.get_field_name("
class")]} is not mapped to any Evolve Profile class')
        self.ns.add(eq)
        return eq

    def add_equipment(self):
        for index, row in self.gdf.iterrows():
            loc = self.add_location(row)
```

```
            eq = self.create_equipment(row, loc)
            if eq is not None:
                self.fdr.add_equipment(eq)
                eq.add_container(self.fdr)
                if row[self.get_field_name("headTermin")] == 1:
                    self.headEqMrid = row[self.get_field_name("mrid")]
                    logger.info(f'Detect head Equipment: {self.
headEqMrid}')
            else:
                logger.error(f'Equipment not mapped to a Evolve Profile
 class: {row[self.get_field_name("mrid")]}')
        self.ns.add(self.fdr)
        self.connect_Colac_equipment()
        return self.ns

    def connect_Colac_equipment(self):
        gdf_b = self.gdf[self.gdf['geometry'].apply(lambda x: x.type ==
 'LineString')]
        for index, row in gdf_b.iterrows():
            if row[self.get_field_name('fromEq')] is not None:
                logger.info(f'Connecting: {(row[self.get_field_name("
fromEq")])} to {row[self.get_field_name("toEq")]} '
                            f'with acls: {row[self.get_field_name("mrid
")]}')

            mrid_equipment = self.ns.get(mrid=str(row[self.
get_field_name('mrid')]))
            eq1 = self.ns.get(mrid=row[self.get_field_name('fromEq')])
            eq2 = self.ns.get(mrid=row[self.get_field_name('toEq')])

            geophases = ev.PhaseCode[row.get(self.get_field_name('
phases'), default="ABC")]
            Terminal_1 = ev.Terminal(conducting_equipment=
mrid_equipment, phases=geophases)
            Terminal_2 = ev.Terminal(conducting_equipment=
mrid_equipment, phases=geophases)

            mrid_equipment.add_terminal(Terminal_1)
            mrid_equipment.add_terminal(Terminal_2)

```

```
159             from_Equipment = ev.Terminal(conducting_equipment=eq1,
    phases=geophases)
160             eq1.add_terminal( from_Equipment)
161 
162             To_Equipment = ev.Terminal(conducting_equipment=eq2, phases
    =geophases)
163             eq2.add_terminal( To_Equipment)
164 
165             if eq1.mrid == self.headEqMrid and self.fdr.
    normal_head_terminal is None:
166                 logger.info(f'Assigning head terminal to Feeder for the
     Equipment {eq1.mrid}')
167                 setattr(self.fdr, 'normal_head_terminal',
    from_Equipment)
168             if eq2.mrid == self.headEqMrid and self.fdr.
    normal_head_terminal is None:
169                 logger.info(f'Assigning head terminal to Feeder for the
     Equipment {eq2.mrid}')
170                 setattr(self.fdr, 'normal_head_terminal', To_Equipment)
171             self.ns.add(Terminal_1)
172             self.ns.add(from_Equipment)
173             self.ns.add(Terminal_2)
174             self.ns.add(To_Equipment)
175             self.ns.connect_terminals(Terminal_1, from_Equipment)
176             self.ns.connect_terminals(Terminal_2, To_Equipment)
177 
178 
179 async def main():
180     parser = argparse.ArgumentParser(description="Zepben_UNSW cimbend
    demo for geoJSON ingestion")
181     parser.add_argument('server', help='Host and port of grpc server',
    metavar="host:port", nargs="?",
182                         default="localhost")
183     parser.add_argument('--rpc-port', help="The gRPC port for the
    server", default="50051")
184     parser.add_argument('--conf-address', help="The address to retrieve
     auth configuration from",
185                         default="http://localhost/auth")
186     parser.add_argument('--client-id', help='Auth0 M2M client id',
    default="")
```

```
187     parser.add_argument('--client-secret', help='Auth0 M2M client
    secret', default="")
188     parser.add_argument('--ca', help='CA trust chain', default="")
189     parser.add_argument('--cert', help='Signed certificate for your
    client', default="")
190     parser.add_argument('--key', help='Private key for signed cert',
    default="")
191     parser.add_argument('--geojson_path', help='Path of the geojson
    input file',
192                         default= "C:/Users/Ali/Desktop/Git projects
    /2021/evolve-python-sdk-tests/src/ColacDemoFeeder31.geojson")
193     args = parser.parse_args()
194     ca = cert = key = client_id = client_secret = None
195     if not args.client_id or not args.client_secret or not args.ca or
    not args.cert or not args.key:
196         logger.warning(
197             f"Using an insecure connection as at least one of (--ca, --
    token, --cert, --key) was not provided.")
198     else:
199         with open(args.key, 'rb') as f:
200             key = f.read()
201         with open(args.ca, 'rb') as f:
202             ca = f.read()
203         with open(args.cert, 'rb') as f:
204             cert = f.read()
205         client_secret = args.client_secret
206         client_id = args.client_id
207     # Creates a Network
208     network = Colac_Network(args.geojson_path).add_equipment()
209 
210     # Connect to a local cimcap instance using credentials if provided.
211     async with connect_async(host=args.server, rpc_port=args.rpc_port,
    conf_address=args.conf_address,
212                              client_id=client_id, client_secret=
    client_secret, pkey=key, cert=cert, ca=ca) as channel:
213         client = ProducerClient(channel)
214         # Send the network to the postbox instance.
215         res = await client.send([network])
216 
217 
```

```
218 if __name__ == "__main__":
219     loop = asyncio.get_event_loop()
220     loop.run_until_complete(main())
```

Bibliography

[1] C. Mateo, F. Postigo, F. de Cuadra, T. G. San Roman, T. Elgindy, P. Dueñas, B.-M. Hodge, V. Krishnan, and B. Palmintier, "Building large-scale us synthetic electric distribution system models," *IEEE Transactions on Smart Grid*, vol. 11, no. 6, pp. 5301–5313, 2020.

[2] "General electric (GE) global transmission and distribution grid challenges," https://www.ge.com/digital/lp/frost-and-sullivan-awards-ge-digital-product-leadership-award, 2021.

[3] E. Y. Dari, A. Bendahmane, and M. Essaaidi, "Verification-based data integrity mechanism in smart grid network," *International Journal of Security and Networks*, vol. 16, no. 1, pp. 1–11, 2021.

[4] J. Houghton and N. Gruen, "The value of research data: Open research data report," https://www.ands.org.au/working-with-data/articulating-the-value-of-open-data, 2020, [Online; accessed 22-October-2021].

[5] H. Ping, J. Stoyanovich, and B. Howe, "Datasynthesizer: Privacy-preserving synthetic datasets," in *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, 2017, pp. 1–5.

[6] N. Anuar, N. Baharin, N. Nizam, A. Fadzilah, S. Nazri, and N. Lip, "Determination of typical electricity load profile by using double clustering of fuzzy c-means and hierarchical method," in *2021 IEEE 12th Control and System Graduate Research Colloquium (ICSGRC)*. IEEE, 2021, pp. 277–280.

[7] Y. Sun, W. Gu, J. Lu, and Z. Yang, "Fuzzy clustering algorithm-based classification of daily electrical load patterns," in *2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*. IEEE, 2015, pp. 50–54.

[8] Z. Ma, A. Clausen, Y. Lin, and B. N. Jørgensen, "An overview of digitalization for the building-to-grid ecosystem," *Energy Informatics*, vol. 4, no. 2, pp. 1–21, 2021.

[9] Y. Pan and L. Zhang, "Data-driven estimation of building energy consumption with multi-source heterogeneous data," *Applied Energy*, vol. 268, p. 114965, 2020.

[10] "IEEE, PES AMPS DSAS Test Feeder Working Group, Test feeders." http://sites.ieee.org/pes-testfeeders/resources/ [Accessed: 1-Aug-2021].

[11] "Texas A&M university electric grid datasets." https://electricgrids.engr.tamu.edu/electric-grid-test-cases/datasets-for-arpa-e-perform-program/ [Accessed: 6-Aug-2021].

[12] "EPRI, (Electric Power Research institute)." https://smartgrid.epri.com/SimulationTool.aspx [Accessed: 6-Aug-2021].

[13] "ENTSO-E transmission system,." https://www.entsoe.eu/data/map/ [Accessed: 6-Aug-2021].

[14] "PNNL (pacific northwest national laboratory), "taxonomy of prototypical feeders,." https://sourceforge.net/p/gridlab-d/code/HEAD/tree/Taxonomy_Feeders/ [Accessed: 6-Aug-2021].

[15] "United kingdom generic distribution system (ukgds),." https://github.com/sedg/ukgds [Accessed: 6-Aug-2021].

[16] "Cigre capacity of distribution feeders for hosting der, 2014," https://www.cigreaustralia.org.au/assets/ITL-SEPT-2014/3.1-Capacity-of-Distribution-Feeders-for-hosting-Distributed-Energy-Resources-DER-abstract.pdf.

[17] "Open energy modelling (OPENMOD) initiative,." https://openmod-initiative.org/ [Accessed: 28-Aug-2021].

[18] "OPSD (open power system data),." https://open-power-system-data.org/data-projects [Accessed: 6-Aug-2021].

[19] "Open energy platform (OEP) ." https://openenergy-platform.org/ [Accessed: 16-Aug-2021].

[20] "Illinois center for a smarter electric grid (ICSEG) ." https://icseg.iti.illinois.edu/power-cases/ [Accessed: 26-Aug-2021].

[21] "MATPOWER test cases, "electric power system simulation and optimization,." https://github.com/MATPOWER/matpower [Accessed: 6-Aug-2021].

[22] "LINES (laboratory for intelligent integrated networks of engineering systems) ." http://amfarid.scripts.mit.edu/Datasets/SPG-Data/index.php [Accessed: 23-Aug-2021].

[23] S. Meinecke, D. Sarajlić, S. R. Drauz, A. Klettke, L.-P. Lauven, C. Rehtanz, A. Moser, and M. Braun, "Simbench—a benchmark dataset of electric power systems to compare innovative solutions based on power flow analysis," *Energies*, vol. 13, no. 12.

[24] V. Ayala-Rivera, P. McDonagh, T. Cerqueus, L. Murphy, C. Thorpe *et al.*, "Enhancing the utility of anonymized data by improving the quality of generalization hierarchies," *Transactions on Data Privacy*, vol. 10, no. 1, pp. 27–59, 2017.

[25] C.-Y. Lin, "Suppression techniques for privacy-preserving trajectory data publishing," *Knowledge-Based Systems*, vol. 206, p. 106354, 2020.

[26] Y. Chen, J.-F. Martínez, P. Castillejo, and L. López, "A privacy-preserving noise addition data aggregation scheme for smart grid," *Energies*, vol. 11, no. 11, p. 2972, 2018.

[27] J. Domingo-Ferrer, K. Muralidhar, and M. Bras-Amorós, "General confidentiality and utility metrics for privacy-preserving data publishing based on the permutation model," *IEEE Transactions on Dependable and Secure Computing*, 2020.

[28] J. Heldal and D.-C. Iancu, "Synthetic data generation for anonymization purposes. application on the norwegian survey on living conditions/ehis," 2019.

[29] K. Kapusta, G. Memmi, and H. Noura, "Secure and resilient scheme for data protection in unattended wireless sensor networks," in *2017 1st Cyber Security in Networking Conference (CSNet)*. IEEE, 2017, pp. 1–8.

[30] M. Dehghani, M. Ghiasi, T. Niknam, A. Kavousi-Fard, E. Tajik, S. Padmanaban, and H. Aliev, "Cyber attack detection based on wavelet singular entropy in ac smart islands: False data injection attack," *IEEE Access*, vol. 9, pp. 16 488–16 507, 2021.

[31] M. A. Ferrag, L. A. Maglaras, H. Janicke, J. Jiang, and L. Shu, "A systematic review of data protection and privacy preservation schemes for smart grid communications," *Sustainable cities and society*, vol. 38, pp. 806–835, 2018.

[32] X. Wang, J.-K. Chou, W. Chen, H. Guan, W. Chen, T. Lao, and K.-L. Ma, "A utility-aware visual approach for anonymizing multi-attribute tabular data," *IEEE transactions on visualization and computer graphics*, vol. 24, no. 1, pp. 351–360, 2017.

[33] J. Jasiūnas, P. D. Lund, and J. Mikkola, "Energy system resilience–a review," *Renewable and Sustainable Energy Reviews*, vol. 150, p. 111476, 2021.

[34] S. Xin, Q. Guo, J. Wang, C. Chen, H. Sun, and B. Zhang, "Information masking theory for data protection in future cloud-based energy management," *IEEE Transactions on Smart Grid*, vol. 9, no. 6, pp. 5664–5676, 2017.

[35] H. Habibzadeh, B. H. Nussbaum, F. Anjomshoa, B. Kantarci, and T. Soyata, "A survey on cybersecurity, data privacy, and policy issues in cyber-physical system deployments in smart cities," *Sustainable Cities and Society*, vol. 50, p. 101660, 2019.

[36] M. Ring, S. Wunderlich, D. Scheuring, D. Landes, and A. Hotho, "A survey of network-based intrusion detection data sets," *Computers & Security*, vol. 86, pp. 147–167, 2019.

[37] T. Ahmad, H. Chen, Y. Guo, and J. Wang, "A comprehensive overview on the data driven and large scale based approaches for forecasting of building energy demand: A review," *Energy and Buildings*, vol. 165, pp. 301–320, 2018.

[38] M. Bourdeau, X. qiang Zhai, E. Nefzaoui, X. Guo, and P. Chatellier, "Modeling and forecasting building energy consumption: A review of data-driven techniques," *Sustainable Cities and Society*, vol. 48, p. 101533, 2019.

[39] J. Arroyo, "Synergy between control theory and machine learning for building energy management," 2022.

[40] M. S. Piscitelli, S. Brandi, and A. Capozzoli, "Recognition and classification of typical load profiles in buildings with non-intrusive learning approach," *Applied Energy*, vol. 255, p. 113727, 2019.

[41] S. Zhong and K.-S. Tam, "Hierarchical classification of load profiles based on their characteristic attributes in frequency domain," *IEEE Transactions on Power Systems*, vol. 30, no. 5, pp. 2434–2441, 2014.

[42] A. A. A. Gassar and S. H. Cha, "Energy prediction techniques for large-scale buildings towards a sustainable built environment: A review," *Energy and Buildings*, p. 110238, 2020.

[43] S. Ikeda and T. Nagai, "A novel optimization method combining metaheuristics and machine learning for daily optimal operations in building energy and storage systems," *Applied Energy*, vol. 289, p. 116716, 2021.

[44] J.-S. Chou and D.-S. Tran, "Forecasting energy consumption time series using machine learning techniques based on usage patterns of residential householders," *Energy*, vol. 165, pp. 709–726, 2018.

[45] C. Yang, W. Shen, Q. Chen, and B. Gunay, "A practical solution for hvac prognostics: Failure mode and effects analysis in building maintenance," *Journal of Building Engineering*, vol. 15, pp. 26–32, 2018.

[46] D. A. Narciso and F. Martins, "Application of machine learning tools for energy efficiency in industry: A review," *Energy Reports*, vol. 6, pp. 1181–1199, 2020.

[47] S. Yilmaz, J. Chambers, and M. K. Patel, "Comparison of clustering approaches for domestic electricity load profile characterisation-implications for demand side management," *Energy*, vol. 180, pp. 665–677, 2019.

[48] M. H. Mohammadi and K. Saleh, "Synthetic benchmarks for power systems," *IEEE Access*, vol. 9, pp. 162 706–162 730, 2021.

[49] J. Y. Park, X. Yang, C. Miller, P. Arjunan, and Z. Nagy, "Apples or oranges? identification of fundamental load shape profiles for benchmarking buildings using a large and diverse dataset," *Applied energy*, vol. 236, pp. 1280–1295, 2019.

[50] Y. Wang, Q. Chen, C. Kang, M. Zhang, K. Wang, and Y. Zhao, "Load profiling and its application to demand response: A review," *Tsinghua Science and Technology*, vol. 20, no. 2, pp. 117–129, 2015.

[51] "Integrating distributed energy resources for the grid of the future," https://www.aemc.gov.au/sites/default/files/2019-09/Final%20report%20-%20ENERFR%202019%20-%20EPR0068.PDF, 2021, [Online; accessed 23-October-2021].

[52] C. Mateo, G. Prettico, T. Gómez, R. Cossent, F. Gangale, P. Frías, and G. Fulli, "European representative electricity distribution networks," *International Journal of Electrical Power & Energy Systems*, vol. 99, pp. 273–280, 2018.

[53] Y. Duan, C. Wang, and W. Zhou, "Topology modeling of distribution network based on open-source gis," in *International Conference on Electric Utility Deregulation and Restructuring and Power Technologies (DRPT)*.

[54] S. Papathanassiou, N. Hatziargyriou, P. Anagnostopoulos, L. Aleixo, B. Buchholz, C. Carter-Brown, N. Drossos, B. Enayati, M. Fan, V. Gabrion *et al.*, "Capacity of distribution feeders for hosting der," *CIGRE Working Group C*, vol. 6, 2014.

[55] M. Uslar, M. Specht, S. Rohjans, J. Trefke, and J. M. González, *The Common Information Model CIM: IEC 61968/61970 and 62325-A practical introduction to the CIM*. Springer Science & Business Media, 2012.

[56] L. Waltman and N. J. Van Eck, "A new methodology for constructing a publication-level classification system of science," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 12, pp. 2378–2392, 2012.

[57] N. J. Van Eck and L. Waltman, "Citation-based clustering of publications using citnetexplorer and vosviewer," *Scientometrics*, vol. 111, no. 2, pp. 1053–1070, 2017.

[58] L. Oneto, S. Ridella, and D. Anguita, "Differential privacy and generalization: Sharper bounds with applications," *Pattern Recognition Letters*, vol. 89, pp. 31–38, 2017.

[59] M. Orooji and G. M. Knapp, "Improving suppression to reduce disclosure risk and enhance data utility," *arXiv preprint arXiv:1901.00716*, 2019.

[60] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE, 2008, pp. 111–125.

[61] D. Li, X. He, L. Cao, and H. Chen, "Permutation anonymization," *Journal of Intelligent Information Systems*, vol. 47, no. 3, pp. 427–445, 2016.

[62] J. Yoon, L. N. Drumright, and M. Van Der Schaar, "Anonymization through data synthesis using generative adversarial networks (ads-gan)," *IEEE journal of biomedical and health informatics*, vol. 24, no. 8, pp. 2378–2388, 2020.

[63] S. U. Bazai, J. Jang-Jaccard, and H. Alavizadeh, "Scalable, high-performance, and generalized subtree data anonymization approach for apache spark," *Electronics*, vol. 10, no. 5, p. 589, 2021.

[64] C. Vigurs, C. Maidment, M. Fell, and D. Shipworth, "Customer privacy concerns as a barrier to sharing data about energy use in smart local energy systems: A rapid realist review," *Energies*, 2021.

[65] Y. Li and P. Wolfs, "Statistical identification of prototypical low voltage distribution feeders in western australia," in *2012 IEEE Power and Energy Society General Meeting*. IEEE, 2012, pp. 1–8.

[66] S. Xin, Q. Guo, J. Wang, C. Chen, H. Sun, and B. Zhang, "Information masking theory for data protection in future cloud-based energy management," *IEEE Transactions on Smart Grid*, vol. 9, no. 6, pp. 5664–5676, 2018.

[67] S. Shaham, M. Ding, B. Liu, S. Dang, Z. Lin, and J. Li, "Privacy preserving location data publishing: A machine learning approach," *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[68] S. Armoogum and V. Bassoo, "Privacy of energy consumption data of a household in a smart grid," in *Smart Power Distribution Systems*. Elsevier, 2019, pp. 163–177.

[69] A. Cui, H. Zhao, X. Zhang, B. Zhao, and Z. Li, "Power system real time data encryption system based on des algorithm," in *2021 13th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*. IEEE, 2021, pp. 220–228.

[70] S. Belguith, N. Kaaniche, and M. Hammoudeh, "Analysis of attribute-based cryptographic techniques and their application to protect cloud services," *Transactions on Emerging Telecommunications Technologies*, p. e3667, 2019.

[71] V. L. Narayana and C. Bharathi, "Multi-mode routing mechanism with cryptographic techniques and reduction of packet drop using 2ack scheme manets," in *Smart Intelligent Computing and Applications*. Springer, 2019, pp. 649–658.

[72] Z. Guan, Y. Zhang, L. Zhu, L. Wu, and S. Yu, "Effect: An efficient flexible privacy-preserving data aggregation scheme with authentication in smart grid," *Science China Information Sciences*, vol. 62, no. 3, p. 32103, 2019.

[73] Y. Su, Y. Li, J. Li, and K. Zhang, "Lceda: Lightweight and communication efficient data aggregation scheme for smart grid," *IEEE Internet of Things Journal*, 2021.

[74] G. Minello, L. Rossi, and A. Torsello, "k-anonymity on graphs using the szemerédi regularity lemma," *IEEE Transactions on Network Science and Engineering*, 2020.

[75] M. I. Pramanik, R. Y. Lau, M. S. Hossain, M. M. Rahoman, S. K. Debnath, M. G. Rashed, and M. Z. Uddin, "Privacy preserving big data analytics: A critical analysis of state-of-the-art," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 11, no. 1, p. e1387, 2021.

[76] L. El Haourani, A. A. El Kalam, and A. A. Ouahman, "Big data security and privacy techniques," in *Proceedings of the 3rd International Conference on Networking, Information Systems & Security*, 2020, pp. 1–9.

[77] Y. Sei, H. Okumura, T. Takenouchi, and A. Ohsuga, "Anonymization of sensitive quasi-identifiers for l-diversity and t-closeness," *IEEE transactions on dependable and secure computing*, 2017.

[78] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *2007 IEEE 23rd International Conference on Data Engineering*. IEEE, 2007, pp. 106–115.

[79] M. Yamaç, M. Ahishali, N. Passalis, J. Raitoharju, B. Sankur, and M. Gabbouj, "Multi-level reversible data anonymization via compressive sensing and data hiding," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1014–1028, 2020.

[80] S. Martínez, F. Sebé, and C. Sorge, "Measuring privacy in smart metering anonymized data," *arXiv preprint arXiv:2002.04863*, 2020.

[81] Z. Wang and T. Hong, "Generating realistic building electrical load profiles through the generative adversarial network (gan)," *Energy and Buildings*, vol. 224, p. 110299, 2020.

[82] H. Wei, Z. Hongxuan, D. Yu, W. Yiting, D. Ling, and X. Ming, "Short-term optimal operation of hydro-wind-solar hybrid system with improved generative adversarial networks," *Applied Energy*, vol. 250, pp. 389–403, 2019.

[83] J. Medková, "High-degree noise addition method for the k-degree anonymization algorithm," in *2020 Joint 11th International Conference on Soft Computing and*

Intelligent Systems and 21st International Symposium on Advanced Intelligent Systems (SCIS-ISIS). IEEE, 2020, pp. 1–6.

[84] M. Gough, S. Santos, T. Alskaif, M. Javadi, R. Castro, and J. P. Catalao, "Preserving privacy of smart meter data in a smart grid environment," *IEEE Transactions on Industrial Informatics*, 2021.

[85] X.-Y. Zhang and S. Kuenzel, "Differential privacy for deep learning-based online energy disaggregation system," in *2020 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe)*. IEEE, 2020, pp. 904–908.

[86] T. W. Mak, F. Fioretto, and P. Van Hentenryck, "Privacy-preserving obfuscation for distributed power systems," *Electric Power Systems Research*, vol. 189, p. 106718, 2020.

[87] L. Rocher, J. M. Hendrickx, and Y.-A. De Montjoye, "Estimating the success of re-identifications in incomplete datasets using generative models," *Nature communications*, vol. 10, no. 1, pp. 1–9, 2019.

[88] S. Virupaksha and V. Dondeti, "Anonymized noise addition in subspaces for privacy preserved data mining in high dimensional continuous data," *Peer-to-Peer Networking and Applications*, vol. 14, no. 3, pp. 1608–1628, 2021.

[89] B. Denham, R. Pears, and M. A. Naeem, "Enhancing random projection with independent and cumulative additive noise for privacy-preserving data stream mining," *Expert Systems with Applications*, vol. 152, p. 113380, 2020.

[90] J. Yan, G.-H. Yang, and Y. Wang, "Dynamic reduced-order observer-based detection of false data injection attacks with application to smart grid systems," *IEEE Transactions on Industrial Informatics*, 2022.

[91] H. Kim, D. Olave-Rojas, E. Álvarez-Miranda, and S.-W. Son, "In-depth data on the network structure and hourly activity of the central chilean power grid," *Scientific data*, vol. 5, no. 1, pp. 1–10, 2018.

[92] H. Li, J. H. Yeo, J. L. Wert, and T. J. Overbye, "Steady-state scenario development for synthetic transmission systems," in *2020 IEEE Texas Power and Energy Conference (TPEC)*. IEEE, 2020, pp. 1–6.

[93] W. H. Kersting, "Radial distribution test feeders," *IEEE Transactions on Power Systems*, vol. 6, no. 3, pp. 975–985, 1991.

[94] K. P. Schneider, Y. Chen, D. Engle, and D. Chassin, "A taxonomy of north american radial distribution feeders," in *2009 IEEE Power Energy Society General Meeting*, 2009, pp. 1–6.

[95] K. M. Gegner, A. B. Birchfield, Ti Xu, K. S. Shetye, and T. J. Overbye, "A methodology for the creation of geographically realistic synthetic power flow models," in *2016 IEEE Power and Energy Conference at Illinois (PECI)*, 2016.

[96] A. B. Birchfield, T. Xu, and T. J. Overbye, "Power flow convergence and reactive power planning in the creation of large synthetic grids," *IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 6667–6674, 2018.

[97] H. Li, A. L. Bornsheuer, T. Xu, A. B. Birchfield, and T. J. Overbye, "Load modeling in synthetic electric grids," in *2018 IEEE Texas Power and Energy Conference (TPEC)*. IEEE, 2018, pp. 1–6.

[98] A. B. Birchfield, T. Xu, K. M. Gegner, K. S. Shetye, and T. J. Overbye, "Grid structural characteristics as validation criteria for synthetic networks," *IEEE Transactions on Power Systems*, vol. 32, no. 4, pp. 3258–3265, 2017.

[99] Z. Wang, S. H. Elyas, and R. J. Thomas, "Generating synthetic electric power system data with accurate electric topology and parameters," in *2016 51st International Universities Power Engineering Conference (UPEC)*. IEEE, 2016, pp. 1–6.

[100] H. Li, J. L. Wert, A. B. Birchfield, T. J. Overbye, T. G. San Roman, C. M. Domingo, F. E. P. Marcos, P. D. Martinez, T. Elgindy, and B. Palmintier, "Building highly detailed synthetic electric grid data sets for combined transmission and distribution

systems," *IEEE Open Access Journal of Power and Energy*, vol. 7, pp. 478–488, 2020.

[101] R. Espejo, S. Lumbreras, and A. Ramos, "A complex-network approach to the generation of synthetic power transmission networks," *IEEE Systems Journal*, vol. 13, no. 3, pp. 3050–3058, 2018.

[102] V. Rosato, S. Bologna, and F. Tiriticco, "Topological properties of high-voltage electrical transmission networks," *Electric Power Systems Research*, vol. 77, no. 2, pp. 99–105, 2007.

[103] M. H. Athari and Z. Wang, "Introducing voltage-level dependent parameters to synthetic grid electrical topology," *IEEE Transactions on Smart Grid*, vol. 10, no. 4, pp. 4048–4056, 2018.

[104] P. Wlazlo, K. Price, C. Veloz, A. Sahu, H. Huang, A. Goulart, K. Davis, and S. Zounouz, "A cyber topology model for the texas 2000 synthetic electric power grid," in *2019 Principles, Systems and Applications of IP Telecommunications (IPTComm)*. IEEE, 2019, pp. 1–8.

[105] E. Schweitzer, A. Scaglione, A. Monti, and G. A. Pagani, "Automated generation algorithm for synthetic medium voltage radial distribution systems," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 7, no. 2, pp. 271–284, 2017.

[106] R. Kadavil, T. M. Hansen, and S. Suryanarayanan, "An algorithmic approach for creating diverse stochastic feeder datasets for power systems co-simulations," in *2016 IEEE Power and Energy Society General Meeting (PESGM)*, 2016, pp. 1–5.

[107] A. Trpovski, D. Recalde, and T. Hamacher, "Synthetic distribution grid generation using power system planning: Case study of singapore," in *2018 53rd International Universities Power Engineering Conference (UPEC)*, 2018, pp. 1–6.

[108] F. Postigo, C. Mateo, T. Gómez, F. de Cuadra, P. Dueñas, T. Elgindy, B.-M. Hodge, B. Palmintier, and V. Krishnan, "Phase-selection algorithms to minimize cost and

imbalance in us synthetic distribution systems," *International Journal of Electrical Power & Energy Systems*, vol. 120, p. 106042, 2020.

[109] Y. Liu, Z. Mao, H. Li, K. S. Shetye, and T. J. Overbye, "Integration of renewable generators in synthetic electric grids for dynamic analysis," *arXiv preprint arXiv:2101.02319*, 2021.

[110] H. Li, J. H. Yeo, A. L. Bornsheuer, and T. J. Overbye, "The creation and validation of load time series for synthetic electric power systems," *IEEE Transactions on Power Systems*, vol. 36, no. 2, pp. 961–969, 2020.

[111] F. E. Postigo Marcos, C. Mateo Domingo, T. Gomez San Roman, B. Palmintier, B.-M. Hodge, V. Krishnan, F. de Cuadra García, and B. Mather, "A review of power distribution test feeders in the united states and the need for synthetic representative networks," *Energies*, vol. 10, no. 11, p. 1896, 2017.

[112] K. Schneider, P. Phanivong, and J. Lacroix, "Ieee 342-node low voltage networked test system," in *2014 IEEE PES General Meeting — Conference Exposition*, 2014, pp. 1–5.

[113] "Scientific Grid (SciGrid) (pacific northwest national laboratory), "taxonomy of prototypical feeders,." http://git.scigrid.de/ [Accessed: 6-Aug-2021].

[114] "CIGRE (conseil international des grands réseaux electriques) power system test cases." https://e-cigre.org/publication/736-power-system-test-cases-for-emt-type-simulation-studies.

[115] "SimBench - benchmark data set for grid analysis, grid planning and grid operation management," 2020," https://simbench.de/en/ [Accessed: 26-Aug-2021].

[116] R. Atat, M. Ismail, M. F. Shaaban, E. Serpedin, and T. Overbye, "Stochastic geometry-based model for dynamic allocation of metering equipment in spatio-temporal expanding power grids," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 2080–2091, 2019.

[117] A. B. Birchfield, K. M. Gegner, T. Xu, K. S. Shetye, and T. J. Overbye, "Statistical considerations in the creation of realistic synthetic power grids for geomagnetic disturbance studies," *IEEE Transactions on Power Systems*, vol. 32, no. 2, pp. 1502–1510, 2016.

[118] K. M. Gegner, A. B. Birchfield, T. Xu, K. S. Shetye, and T. J. Overbye, "A methodology for the creation of geographically realistic synthetic power flow models," in *2016 IEEE Power and Energy Conference at Illinois (PECI)*.

[119] S. Soltan and G. Zussman, "Generation of synthetic spatially embedded power grid networks," in *2016 IEEE Power and Energy Society General Meeting (PESGM)*. IEEE, 2016, pp. 1–5.

[120] H. Sadeghian and Z. Wang, "Autosyngrid: A matlab-based toolkit for automatic generation of synthetic power grids," *International Journal of Electrical Power & Energy Systems*, vol. 118, p. 105757, 2020.

[121] T. E. McDermott, "A test feeder for dg protection analysis," in *2011 IEEE/PES Power Systems Conference and Exposition*, 2011, pp. 1–7.

[122] K. P. Schneider and J. C. Fuller, "Voltage control devices on the ieee 8500 node test feeder," in *IEEE PES T&D 2010*. IEEE, 2010, pp. 1–6.

[123] A. Rajabi, M. Eskandari, M. J. Ghadi, L. Li, J. Zhang, and P. Siano, "A comparative study of clustering techniques for electrical load pattern segmentation," *Renewable and Sustainable Energy Reviews*, vol. 120, p. 109628, 2020.

[124] M. Bourdeau, P. Basset, S. Beauchêne, D. Da Silva, T. Guiot, D. Werner, and E. Nefzaoui, "Classification of daily electric load profiles of non-residential buildings," *Energy and Buildings*, vol. 233, p. 110670, 2021.

[125] J. Yang, J. Zhao, F. Wen, and Z. Dong, "A model of customizing electricity retail prices based on load profile clustering analysis," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 3374–3386, 2018.

[126] K. Li, Z. Ma, D. Robinson, and J. Ma, "Identification of typical building daily electricity usage profiles using gaussian mixture model-based clustering and hierarchical clustering," *Applied energy*, vol. 231, pp. 331–342, 2018.

[127] L. Wen, K. Zhou, and S. Yang, "A shape-based clustering method for pattern recognition of residential electricity consumption," *Journal of cleaner production*, vol. 212, pp. 475–488, 2019.

[128] S. Zhan, Z. Liu, A. Chong, and D. Yan, "Building categorization revisited: A clustering-based approach to using smart meter data for building energy benchmarking," *Applied Energy*, vol. 269, p. 114920, 2020.

[129] C. Jie, Z. Jiyue, W. Junhui, W. Yusheng, S. Huiping, and L. Kaiyan, "Review on the research of k-means clustering algorithm in big data," in *2020 IEEE 3rd International Conference on Electronics and Communication Engineering (ICECE)*. IEEE, 2020, pp. 107–111.

[130] Y. P. Raykov, A. Boukouvalas, F. Baig, and M. A. Little, "What to do when k-means clustering fails: a simple yet principled alternative algorithm," *PloS one*, vol. 11, no. 9, p. e0162259, 2016.

[131] B. A. Hassan and T. A. Rashid, "A multidisciplinary ensemble algorithm for clustering heterogeneous datasets," *Neural Computing and Applications*, pp. 1–24, 2021.

[132] X. Wang and X. Wang, "A fast k-medoids clustering algorithm for image segmentation based object recognition," *J Robotics Autom*, vol. 4, no. 1, pp. 202–211, 2020.

[133] S. Paudel, M. Elmitri, S. Couturier, P. H. Nguyen, R. Kamphuis, B. Lacarrière, and O. Le Corre, "A relevant data selection method for energy consumption prediction of low energy building based on support vector machine," *Energy and Buildings*, vol. 138, pp. 240–256, 2017.

[134] M. Mansoor, F. Grimaccia, S. Leva, and M. Mussetta, "Comparison of echo state network and feed-forward neural networks in electrical load forecasting for de-

mand response programs," *Mathematics and Computers in Simulation*, vol. 184, pp. 282–293, 2021.

[135] B. Najafi, M. Depalo, F. Rinaldi, and R. Arghandeh, "Building characterization through smart meter data analytics: Determination of the most influential temporal and importance-in-prediction based features," *Energy and Buildings*, vol. 234, p. 110671, Mar. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378778820334575

[136] M. R. Haq and Z. Ni, "Classification of Electricity Load Profile Data and The Prediction of Load Demand Variability," in *2019 IEEE International Conference on Electro Information Technology (EIT)*, May 2019, pp. 304–309, iSSN: 2154-0373.

[137] M. Fahim and A. Sillitti, "Analyzing Load Profiles of Energy Consumption to Infer Household Characteristics Using Smart Meters," *Energies*, vol. 12, no. 5, p. 773, Jan. 2019, number: 5 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1996-1073/12/5/773

[138] C. Si, S. Xu, C. Wan, D. Chen, W. Cui, and J. Zhao, "Electric load clustering in smart grid: Methodologies, applications, and future trends," *Journal of Modern Power Systems and Clean Energy*, vol. 9, no. 2, pp. 237–252, 2021.

[139] S. Zhong and K.-S. Tam, "A frequency domain approach to characterize and analyze load profiles," *IEEE Transactions on Power Systems*, vol. 27, no. 2, pp. 857–865, 2011.

[140] M. Sanabria-Villamizar, M. Bueno-López, J. C. Hernández, and D. Vera, "Characterization of household-consumption load profiles in the time and frequency domain," *International Journal of Electrical Power & Energy Systems*, vol. 137, p. 107756, 2022.

[141] S. Zhong, "Electricity load modeling in frequency domain," Ph.D. dissertation, Virginia Tech, 2017.

[142] E. Carpaneto, G. Chicco, R. Napoli, and M. Scutariu, "Electricity customer classification using frequency–domain load pattern data," *International Journal of Electrical Power & Energy Systems*, vol. 28, no. 1, pp. 13–20, 2006.

[143] S. Yan, K. Li, F. Wang, X. Ge, X. Lu, Z. Mi, H. Chen, and S. Chang, "Time–frequency feature combination based household characteristic identification approach using smart meter data," *IEEE Transactions on Industry Applications*, vol. 56, no. 3, pp. 2251–2262, 2020.

[144] X. Wu, X. Han, L. Liu, and B. Qi, "A load identification algorithm of frequency domain filtering under current underdetermined separation," *IEEE Access*, vol. 6, pp. 37 094–37 107, 2018.

[145] A. M. Tureczek and P. S. Nielsen, "Structured literature review of electricity consumption classification using smart meter data," *Energies*, vol. 10, no. 5, p. 584, 2017.

[146] Y. Zhao, C. Zhang, Y. Zhang, Z. Wang, and J. Li, "A review of data mining technologies in building energy systems: Load prediction, pattern identification, fault detection and diagnosis," *Energy and Built Environment*, vol. 1, no. 2, pp. 149–164, 2020.

[147] K. Grolinger, A. L'Heureux, M. A. Capretz, and L. Seewald, "Energy forecasting for event venues: Big data and prediction accuracy," *Energy and buildings*, vol. 112, pp. 222–233, 2016.

[148] J. A. Jardini, C. M. Tahan, M. R. Gouvea, S. U. Ahn, and F. Figueiredo, "Daily load profiles for residential, commercial and industrial low voltage consumers," *IEEE Transactions on power delivery*, vol. 15, no. 1, pp. 375–380, 2000.

[149] V. Grimm, A. S. Johnston, H.-H. Thulke, V. Forbes, and P. Thorbek, "Three questions to ask before using model outputs for decision support," *Nature communications*, vol. 11, no. 1, pp. 1–3, 2020.

[150] R. Markovič, M. Gosak, V. Grubelnik, M. Marhl, and P. Virtič, "Data-driven classification of residential energy consumption patterns by means of functional connectivity networks," *Applied energy*, vol. 242, pp. 506–515, 2019.

[151] M. A. Devlin and B. P. Hayes, "Non-intrusive load monitoring and classification of activities of daily living using residential smart meter data," *IEEE Transactions on Consumer Electronics*, vol. 65, no. 3, pp. 339–348, 2019.

[152] B. Yildiz, J. I. Bilbao, J. Dore, and A. Sproul, "Household electricity load forecasting using historical smart meter data with clustering and classification techniques," in *2018 IEEE Innovative Smart Grid Technologies-Asia (ISGT Asia)*. IEEE, 2018, pp. 873–879.

[153] B. Zhao, M. Ye, L. Stankovic, and V. Stankovic, "Non-intrusive load disaggregation solutions for very low-rate smart meter data," *Applied Energy*, vol. 268, p. 114949, 2020.

[154] K. Foteinaki, R. Li, C. Rode, and R. K. Andersen, "Modelling household electricity load profiles based on danish time-use survey data," *Energy and Buildings*, vol. 202, p. 109355, 2019.

[155] Y. Q. Ang, Z. M. Berzolla, and C. F. Reinhart, "From concept to application: A review of use cases in urban building energy modeling," *Applied Energy*, vol. 279, p. 115738, 2020.

[156] N. Huyghues-Beaufond, S. Tindemans, P. Falugi, M. Sun, and G. Strbac, "Robust and automatic data cleansing method for short-term load forecasting of distribution feeders," *Applied Energy*, 2020.

[157] F. Wang, L. Li, C. Li, Q. Wu, Y. Cao, B. Zhou, and B. Fang, "Fractal characteristics analysis of blackouts in interconnected power grid," *IEEE Transactions on Power Systems*, vol. 33, no. 1, pp. 1085–1086, 2017.

[158] A. C. Tamhane, *Statistical Analysis of Designed Experiments: Theory and Applications, Appendix C: Statistical Tables*, 1st ed., ser. Wiley Series

in Probability and Statistics. Wiley, Mar. 2009. [Online]. Available: https://onlinelibrary.wiley.com/doi/book/10.1002/9781118491621

[159] Z. Charles, "Kolmogorov-Smirnov Table | Real Statistics Using Excel," 2020. [Online]. Available: https://www.real-statistics.com/statistics-tables/kolmogorov-smirnov-table/

[160] T. Matsumura, M. Tsukamoto, A. Tsusaka, K. Yukita, Y. Goto, Y. Yokomizu, K. Tatewaki, D. Iioka, H. Shimizu, Y. Kanazawa, H. Ishikawa, A. Matsuo, and H. Iwatsuki, "Line-End Voltage and Voltage Profile along Power Distribution Line with Large-Power Photovoltaic Generation System," *International Journal of Photoenergy*, vol. 2019, pp. 1–8, Mar. 2019. [Online]. Available: https://www.hindawi.com/journals/ijp/2019/1263480/

[161] N. Balakrishnan, V. Voinov, and M. S. Nikulin, *Chi-squared goodness of fit tests with applications*. Academic Press, 2013.

[162] Y. Pawitan, *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press, 2013.

[163] B. F. Ginos, "Parameter estimation for the lognormal distribution," 2009.

[164] A. R. Joseph Hilbe, *Methods of Statistical Model Estimation*. Chapman and Hall/CRC, 2013.

[165] "EPRI | Smart Grid Resource Center > Simulation Tool – OpenDSS," https://smartgrid.epri.com/SimulationTool.aspx, last accessed 2021-03-29.

[166] H. Li, A. Zhang, X. Shen, and J. Xu, "A load flow method for weakly meshed distribution networks using powers as flow variables," *International Journal of Electrical Power & Energy Systems*, vol. 58, pp. 291–299, Jun. 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0142061514000283

[167] W.-T. Huang and K.-C. Yao, "New network sensitivity-based approach for real-time complex power flow calculation," *IET Generation, Transmission*

& Distribution, vol. 6, no. 2, p. 109, 2012. [Online]. Available: https://digital-library.theiet.org/content/journals/10.1049/iet-gtd.2011.0429

[168] Y.-T. Tsou, H.-L. Chen, and Y.-H. Chang, "Rod: Evaluating the risk of data disclosure using noise estimation for differential privacy," *IEEE Transactions on Big Data*, 2019.

[169] K. Mivule, "Utilizing noise addition for data privacy, an overview," 2013.

[170] S. Hamzehzadeh and S. M. Mazinani, "Annm: A new method for adding noise nodes which are used recently in anonymization methods in social networks," *Wireless Personal Communications*, vol. 107, no. 4, pp. 1995–2017, 2019.

[171] C. Eyupoglu, M. A. Aydin, A. H. Zaim, and A. Sertbas, "An efficient big data anonymization algorithm based on chaos and perturbation techniques," *Entropy*, vol. 20, no. 5, p. 373, 2018.

[172] L. Qi, X. Zhang, S. Li, S. Wan, Y. Wen, and W. Gong, "Spatial-temporal data-driven service recommendation with privacy-preservation," *Information Sciences*, vol. 515, pp. 91–102, 2020.

[173] L. Qi, C. Hu, X. Zhang, M. R. Khosravi, S. Sharma, S. Pang, and T. Wang, "Privacy-aware data fusion and prediction with spatial-temporal context for smart city industrial environment," *IEEE Transactions on Industrial Informatics*, 2020.

[174] M. S. Roy, R. Gupta, J. K. Chandra, K. D. Sharma, and A. Talukdar, "Improving photoplethysmographic measurements under motion artifacts using artificial neural network for personal healthcare," *IEEE Transactions on Instrumentation and Measurement*, vol. 67, no. 12, pp. 2820–2829, 2018.

[175] H. L. Willis, *Power distribution planning reference book*. CRC press, 2004.

[176] "Department of environment, land, water and planning, victoria, australia," [Online].Available:http://services.land.vic.gov.au/SpatialDatamart/index.jsp#.[Accessed:1-Aug-2021].

[177] QGIS Development Team, *QGIS Geographic Information System*, Open Source Geospatial Foundation, 2021. [Online]. Available: http://qgis.org

[178] M. Raifer *et al.*, "Overpass turbo, 2021," https://overpass-turbo.eu.

[179] "GIS network extractor from GIS json files,." https://zepben.github.io/evolve/docs/gis-network-extractor/2.12.0.

[180] H. L. Willis, *Spatial electric load forecasting*. CRC Press, 2002.

[181] "Vector geometry-qgis documentation 23.1.15." [Online].Available:https://docs.qgis.org/3.10/en/docs/user_manual/processing_algs/qgis/vectorgeometry.html#add-geometry-attribute.[Accessed:1-Aug-2021].

[182] "Join by lines (hub lines), qgis, "23.1.12. vector analysis — qgis documentation. ," 2020," https://docs.qgis.org/3.10/en/docs/user_manual/processing_algs/qgis/vectoranalysis.html#join-by-lines-hub-lines.[Accessed:1-Aug-2021].

[183] "Creating points along a line, 2021," https://desktop.arcgis.com/en/arcmap/latest/manage-data/creating-new-features/creating-new-points-along-a-line.htm.

[184] D. A. W. McMorran, "An Introduction to IEC 61970-301 & 61968-11: The Common Information Model," p. 41.

[185] "Standard for distribution line design overhead, 2020," https://www.ergon.com.au/__data/assets/pdf_file/0020/326630/STNW3361-Distribution-Line-Design-Overhead.pdf.

[186] "Pyqgis, qgsgeometry interpolate method, 2021," https://qgis.org/api/classQgsGeometry.html#a58b57cc606fabaf4e26c97092cba345b.

[187] "Network topology checker, qgis, 2021," https://docs.qgis.org/3.16/en/docs/user_manual/plugins/core_plugins/plugins_topology_checker.html.

[188] "Sql functions used with st_geometry, 2021," https://desktop.arcgis.com/en/arcmap/latest/manage-data/using-sql-with-gdbs/a-quick-tour-of-sql-functions-used-with-st-geometry.htm.

[189] A. Xenophon and D. Hill, "Open grid model of australia's national electricity market allowing backtesting against historic data," *Scientific data*, vol. 5, no. 1, pp. 1–21, 2018.

[190] B. Zeppelin, "Evolve cim profile, energy workbench server," https://zepben.github.io/evolve/docs/, 2021.

[191] F. Angizeh, A. Ghofrani, and M. Jafari, "Dataset on hourly load pro-files for a set of 24 facilities from industrial commercial and residen-tial end-use sectors," *Mendeley Data*, vol. 1, 2020.

[192] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, "Data imbalance in classification: Experimental evaluation," *Information Sciences*, vol. 513, pp. 429–441, 2020.

[193] S. S. Mullick, S. Datta, S. G. Dhekane, and S. Das, "Appropriateness of performance indices for imbalanced data classification: An analysis," *Pattern Recognition*, vol. 102, p. 107197, 2020.

[194] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[195] R. Atienza, *Advanced Deep Learning with Keras: Apply deep learning techniques, autoencoders, GANs, variational autoencoders, deep reinforcement learning, policy gradients, and more*. Packt Publishing Ltd, 2018.

[196] G. Zhong, L.-N. Wang, X. Ling, and J. Dong, "An overview on data representation learning: From traditional feature learning to recent deep learning," *The Journal of Finance and Data Science*, vol. 2, no. 4, pp. 265–278, 2016.

[197] M. F. I. Ibrahim and A. A. Al-jumaily, "Ica based feature learning and feature selection," in *2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA)*. IEEE, 2016, pp. 1–4.

[198] F. Luo, B. Du, L. Zhang, L. Zhang, and D. Tao, "Feature learning using spatial-spectral hypergraph discriminant analysis for hyperspectral image," *IEEE transactions on cybernetics*, vol. 49, no. 7, pp. 2406–2419, 2018.

[199] A. C. Belkina, C. O. Ciccolella, R. Anno, R. Halpert, J. Spidlen, and J. E. Snyder-Cappione, "Automated optimized parameters for t-distributed stochastic neighbor embedding improve visualization and analysis of large datasets," *Nature communications*, vol. 10, no. 1, pp. 1–12, 2019.

[200] J. Fontaine, M. Ridolfi, B. Van Herbruggen, A. Shahid, and E. De Poorter, "Edge inference for uwb ranging error correction using autoencoders," *IEEE Access*, vol. 8, pp. 139 143–139 155, 2020.

[201] J. Chen, Z. Wu, and J. Zhang, "Driving safety risk prediction using cost-sensitive with nonnegativity-constrained autoencoders based on imbalanced naturalistic driving data," *IEEE transactions on intelligent transportation systems*, vol. 20, no. 12, pp. 4450–4465, 2019.

[202] S. Lander and Y. Shang, "Evoae–a new evolutionary method for training autoencoders for deep learning networks," in *2015 IEEE 39th Annual Computer Software and Applications Conference*, vol. 2. IEEE, 2015, pp. 790–795.

[203] C. Battey, G. C. Coffing, and A. D. Kern, "Visualizing population structure with variational autoencoders," *G3*, vol. 11, no. 1, pp. 1–11, 2021.

[204] F. Anowar, S. Sadaoui, and B. Selim, "Conceptual and empirical comparison of dimensionality reduction algorithms (pca, kpca, lda, mds, svd, lle, isomap, le, ica, t-sne)," *Computer Science Review*, vol. 40, p. 100378, 2021.

[205] B. Janakiramaiah, G. Kalyani, S. Narayana, and T. B. M. Krishna, "Reducing dimensionality of data using autoencoders," in *Smart Intelligent Computing and Applications*. Springer, 2020, pp. 51–58.

[206] P. P. Ippolito, "Feature Extraction Techniques," https://pierpaolo28.github.io/blog/blog29/, 2019, [Online; accessed 22-June-2021].

[207] S. Theodoridis, "Chapter 2 - Probability and Stochastic Processes," in *Machine Learning (Second Edition)*, S. Theodoridis, Ed. Academic Press, Jan. 2020, pp. 19–65. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128188033000118

[208] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[209] P. Cerda and G. Varoquaux, "Encoding high-cardinality string categorical variables," *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[210] M. H. D. M. Ribeiro and L. dos Santos Coelho, "Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series," *Applied Soft Computing*, vol. 86, p. 105837, 2020.

[211] J. Ke, Z. Zhengxuan, Y. Zhe, F. Yu, B. Tianshu, and Z. Jiankang, "Intelligent islanding detection method for photovoltaic power system based on adaboost algorithm," *IET Generation, Transmission & Distribution*, vol. 14, no. 18, pp. 3630–3640, 2020.

[212] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits and systems magazine*, vol. 6, no. 3, pp. 21–45, 2006.

[213] B. Shinde, S. Wang, P. Dehghanian, and M. Babakmehr, "Real-time detection of critical generators in power systems: A deep learning hcp approach," in *2020 IEEE Texas Power and Energy Conference (TPEC)*. IEEE, 2020, pp. 1–6.

[214] J. D. Novaković, A. Veljović, S. S. Ilić, Ž. Papić, and T. Milica, "Evaluation of classification models in machine learning," *Theory and Applications of Mathematics & Computer Science*, vol. 7, no. 1, pp. 39–46, 2017.

[215] L. Chen and Y. Zhu, "A composite cost-sensitive neural network for imbalanced classification," in *2020 39th Chinese Control Conference (CCC)*. IEEE, 2020, pp. 7264–7268.

[216] H. A. Güvenir and M. Kurtcephe, "Ranking instances by maximizing the area under roc curve," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 10, pp. 2356–2366, 2012.

[217] L. Villalobos-Arias, C. Quesada-López, A. Martínez, and M. Jenkins, "Hyperparameter tuning of classification and regression trees for software effort estimation," in *World Conference on Information Systems and Technologies*. Springer, 2021, pp. 589–598.

[218] P. B. Weerakody, K. W. Wong, G. Wang, and W. Ela, "A review of irregular time series data handling with gated recurrent neural networks," *Neurocomputing*, 2021.

[219] Z. Song, Y. Guo, Y. Wu, and J. Ma, "Short-term traffic speed prediction under different data collection time intervals using a sarima-sdgm hybrid prediction model," *PloS one*, vol. 14, no. 6, p. e0218626, 2019.

[220] J. Guo, Z. Liu, W. Huang, Y. Wei, and J. Cao, "Short-term traffic flow prediction using fuzzy information granulation approach under different time intervals," *IET Intelligent Transport Systems*, vol. 12, no. 2, pp. 143–150, 2018.

[221] "Evolve - energy innovation industry data platform," https://www.ausgrid.com.au/Industry/Our-Research/Evolve, 2021.