American International University-Bangladesh (AIUB)

Fraud Detection for Advanced Metering Infrastructure Using Isolation Forest Algorithm

Mohammad Mushfiq Us Saleheen (19-39282-1)

Asif Mahmud (19-39764-1)

Mohammad Golam Faruk Ovi (19-40306-1)

Nafees Fuad Rahman (19-39600-1)

A Thesis submitted for the degree of Bachelor of Science (BSc)
in Computer Science and Engineering (CSE) at
American International University Bangladesh in August, 2023

Faculty of Science and Technology (FST)

Abstract

The Advanced Metering Infrastructure (AMI) plays a crucial role in modern energy systems by providing the means to monitor and regulate energy consumption. The widespread adoption of smart meters has revolutionized the energy sector, but it has also introduced challenges in securing the data collected by these devices. Protecting the integrity and privacy of the data has become a significant concern, necessitating the development of robust fraud detection mechanisms. This literature proposes an efficient fraud detection system for securing smart meters in the AMI. To that end, we utilize the isolation forest algorithm, which is a powerful machine-learning technique for detecting abnormalities. In the AMI, data is collected from multiple smart meters through a smart communication system using a point-to-multipoint topology. By leveraging the isolation forest algorithm, the proposed system aims to enhance precision, efficacy, and data integrity. By implementing the proposed system, anomalies can be promptly detected, and unauthorized access to the system can be prevented, thereby safeguarding the security and integrity of the data collected from smart meters. To validate the performance of the fraud detection system, it is compared with three other widely used machine learning models: DTC, XGBoost, and RTC. The experimental results indicate that the fraud detection system considering the isolation forest algorithm outperforms the other algorithms.

Declaration by author

This thesis is composed of our original work, and contains no material previously published or written by another person except where due reference has been made in the text. We have clearly stated the contribution of others to our thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, financial support and any other original research work used or reported in our thesis. The content of our thesis is the result of work we have carried out since the commencement of Thesis.

We acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate we have obtained copyright permission from the copyright holder to reproduce material in this thesis and have sought permission from co-authors for any jointly authored works included in the thesis.

. . . . . .
Mohammad Mushfiq Us Saleheen
ID: 19-39282-1
BSc in Computer Science and Engineering
American International University-Bangladesh

. . . . . .
Asif Mahmud
ID: 19-39764-1
BSc in Computer Science and Engineering
American International University-Bangladesh

. . . . . .
Nafees Fuad Rahman
ID: 19-39600-1
BSc in Computer Science and Engineering
American International University-Bangladesh

. . . . . .
Mohammad Golam Faruk Ovi
ID: 19-40306-1
BSc in Computer Science and Engineering
American International University-Bangladesh

Approval

The thesis titled **"Fraud Detection for Advanced Metering Infrastructure Using Isolation Forest Algorithm"** has been submitted to the following respected members of the board of examiners of the department of computer science in partial fulfilment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering on **18th May, 2023** and has been accepted as satisfactory.

. . . . . .
Dr. Md Mehedi Hasan
Assistant Professor & Supervisor
Department of Computer Science
American International University-Bangladesh

. . . . . .
Supta Richard Philip
Lecturer
Department of Computer Science
American International University-Bangladesh

. . . . . .
Dr. Akinul Islam Jony
Associate Professor & Head (UG)
Department of Computer Science
American International University-Bangladesh

. . . . . .
Dr. Md. Abdullah - Al - Jubair
Assistant Professor & Director
Faculty of Science & Technology
American International University-Bangladesh

. . . . . .
Prof. Dr. Dip Nandi
Professor & Associate Dean
Faculty of Science & Technology
American International University-Bangladesh

Publications included in this thesis

No publications included

Submitted manuscripts included in this thesis

No manuscripts submitted for publication

Other publications during candidature

No other publications

Research involving human or animal subjects

No animal or human subjects were involved in this research

Contributions by authors to the thesis

	Mohammad Mushfiq Us Saleheen	**Asif Mahmud**	**Mohammad Golam Faruk Ovi**	**Nafees Fuad Rahman**	**Contribution (%)**
	19-39282-1	*19-39764-1*	*19-40306-1*	*19-39600-1*	
Conceptualisation	25%	25%	25%	25%	100(%)
Data curation	25%	25%	25%	25%	100(%)
Formal analysis	25%	25%	25%	25%	100(%)
Investigation	25%	25%	25%	25%	100(%)
Methodology	25%	25%	25%	25%	100(%)
Validation	25%	25%	25%	25%	100(%)
Theoretical derivations	25%	25%	25%	25%	100(%)
Preparation of figures	25%	25%	25%	25%	100(%)
Writing – original draft	25%	25%	25%	25%	100(%)
Writing – review & editing	25%	25%	25%	25%	100(%)

Acknowledgments

We would like to thank our honorable supervisor, Dr. Md Mehedi Hasan, sir, for exposing us to this fascinating topic and mentoring us. His extensive expertise in this sector, intense attention, patience, and unwavering support enabled us to complete our task. His directions have made a significant contribution to every part of the thesis. Respected parents, honorable instructors, fellow students, and supportive friends are appreciated for sharing their expertise and thoughts that aided in the preparation of this thesis, as are all who participated in various ways throughout his thesis.

Keywords

Advance Metering Infrastructure, smart meter, machine learning, point-to-multi-point topology, Micro grid,

Contents

Abstract ii

Contents **ix**

List of Figures **x**

List of Tables **xi**

List of Abbreviations and Symbols **xiii**

1 Introduction **1**

2 Literature review **5**

3 System Model **9**

3.1 Architecture of AMI 9

3.2 Point to Multipoint Topology 11

3.3 AMI working procedure with MDMS 13

4 Clustering with Isolation Forest **15**

5 Proposed Method **19**

5.1 Data Collection 21

5.2 Data Preprocessing: 23

5.3 Build the model using Isolation Forest algorithm: 29

5.4 Evaluate the model: 30

6 Results Analysis **33**

6.1 Metrics for Evaluating Performance 34

6.2 Results obtained using the suggested Method 36

6.3 A Comparison of the Performance 41

7 Conclusion **43**

Bibliography **45**

List of Figures

3.1 Smart Communication System 10
3.2 Point to Multipoint Topology 12
3.3 Meter Data Management System with Advanced Metering Infrastructure 14

4.1 Isolation Forest Algorithm diagram 17

5.1 Model evaluation steps 20
5.2 Feature Engineering Heat-map 25

6.1 Performance Evaluation Results Heatmaps 35
6.2 Performance Evaluation Bar Graphs for Different Test Percentages 37
6.3 Performance Evaluation Line Charts for Different Test Sets 38
6.4 Classification Algorithm Performance By Test Set Size 39

List of Tables

6.1 Confusion matrix applied in smart meter theft detection. 36
6.2 Results of considering different performance evaluations with different test sets. 40

List of Abbreviations and Symbols

Abbreviations	
AMI	The Advanced Metering Infrastructure
HAN	Home Area Network
MDMS	Meter Data Management System
DoS	Denial of Service
KNN	K-Nearest Neighbors
EMD	Empirical Mode Decomposition
MAD	Median Absolute Deviation
ABE	Attribute-Based encryption
HEMS	Home Energy Management System
IoT	Internet of Things
FAN	Family Access Network
WAN	Wide Area Network
LAN	Local Area Network
API	Application Programming Interface
AI	Artificial Intelligence
CPU	Central Processing Unit
RAM	Random-Access Memory
FPR	False Positive Rate
RFC	Random Forest Classifier
DTC	Decision Tree Classifier
XGBOOST	Extreme Gradient Boosting
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
etc.	*etc.*

Symbols	
$\hat{\rho}$	Density operator
$\circledast$	Convolution
etc.	*etc.*

Chapter 1

Introduction

The Advanced Metering Infrastructure (AMI) has emerged as a critical component of modern energy systems that enables utilities to monitor and regulate energy consumption in real-time [Halle and Shiyamala, 2021]. An advanced and contemporary electrical power system known as a ”smart grid” makes use of digital communication technologies, real-time monitoring, and intelligent control to improve the efficiency, dependability, and sustainability of energy production, distribution, and consumption. A smart grid facilitates bidirectional communication and information exchange between various components within the grid, in contrast to conventional power grids, which operate in a one-way flow of electricity from power plants to customers. As smart meters become more popular, there is a growing need to protect the security and integrity of the data collected by these devices.

A microgrid is a small-scale, autonomous energy system that can function on its own or in cooperation with the national power grid. It is made up of sophisticated control systems, batteries, and distributed energy sources like solar cells, wind turbines, and backup generators. Local electricity generation, storage, and distribution are all capabilities of microgrids, which provide a number of advantages including enhanced energy resilience, decreased dependency on the main grid, and the capacity to function independently during grid disruptions. To provide continuous power supply and energy optimisation, microgrids are frequently installed in remote or unreliable grid access locations as well as in essential buildings like hospitals, military sites, and industrial complexes.

An Advanced Metering Infrastructure (AMI) system uses a metre data management system (MDMS) to manage, process, and analyse the enormous volumes of data generated by smart metres and other metering devices. MDMS acts as a focal point for gathering, archiving, and validating metre data, assuring its dependability and accuracy. It offers features including data aggregation, validation, estimating, and editing (AVEE), as well as reporting and data visualisation. Utility companies need MDMS to effectively track and manage energy use patterns, billing, and customer services. Additionally, it makes it easier to integrate different data sources, which empowers utilities to decide wisely about energy distribution, load control, and consumer engagement. Additionally, MDMS improves security by guaranteeing the confidentiality and integrity of metre data and by facilitating compliance with legal standards for data management and consumer privacy.

In comparison to conventional analogue metres, a smart metre uses digital technology to measure and record energy usage for gas, electricity, and water. For the purposes of billing, monitoring, and management, it offers real-time data on energy usage and transmits this information remotely to utility providers. Smart metres have advantages like precise and automated metre readings, real-time energy usage tracking for consumers, and support for demand-response programmes for effective energy use. Utility companies can use these tools to swiftly identify and resolve problems like outages or unusual consumption trends. Smart metres also aid in the modernisation of energy systems by being an essential component of Advanced Metering Infrastructure (AMI), facilitating the development of smarter grids, and encouraging energy efficiency. Fraud detection is an essential part of smart meter security because it can detect irregularities and prevent unauthorized system access [Shehzad et al., 2021]. Smart meter theft is a significant problem that poses various challenges and concerns in the field of energy distribution and metering. It involves the unauthorized manipulation or tampering of smart meters by individuals to avoid paying for their actual electricity consumption. According to a study by [Mendel, 2017], smart meter theft has been a persistent issue in many regions. The research highlights that theft occurs at different points in the network, including the connection between the smart meter and the utility company, as well as within the Home Area Network (HAN) where the meter is located. This highlights the vulnerability of smart meters to unauthorized access and tampering, leading to inaccurate billing and revenue loss for energy providers. One of the primary concerns with smart meter theft is the financial impact on utility companies. The study by [Anand et al., 2021] explains that meter tampering and energy theft result in significant revenue deficits for energy providers. Consumers who manipulate meters or bypass them consume electricity without proper payment, leading to financial losses that affect the sustainability of energy operations and the ability to provide reliable services. In addition to financial losses, smart meter theft also has implications for accurate billing and fair distribution of energy resources. According to research by [Khattak et al., 2019], tampered meters provide incorrect readings, leading to billing discrepancies for both honest consumers and energy providers. This undermines the fairness and transparency of the billing process, creating confusion and dissatisfaction among customers. Furthermore, smart meter theft raises security and privacy concerns. A study by [Al-Fuqaha et al., 2015] emphasizes that manipulated meters may allow unauthorized access to private information or create vulnerabilities in the system that could be exploited for malicious purposes. This highlights the need for robust security measures to prevent tampering and protect the integrity of data exchanged between smart meters and utility providers. Addressing the issue of smart meter theft requires a comprehensive approach. The research also suggests the implementation of advanced security mechanisms, such as encryption and intrusion detection systems, to prevent unauthorized access and tampering. Regular inspections and audits of metering systems, as recommended by [Patruni et al., 2022], can help identify and address potential vulnerabilities. Additionally, raising consumer awareness about the consequences of meter theft, as discussed, can foster a culture of ethical energy consumption and contribute to mitigating this problem.

The importance of security in AMI cannot be overstated. The data collected by AMI systems includes sensitive customer information, such as personal identifying information, energy usage

patterns, and payment information. This data can be highly valuable to attackers who seek to exploit it for financial gain or other malicious purposes. Additionally, AMI systems are vulnerable to cyber-attacks that can disrupt their functionality, causing service outages and potentially putting public safety at risk [Iorga and Shorter, 2015]. Security threats to AMI systems come in many forms. One of the most common threats is physical attacks on the network infrastructure. Attackers may physically access the network equipment or tamper with the meters themselves to intercept data or manipulate readings. Other threats to AMI security include malware, phishing attacks, social engineering, and denial-of-service attacks [Mishra and Tiwari, 2020]. To mitigate these risks, AMI systems must be designed with strong security controls from the ground up. Security should be integrated into the system architecture, with secure coding practices, encryption, and secure data storage and transmission protocols. To guarantee that only authorized people may access the system, access restrictions and authentication measures should be installed. Vulnerability assessments and penetration testing should be performed on a regular basis to discover and correct any security flaws. [Eder-Neuhauser et al., 2016] Furthermore, it is essential to establish robust governance and oversight frameworks to manage AMI security effectively. This includes implementing policies and procedures that address security risks, establishing incident response plans, and ensuring that all stakeholders are aware of their roles and responsibilities for maintaining AMI security. Regulatory compliance requirements must also be considered, as many countries have specific regulations governing AMI security [Toftegaard et al., 2022].

Therefore, security is essential to the successful implementation and operation of AMI systems. With the increasing number of data breaches and cyber-attacks, energy companies must take proactive steps to protect customer data and ensure the continued functionality of their AMI systems. By implementing robust security controls and governance frameworks, energy companies can mitigate these risks and provide safe, reliable, and secure services to their customers.

In conclusion, security is essential to the successful implementation and operation of AMI systems. With the increasing number of data breaches and cyber attacks, energy companies must take proactive steps to protect customer data and ensure the continued functionality of their AMI systems. By implementing robust security controls and governance frameworks, energy companies can mitigate these risks and provide safe, reliable, and secure services to their customers.

In this research, we propose an efficient fraud detection system for AMI smart meter security based on the isolation forest algorithm [Liu et al., 2008]. Our focus will be on network design and topology, particularly point-to-multipoint topology, as well as the smart communication system between the smart meter and the Meter Data Management System (MDMS) [Ford et al., 2014]. We chose the point-to-multipoint architecture for our proposed fraud detection system after carefully considering the various possibilities.

Chapter 2

Literature review

In recent years, there has been a heightened focus on the safety of smart meters used in contemporary metering infrastructure, in particular Advanced Metering Infrastructure (AMI) systems. Dishonest customers who are trying to avoid paying for their electricity use have found a new way to do so: by manipulating these gadgets, which play a critical role in monitoring energy use and enabling interaction among utility companies and customers. According to [MarketScreener, 2023], a significant occurrence happened in 2022 that illustrates the possible repercussions of meter tampering. As a result of lax data protection, the personal information of 51 million Americans has been compromised in the healthcare sector. This incident highlights the critical need of implementing stringent security policies to protect the confidentiality and precision of meter data from manipulation and unauthorized access. Smart meters include security flaws that dishonest customers might use to manipulate meter readings, costing utilities money. It is crucial to create and deploy efficient security measures and fraud detection techniques to protect the stability and sustainability of the power grid from these threats. By strengthening the safety of smart meters, utilities can secure their income and provide consumers' confidence in their bills' accuracy. A resilient and environmentally friendly energy infrastructure is supported by smart meters that are both easy to use and secure from outside threats.

Since 2013, as mentioned in reference [Raciti and Nadjm-Tehrani, 2013], researchers have been working to improve smart meter security and identify abnormalities brought on by physical-cyber intrusions. Raciti and Nadjm-Tehrani explore the use of embedded technology and cyber-physical systems in smart meter anomaly detection in their research. This book gives a thorough introduction to the use of such technologies in the detection of theft and tampering. The use of machine learning methods, and more specifically unsupervised learning, for anomaly identification in smart meters is one of the primary topics investigated by the study. Researchers want to use these techniques to create algorithms that can spot abnormalities in meter readings that might indicate a breach in security. The research of Raciti and Nadjm-Tehrani on integrated cyber-physical anomaly detection in the context of smart meters is quite helpful. Their findings provide a solid basis for addressing the challenges of safeguarding smart meter systems and introducing reliable anomaly detection procedures. Energy companies may improve the reliability and safety of their smart meter networks by implementing

the lessons learned from this study. Safeguarding the integrity of energy consumption data, which is essential for assuring fair billing and dependable functioning of the smart grid infrastructure, may be achieved by the timely detection and response to abnormal activity.

According to [Kalogridis et al., 2013], in 2014, a uniform architecture was developed to handle privacy and security issues in smart meter networks. Anonymization and differential privacy were two privacy protection methods that this framework sought to combine with other security features including encryption, access control, and intrusion detection. By bringing these methods together, the framework aims to provide a holistic method of protecting the private information gathered by smart meters. Data sharing, security, and traffic analysis were at the forefront of efforts to create an open-source smart meter platform in 2018. This work was highlighted in [Caropreso et al., 2018]. For effective sharing of smart meter data, the research presented an open-source solution based on a message broker architecture. The stability and scalability of the system were guaranteed by this platform's real-time traffic analysis capabilities. Potential security issues might be recognized quickly via traffic analysis, improving the smart meter network's security overall. Smart meter network security and privacy have benefited from the research described in references [Kalogridis et al., 2013] and [Caropreso et al., 2018]. By bringing together privacy-enhancing measures and security approaches, the unified framework suggested in [Kalogridis et al., 2013] aims to provide a secure setting for smart meter data. In contrast, the work discussed in [Caropreso et al., 2018] focuses on creating an open-source platform that facilitates secure data exchange and allows for real-time analysis. The research presented here is indicative of the continuous work being done to resolve security and privacy issues in smart meter networks. Energy suppliers and stakeholders may better safeguard sensitive data, strengthen network security, and encourage trustworthy smart grid operations by adopting the recommended frameworks and platforms.

Smart grid metering network privacy and security has been the subject of extensive study in 2019. One excellent research [Kumar et al., 2019] analysed all the potential dangers of such networks in detail. In addition to delving into privacy and security concerns, the report also looks at potential threats to smart grid metering network infrastructure. Replay attacks, DoS attacks, and metre manipulation were only some of the methods used. The study's primary goals were to increase public understanding of these threats and to encourage more study into their causes and possible remedies. Related to the topic of smart metre security in the context of AMI is another study work cited as [Khattak et al., 2019]. This research looked at the downsides and dangers of smart metres and suggested solutions to these problems. The effects of various assaults, such as tampering with metres, on energy providers and end users were underlined. The essay emphasised the need of secure communication pathways and encryption mechanisms to ensure the privacy and security of data transmitted between smart metres and energy suppliers. The author also suggested using machine learning and intrusion detection technologies to spot and stop any fraudulent activities.

New this year are a slew of studies that advance our knowledge of and methods for bolstering smart grid metering network security. Researchers, industry experts, and politicians may all learn from the in-depth investigation of privacy and security threats and the suggested responses. Stakeholders may

improve the robustness and security of smart metres by introducing secure communication protocols, encryption methods, and cutting-edge detection systems. These studies provide the groundwork for future investigations and innovations related to the safety of smart metres. Smart metre research grew significantly in 2020 due to modernization. Smart grid management requires smart metre security [Marah et al., 2020]. Smart metres may leak electricity and sensitive data, according to the authors. AMI transmits important data, hence the authors stressed the necessity for strong security. They suggested addressing these security issues using machine learning.

In 2020, [Aziz et al., 2020] released a new power theft detection method. KNN and EMD were used in the suggested technique. EMD was used to decompose smart metre load profile data into IMFs. KNN analysed the IMFs to determine energy use trends. The suggested method detected energy theft accurately and efficiently using actual data from a Pakistani distribution firm. 2020 study showed the growing importance of smart metre security in smart grid management. Security was highlighted by smart metre weaknesses and possible attackers. Machine learning, empirical mode decomposition, and KNN might improve energy theft detection and prevention. These advances help contemporary energy systems operate reliably and efficiently by ensuring smart metre data integrity, privacy, and security. In 2022, two studies were conducted on the topic of smart grid technology. One study [Prabhakar et al., 2022] looked at the security of the smart metering system and recommended the Median Absolute Deviation (MAD) method for detecting and stopping cyberattacks. The second study [Abdalzaher et al., 2022] proposed a strategy to smart metering data privacy and security based on attribute-based encryption (ABE) and secure data aggregation. Both research highlighted the need of taking necessary security precautions for smart grid technologies to prevent cyberattacks and preserve data privacy. New concerns concerning privacy in smart metering systems are highlighted in a 2023 article [Polčák, 2023]. It highlights the need of protecting sensitive information and discusses privacy-enhancing tools including encryption, pseudonymization, and differential privacy.

Chapter 3

System Model

3.1 Architecture of AMI

The fraud detection system may accomplish efficient and accurate anomaly detection by carefully choosing and executing the right network architecture, including the point-to-multipoint topology and a strong smart communication system. The network design is a critical component of the overall architecture since it directly affects the efficiency and dependability of the system. By taking these factors into account, an effective and reliable AMI system can be created, guaranteeing the security and accuracy of the data gathered from smart meters.

The Home Energy Management System (HEMS) and various Internet of Things (IoT) devices make up the Home Area Network (HAN) in this network design. The HEMS is crucial in controlling energy use at home and increasing efficiency. Each residence in the HAN is represented by its own smart meter. Smart meters are installed in houses and are used to monitor and record energy consumption. Next, the HAN is linked to either the FAN or the WAN. Smart meters may talk to one another over longer distances when connected to a wide area network (WAN), as opposed to a local area network (LAN). To ensure reliable transmission of energy consumption data from the HAN to the LAN or WAN, smart meters act as a connecting link between the two networks.

A meter data management systems (MDMSs), are vital to the infrastructure of modern networks. They are made up of numerous modules that work together to optimise power grids, manage service providers' utility expenditures, and improve customers' experiences. Customers, for instance, can use individualized websites to keep tabs on and control their electrical consumption, while utilities can make better use of consumption data for in-depth research and strategic planning.

Smart meters in the home area network (HAN) are connected to the provider's meter data management system (MDMS) via a communication mechanism. As shown in Fig. 3.1, this communication system acts as a data collector within a point-to-multipoint topology. It makes it easier for smart meters to transmit energy consumption data to the MDMS, which in turn improves the speed and accuracy of data collection, analysis, and application.

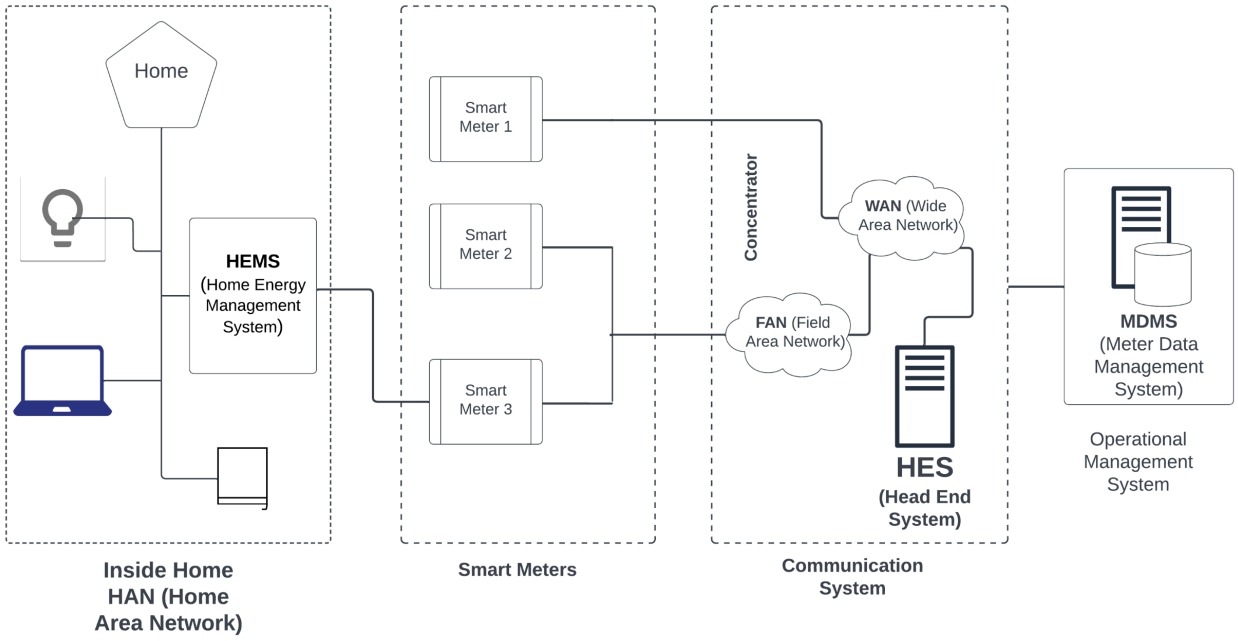


Figure 3.1: Smart Communication System

Therefore, Household energy consumption can be managed and optimized, and communication and data exchange between consumers, utilities, and service providers can run smoothly. By bolstering the smart metering system's general functionality and efficiency, this design makes possible cutting-edge energy management and encourages better energy practices on the part of both individuals and communities.

3.2 Point to Multipoint Topology

In this research, we propose an efficient fraud detection system for AMI smart meter security based on the isolation forest algorithm [Liu et al., 2008]. Our focus will be on network design and topology, particularly point-to-multipoint topology, as well as the smart communication system between the smart meter and the Meter Data Management System (MDMS) [Ford et al., 2014]. We chose the point-to-multipoint architecture for our proposed fraud detection system after carefully considering the various possibilities.

Firstly, the dedicated communication lines used in this topology can help ensure data integrity by reducing the risk of data loss, corruption, or manipulation that may occur with shared communication paths [6]. This is because each connection between the central hub and the endpoints is exclusive and independent of other connections, minimizing the chances of interference or interception from third parties. Moreover, the point-to-multipoint topology's streamlined data transport and reduced noise can significantly improve the precision and efficacy of the isolation forest method employed in your fraud detection system [6]. The isolation forest algorithm uses decision trees to identify anomalies in data, and the accuracy of its predictions depends heavily on the quality and quantity of the input data. By minimizing noise and optimizing data transport, you can enhance the model's performance and reduce the number of false positives or false negatives. Another significant advantage of using a point-to-multipoint architecture for your fraud detection system is that it allows for better control and management of data collection. As you mentioned, you plan to use the smart communication system as a data collector. By doing so, you can ensure that all data collected is clean and accurate, limiting the possibility of false positives and false negatives caused by noisy data [6]. Additionally, with a dedicated communication line, you can monitor and manage data flow more effectively, pinpointing any potential issues and addressing them promptly. The point-to-multipoint architecture is an excellent choice for your proposed fraud detection system due to its many advantages over other topologies. Its dedicated communication lines enhance data integrity, streamline data transport, and reduce noise, improving the efficacy of the isolation forest algorithm. Furthermore, it provides better control and management of data collection, ensuring that all data collected is clean and accurate.

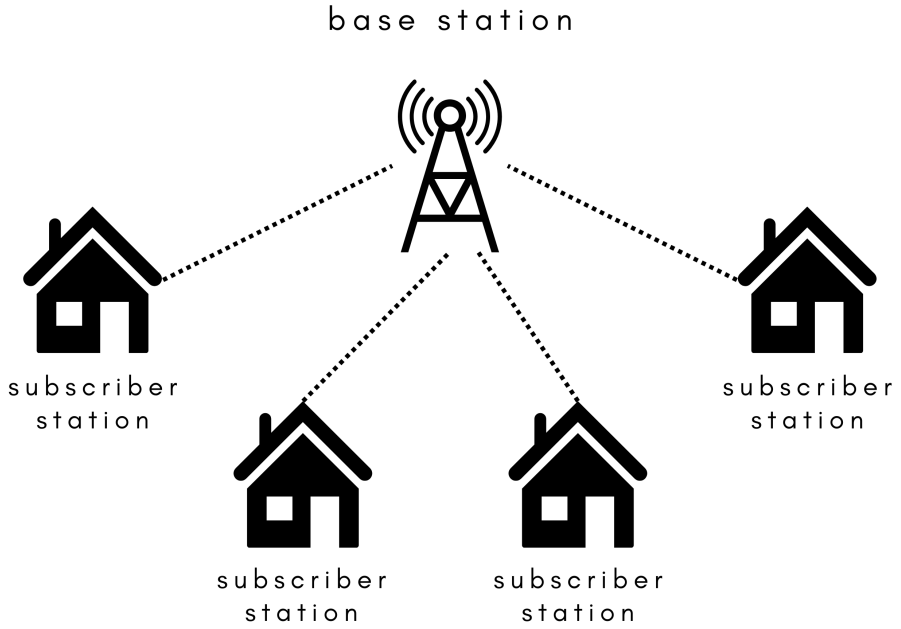


Figure 3.2: Point to Multipoint Topology

3.3 AMI working procedure with MDMS

Smart meters, unlike traditional meters, can collect and transmit data about energy consumption in real-time. This data is then sent to the meter data management system (MDMS), which is responsible for processing, storing, and analyzing the data. The AMI system works by using a combination of technologies such as wireless mesh networks, power-line communications, cellular networks, and Wi-Fi to transmit data from smart meters to the MDMS [Shehzad et al., 2021]. The communication network created by these technologies ensures that the data is transmitted securely and reliably. Once the data has been collected by the smart meters, it is transmitted to the MDMS in either real-time or in batches depending on the network type used. The MDMS then stores the data in a database where it can be accessed by the utility company for billing, load forecasting, outage management, and other purposes [Liu et al., 2008].

To maintain the correctness and integrity of the data, the MDMS also conducts activities such as data validation, estimate, and editing. For instance, when there is a missing or incomplete data point, the MDMS may use algorithms to estimate the missing value based on past consumption patterns [Ford et al., 2014]. One advantage of combining the AMI system and the MDMS is the ability to build demand response programs. Demand response systems help utilities lower peak demand by rewarding consumers to reduce their energy consumption during peak periods. Utilities can use the AMI system to transmit signals to smart meters to alter customers' energy usage during these times [Bentéjac et al., 2019]. Finally, the AMI system collaborates with the MDMS to provide utilities with precise, real-time data on the energy use of their consumers. This data allows utilities to make educated decisions about energy generation, distribution, and consumption. Utilities may enhance their operations, cut expenses, and encourage energy conservation with the help of the AMI system and the MDMS. The Advanced Metering Infrastructure (AMI) has emerged as a critical component of modern energy systems that enables utilities to monitor and regulate energy consumption in real-time [1].

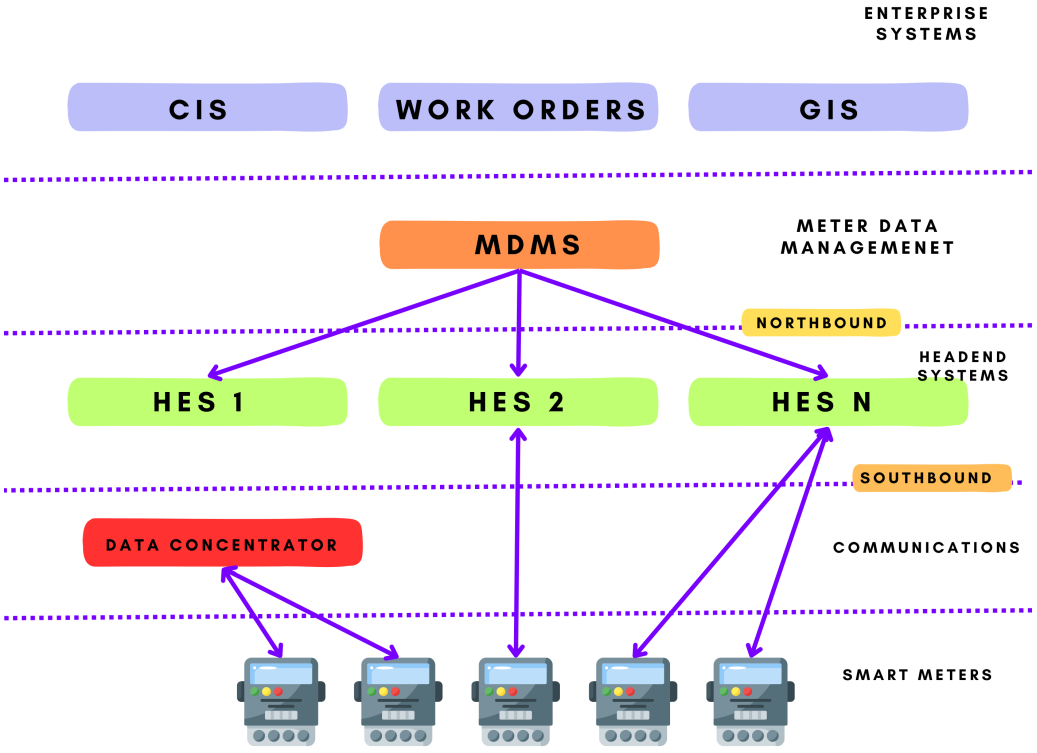


Figure 3.3: Meter Data Management System with Advanced Metering Infrastructure

Chapter 4

Clustering with Isolation Forest

When it comes to finding outliers and anomalies in a dataset, the Isolation Forest algorithm is a powerful unsupervised machine learning tool. Anomalies are distinguished as instances that are isolated or stand out from the majority of data. To execute the Isolation Forest algorithm, you must first prepare the dataset by ensuring that it is in an appropriate format and removing any missing values or irrelevant features. Then, you import the essential libraries or packages, such as scikit-learn in Python, that provide the algorithm's implementation. Next, you choose the features that will be utilized for anomaly detection. This can be accomplished based on domain expertise, data exploration, or the nature of the issue at hand. By contemplating the most informative characteristics, you improve the algorithm's ability to accurately identify anomalies. Once the data and features have been prepared, the Isolation Forest model can be trained on the training set. The algorithm constructs an ensemble of isolation trees, which are binary trees that partition the data recursively by selecting at each node a random feature and split value. The objective of the isolation trees is to isolate anomalies in smaller partitions, whereas normal instances are anticipated to require more partitions. Once the model has been trained, it can be applied to the testing set or to new, unobserved data for anomaly detection. Each instance is assigned an anomaly score by the Isolation Forest algorithm, which represents the likelihood that it is an aberration. As they were isolated faster during the partitioning procedure, instances with lower scores are more likely to be anomalies. To classify instances as anomalies or normal data points, you set an anomaly score threshold. This threshold can be established using domain expertise, statistical analysis, or an examination of the distribution of anomaly scores. Instances with scores above the threshold are considered normal, whereas those with scores below the threshold are considered anomalous. These metrics offer insight into the algorithm's ability to identify anomalies with precision. The results of the Isolation Forest algorithm must be interpreted by analyzing the detected anomalies within the context of the problem domain. Understanding the fundamental causes of their abnormality can provide valuable insights for additional analysis or decision-making. In this section, we will go through the fundamentals of the Isolation Forest algorithm. To identify anomalies, the data is often passed through a random-generated forest of isolation trees. A subset of the investigated data set is used to develop each tree individually. Isolating trees are so named because they subdivide the

study area over and over again, until either all of the areas of interest fit within the tree's leaves or the depth limit is reached, whichever comes first 4.1. In order to discover outliers, data is submitted to a developing forest, which then calculates an anomaly measure for each data point. The formula is used to compute the value.

$$s(x,n) = 2^{(-E(h(x)))/(C(n)))} \tag{4.1}$$

Where

$$C(n) = 2H(n-1) - (2(n-1))/n \tag{4.2}$$

$$H(n) = ln(n) + 0.57772156649 \tag{4.3}$$

The E(x) term represents the average lengths of all the pathways that must be traversed in the isolating trees in the aforementioned equations 4.1, 4.2, 4.3. The theoretical average path length of a failed search in a binary search tree is denoted by the constant C(n).

The publication can be consulted by a reader who is particularly interested because it contains the original Isolation Forest building procedure and the formula for determining the anomalous score. However, the addition in this study focuses on the algorithm for creating an isolation tree. As a result, we remember it as pseudocode; for more information,

Algorithm 1 : iTree(X,e,l) Inputs: X – input data, e – current tree height, l – height limit

Output: iTree

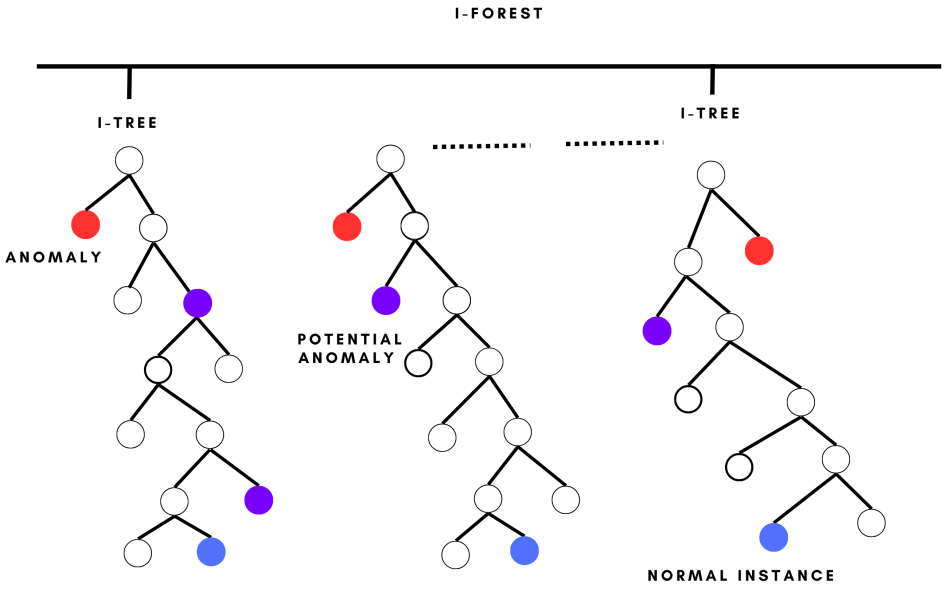


Figure 4.1: Isolation Forest Algorithm diagram

Chapter 5

Proposed Method

This thesis aims to develop and evaluate a machine learning-based methodology that can be applied to a wide range of problems and domains, leveraging diverse data sources and advanced machine learning techniques. The methodology will provide valuable insights, accurate predictions, and effective decision-making capabilities, contributing to the advancement of machine learning research and offering practical applications across various industries and sectors.

A machine learning approach called isolation forest excels in finding outliers or abnormalities in a dataset. Given that it doesn't require prior knowledge of typical or aberrant behavior, it is especially well suited for tasks involving anomaly identification. Instead, it focuses on separating anomalous observations from the rest of the data by repeatedly splitting it. You can use isolation forests to analyze a variety of energy-related data, including power usage, voltage levels, temperature, and other pertinent characteristics. During the training phase, the algorithm discovers the typical patterns seen in the data, building a model that encapsulates the fundamental structure of the typical behavior. Then, during the testing or deployment phase, this model is used to find deviations from the learnt patterns.

The isolation forest algorithm chooses a feature and a random split value within the feature's range at random. The data is then divided into groups according to whether the feature value is higher or lower than the split value. The result is a binary tree structure with shorter average path lengths for anomalies. This method is done recursively. Each observation's average path length can be calculated, and anomalies can be distinguished by their notably shorter lengths. Utilizing isolation forests to detect energy anomalies has a number of benefits, one of which is their effectiveness at handling high-dimensional data. Huge volumes of data are produced by energy systems frequently containing many different variables that interact in intricate ways. Even with its complexity, isolation forest can handle anomalies even in high-dimensional feature spaces.

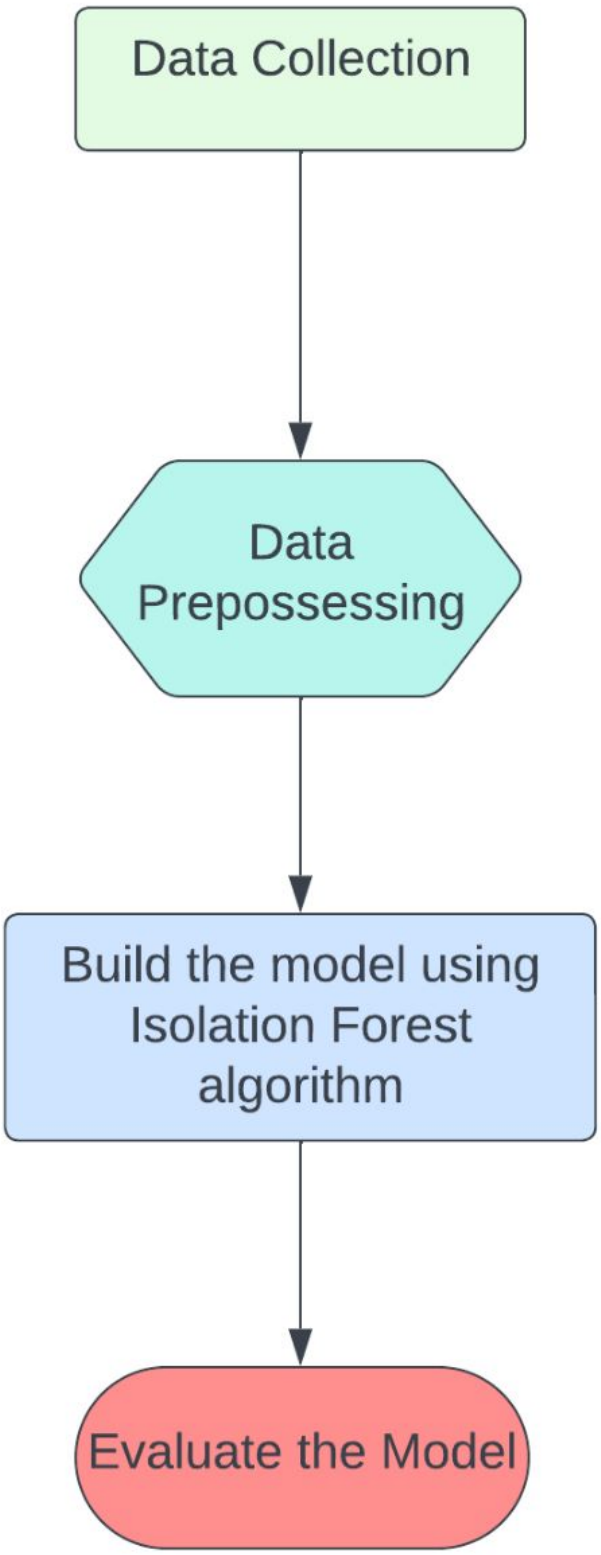


Figure 5.1: Model evaluation steps

5.1 Data Collection

The fundamental process that underpins the development and efficacy of algorithms by supplying the necessary training and validation datasets, data collecting is at the core of machine learning. The performance, fairness, and generalizability of machine learning models are all significantly impacted by the careful selection, acquisition, preprocessing, and curation of the data as well as the consideration of various factors like data quality, representativeness, bias, and privacy. In order to ensure the dependability, robustness, and ethical implications of machine learning systems as well as to advance the responsible and equitable deployment of artificial intelligence technologies in a rapidly evolving and data-driven world, it is crucial to understand the complexities and challenges associated with data collection.

In the discipline of machine learning, which tries to create algorithms capable of automatically understanding patterns and making predictions or choices from data, data collecting is crucial. The caliber and volume of the data used for training strongly influences how well these algorithms perform. The identification and definition of the current problem is the first step in the process of data collection. The goals of a machine learning task must be clearly defined by researchers or practitioners, together with the precise data requirements required to meet those goals.

The next step in data collecting is to choose the right sources from which to collect the required data after the problem has been established. These sources can change based on the nature of the issue and the accessibility of pertinent information. Public datasets, private databases, sensor networks, social media platforms, and other internet sources are typical sources. To make sure the information gathered is in line with the goals of the issue, it is critical to assess the validity, accuracy, and relevance of these sources. The process of data collection entails obtaining the chosen data from the listed sources. This procedure may entail a number of different strategies, including web scraping, APIs, data feeds, surveys, tests, or partnerships with other businesses. The integrity and legality of data collection techniques must be carefully considered, with respect for copyright regulations, terms of service, and user privacy agreements.

Preprocessing the data once it has been obtained is an essential step in the data collection process. To prepare the raw data for analysis and model training, this stage entails cleaning, converting, and organizing it. Handling missing values, eliminating outliers, normalizing or standardizing variables, and encoding categorical variables are examples of frequent preprocessing jobs. In order to minimize dimensionality and enhance model performance, feature selection or extraction techniques may also be used during data preprocessing.

Another crucial component of data acquisition for machine learning is data curation. To give useful context and metadata, it involves organizing and annotating the gathered data. For supervised learning tasks, accurate annotation and labelling of data, such as class labels or target variables, is essential. Curation also entails protecting sensitive data's privacy and security, especially when working with private or confidential data. To defend people's right to privacy anonymity. techniques or data protection measures may be required. The acquisition of data for machine learning is fundamentally concerned

with data quality. The performance and dependability of the generated models are significantly impacted by the correctness, completeness, and consistency of the acquired data. Unreliable data, such as noisy or biased data, can produce inaccurate forecasts, deceptive insights, or unfair results. Techniques for assessing and ensuring the quality of the data, such as data audits, validation, and verification, should be used to spot and address potential problems that can negatively affect the machine learning process. Another important factor is the representativeness of the data that was obtained. To ensure generalizability, the data used to train machine learning models should accurately reflect the target population or the issue domain. When particular groups are overrepresented in the data or when there is a lack of coverage of all possible scenarios or cases, biases might occur. To solve these issues and advance fairness in machine learning applications, bias identification, and mitigation strategies, such as oversampling underrepresented groups or carefully balancing datasets, are essential. Data collecting for machine learning is associated with a substantial ethical concern regarding privacy. As a result of the massive collection of personal data and model training, it becomes crucial to protect people's privacy. To protect people's rights and stop unauthorized access to or exploitation of personal information, compliance with data protection laws, informed consent, and data anonymity techniques are crucial. Building trust and making sure machine learning technologies are used responsibly both depend on privacy-aware data collection practices. For academics, practitioners, and policymakers alike, it is crucial to comprehend the intricacies and difficulties of data collecting in machine learning. It necessitates a multidisciplinary approach that brings together knowledge of data management, ethics, and legal and ethical frameworks. Creating standards and recommended practices for data collection in machine learning can encourage openness, responsibility, and the ethical application of AI technologies. In addition, encouraging collaborations between regulatory agencies, business, and academia can result in the creation of strong standards and frameworks that address the sociological, ethical, and legal consequences of data collection for machine learning. Due to a number of reasons, such as the complexity and volume of data, as well as problems with data accessibility and quality, the collecting and analysis of energy data has proven to be difficult. However, we are investigating the use of publicly accessible datasets to alleviate these issues. The Energy Anomaly Detection competition dataset, which is available on Kaggle, is one such dataset that has piqued our curiosity. This dataset offers a useful resource because it includes a significant quantity of data on the energy use of about 200 commercial buildings. It provides a variety of information including energy use, meter-readings, square-foot, temperature, humidity, and other pertinent factors. With a total sample count of 1,048,575 in the dataset, we are able to do a thorough analysis.

Our research study intends to get over the difficulties usually encountered when collecting and analyzing energy statistics by making use of this dataset. This dataset's accessibility and depth provide chances to investigate energy patterns, spot abnormalities, and learn important lessons about how energy is used in commercial buildings.

5.2 Data Preprocessing:

The process of turning raw data into meaningful and representative features that can accurately capture the underlying patterns and relationships in the data is known as feature engineering, and it is a critical stage in the machine learning process. To extract pertinent information and improve the predictive capability of machine learning models, this procedure calls for a combination of domain expertise, creativity, and data exploration techniques.

Feature Engineering: Feature engineering is essential for enhancing model performance, lowering overfitting, and enabling precise predictions by carefully designing and choosing features. A detailed understanding of the issue domain and the available data is the first step in feature engineering. It entails looking at the traits, attributes, and probable connections between the variables in the dataset. In order to increase the model's capacity for generalization, the exploration phase assists in the identification of pertinent features and offers insights into possible feature combinations and transformations. Handling missing data is a popular feature engineering strategy. For a variety of causes, missing values might happen, and their existence can impair model performance. Different approaches, such as imputation techniques that fill in missing values based on statistical measures or sophisticated algorithms, can be used depending on the type and quantity of missing data. Instead, lacking indicators can be developed to clearly record the absence of data, which can give the model important data. Scaling of features is a crucial component of feature engineering. It entails adjusting traits to make them comparable by normalizing or standardizing them. When using machine learning methods that are sensitive to the size of the features, scaling is very important. According to the distribution and features of the data, common scaling methods include z-score normalization, min-max scaling, or resilient scaling.

Another typical work in feature engineering is categorizing variables. Categorical variables must be converted into numerical representations since many machine learning techniques demand numerical inputs. Using methods like one-hot encoding, label encoding, or target encoding, categorical variables can be converted into numeric form while maintaining their information. The effectiveness of machine learning models can also be improved via feature modifications. Mathematical operations like logarithmic, exponential, or polynomial transformations can be used to capture non-linear relationships between characteristics and the target variable. These adjustments can aid linear models in detecting more intricate patterns in the data. A valuable addition to feature engineering is interaction features. Interaction words are able to represent additive and synergistic effects by combining two or more existing properties. These interactions, which can be produced by mathematical operations like multiplication, division, or exponentiation, might offer more details on the underlying patterns in the data. Finding the most pertinent characteristics that considerably increase the predictive capacity of the model is the goal of feature engineering's crucial feature selection process. This procedure aids in reducing dimensionality, removing distracting or superfluous characteristics, and enhancing the generalizability and interpretability of models. Numerous statistical methods, such as correlation analysis, stepwise selection, regularization strategies like L1 or L2 regularization, or sophisticated algorithms, such as genetic algorithms or recursive feature elimination, can be used to choose features.

In feature engineering, inventiveness and subject-matter expertise are crucial. The raw data may occasionally lack the precise characteristics needed for precise forecasts. On the basis of their knowledge of the issue and the underlying systems, domain experts in such situations might deduce new features or produce composite features. These expert-derived features can record important knowledge and past insights that might not be visible in the raw data. Consideration of temporal or sequential information present in the data is another aspect of feature engineering. For instance, trends or autocorrelation are frequently seen in this data. The model's capacity to predict future values more precisely can be increased by transforming the data to capture these temporal patterns, for as by applying lagged features, moving averages, or Fourier transforms.

The most crucial features were found and chosen via feature engineering throughout the data preprocessing step. We initially pinpointed the key qualities by carefully analyzing the data and thoroughly knowing the domain. We created a heatmap of the correlation matrix using data visualization techniques, specifically Seaborn, in order to further analyze the correlations between these features. We could evaluate whether there were any highly associated features in the dataset and decide whether any feature reduction was required thanks to the heatmap. We examined the heatmap and discovered no highly associated features, proving that no more feature reduction was necessary. This procedure made it easier to get a thorough understanding of the dataset and opened the door for later data analysis and modelling.

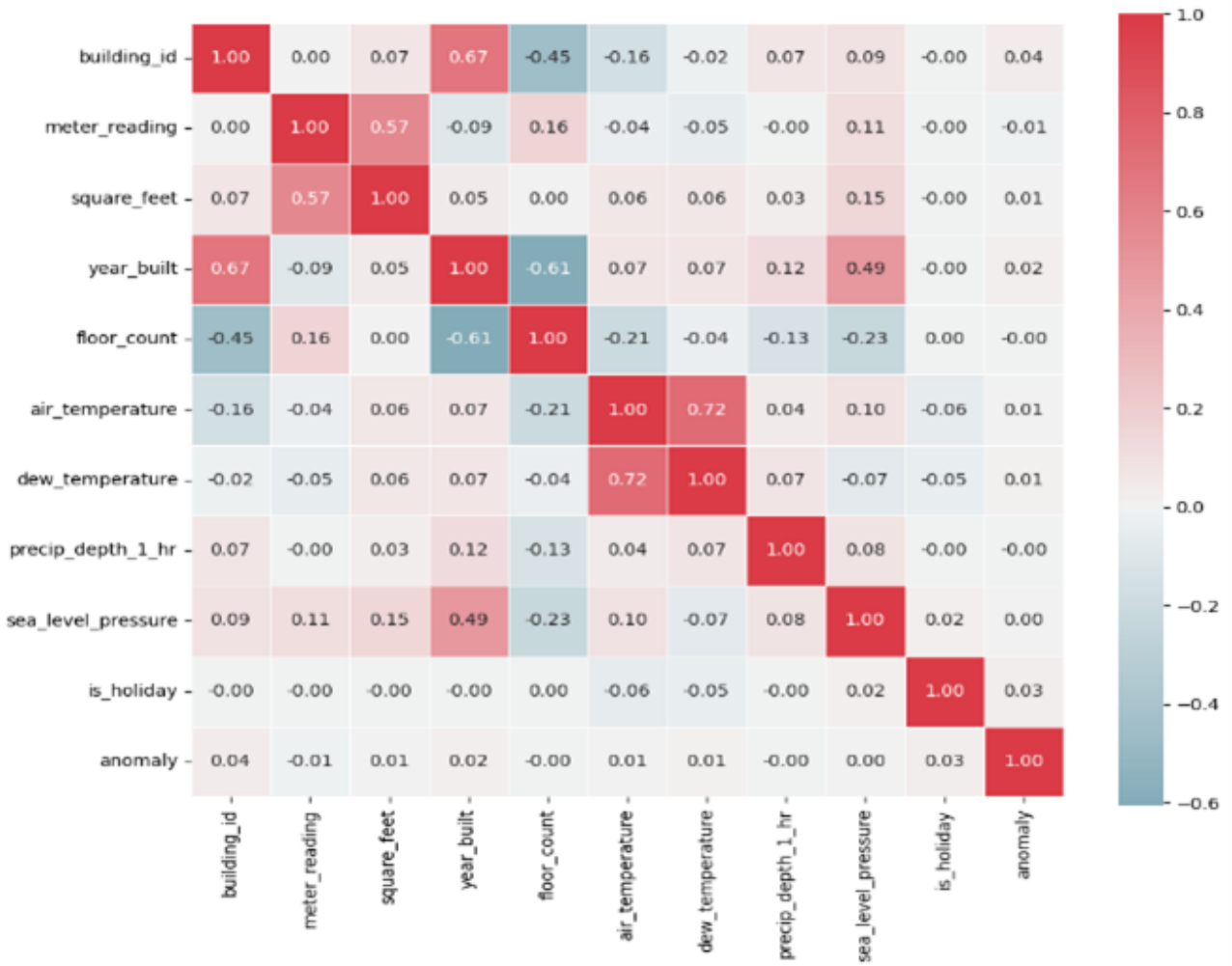


Figure 5.2: Feature Engineering Heat-map

Missing Values Handle: One of the most important steps in the preprocessing of the data for machine learning is handling missing values. Missing values can appear in datasets for a number of reasons, including inaccurate measurement, incorrect data entry, or insufficient data collection. Finding the existence and severity of missing data within the dataset is the first step in addressing missing values. This can be done by looking at the dataset's summary statistics or by utilizing tools like heatmaps or missing value matrices to visualize the patterns of missing data. Formulating effective handling solutions for missing values requires an understanding of the patterns of missing data. Imputation, which substitutes estimated or expected values for missing data, is a popular method for addressing missing values. The three main categories of imputation techniques are mean/mode imputation, regression imputation, and multiple imputation.

1. Mean/mode imputation substitutes the available data's mean (for categorical variables) or mode (for numerical variables) in place of missing values. If the missingness is connected to the target variable or other attributes, the approach may introduce bias because it believes that the absent values are fully random.

2. Regression models are used in regression imputation to forecast missing values based on the existing information. By employing the features with missing values as the target variable and the remaining features as predictors, a regression model is constructed. The missing values are then imputed using the expected values from the regression model. This approach can capture more intricate correlations between variables, but if the regression model is unreliable, it may also increase uncertainty.

3. By repeatedly imputing missing values using an iterative procedure, the multiple imputation technique creates several imputed datasets. Each dataset that is imputed contains random changes that reflect the uncertainty surrounding the missing values. Following imputation, the various datasets are each individually analyses, and the outcomes are then pooled to get a final result. When the missing mechanism is non-ignorable—that is, when the missing is connected to the missing values themselves or to other variables in the dataset—multiple imputation is especially helpful.

Deletion, which entails eliminating instances or variables with missing values from the dataset, is another method for handling missing values in addition to imputation. Pairwise deletion, and feature-wise deletion are the three different categories of deletion techniques.

Complete-case analysis, sometimes referred to as deletion, eliminates entire instances or rows that have at least one missing value. Although this method guarantees entire data for analysis, it may cause a considerable loss of knowledge if the dataset has a lot of missing data.

When doing calculations using such variables, pair deletion keeps the instances of missing values but leaves out the missing values. This strategy enables the use of all accessible data for each computation, but it may cause bias in the findings if the variables utilized in the computations are connected to the missing.Feature-wise deletion eliminates variables or features from the dataset that have missing values. When the missing values are concentrated in a small number of particular variables and the other variables are regarded enough for analysis, this method is appropriate. However, feature-wise deletion could result in the loss of information that might have been present in the excluded

variables. We made the missing values and outliers from the dataset because we were working decision to remove any with a huge dataset, which would ensure the accuracy of the data. Instead of imputing the missing values and outliers, the outcome might not change if the missing values and outliers are dropped.

Data Splitting: Data splitting is a fundamental stage in machine learning involving the separation of a dataset into subsets for training, validation, and testing. The objective of data division is to accurately assess and evaluate the performance of machine learning models, to guarantee correct model generalization, and to prevent overfitting. A typical train-test divide entails allocating between 70 and 80 percent of the data to the training set and the remainder to the testing set. This division ensures that the model learns patterns and associations from a large quantity of data while retaining unseen instances for evaluation. However, relying solely on a train-test split can lead to biased model evaluation and parameter tuning. To address this, a more advanced data splitting strategy called cross-validation is often employed.

In cross-validation, the dataset is divided into multiple subsets, or ”folds,” with each fold functioning as both a training set and a testing set. This method provides a more accurate evaluation of the performance of the model by using the entire dataset for training and testing and aggregating the results across multiple folds. Accurate estimate of the model's performance is obtained by averaging the performance metrics derived from each iteration.Stratified cross-validation is a variant of cross-validation that guarantees the class distribution across folds remains consistent. This is especially beneficial when working with unbalanced datasets in which some classes are underrepresented. Stratified cross-validation prevents bias toward the majority class by conserving proportions of classes in each fold.

In addition to training and testing sets, a validation set is frequently used to fine-tune hyperparameters and assess model performance during training during model development. This is typically accomplished by dividing the data into training, validation, and testing sets using a three-way division. On the basis of performance evaluation, the validation set is used to iteratively modify hyperparameters such as learning rate and regularization strength. On the independent testing set, the performance of the final model is then evaluated to derive an unbiased estimate of its generalization ability. It is crucial to ensure that the data splitting process is random and preserves the underlying distribution and patterns in the data. Random shuffling of the dataset before splitting helps avoid any ordering or grouping biases. Additionally, it is important to avoid data leakage, where information from the testing set inadvertently influences the training process, as this can lead to overly optimistic performance estimates.

By appropriately splitting the data into training, validation, and testing sets, machine learning models can be trained, fine-tuned, and evaluated accurately. This facilitates the selection of optimal models, hyperparameters, and provides a reliable estimate of their performance on unseen data. Data splitting is an essential practice to ensure the robustness and generalization capability of machine learning models in real-world applications. Data splicing is the process of dividing available data into a training set and a testing set. The testing set is used to evaluate the machine learning model's

performance, while the training set is used to educate the model. To facilitate model training and evaluation, we divided the data 70:30 into training and testing sets.

Standardization: A preprocessing method in machine learning called standardization, often referred to as feature scaling or normalization, seeks to scale down numerical features to a uniform scale. It entails scaling the values of various features to make sure that they have comparable ranges and distributions, which can assist machine learning models perform better and remain stable. When attributes are measured in different ways or very much in magnitude, standardization is required. The learning process may be dominated by features with higher values in the absence of standardization, producing biased model results. The relative value of each feature is kept by feature standardization, enhancing the model's capacity for learning. The two most often used standardization techniques are z-score normalization and min-max scaling. 1. Standard score scaling, also known as z-score normalization, modifies the data so that the mean equals zero and the standard deviation equals one. It entails dividing the result after deducting the feature's mean from each data point by its standard deviation. This procedure makes sure that the standardized feature has a spread that can be easily compared to other characteristics and a distribution centered around zero. 2. Min-max scaling, on the other hand, rescales the data to a specific range, typically between 0 and 1. This transformation converts the original feature values to a normalized range to facilitate direct feature comparisons. The choice between z-score normalization and min-max scaling for standardizing features is dependent on the specific needs of the assignment and the quality of the data. The training and testing data are then both transformed using these parameters, maintaining consistency throughout the scaling process. It is crucial to remember that standardization should only be used for numerical features and not for categorical features or features having a predetermined meaning or hierarchy. Additionally, standardization must be carried out separately for each feature and not on the target variable. With the aid of different libraries and frameworks in well-liked programming languages like Python, standardization can be readily applied. Standardization can be easily incorporated into the machine learning pipeline thanks to libraries like scikit-learn that offer pre-built algorithms and classes for scaling data. The performance of machine learning models can be significantly impacted by standardization, particularly those that are sensitive to the scale of the features. It may result in quicker convergence during training, better model interpretability, and less outlier influence. Standardization alone, however, does not ensure better model performance in all cases, and its efficacy may vary based on the particular dataset and modelling methodologies used. Finally, we performed standardization on the numerical variables to normalize their values, ensuring that all variables have the same scale and preventing any variable from dominating the model training. where x is the initial value of the input feature, mean is its mean value in the training set, and standard deviation is its standard deviation. StandardScaler scales the input features have a mean of 0 and a standard deviation of 1 in order to make them more relevant.

5.3 Build the model using Isolation Forest algorithm:

The main tenets of the Isolation Forest algorithm are reviewed in this section. A forest of isolation trees that are constructed randomly is used to find outliers in the data. Each tree is cultivated using a separately collected sample that is a subset of the examined dataset. The trees are known as isolating because they repeatedly divide the space being studied until either all of the distinct points are contained inside their leaves or the depth limit is achieved. By submitting the data to the growing forest and generating an anomaly measure for each data point, outlier detection is carried out. The formula is used to determine the measurement. [Liu et al., 2008]

Random Selection of Subsamples:

1. The Isolation Forest algorithm operates by randomly selecting a subset of the data for building each decision tree in the forest. The subsampling process improves the algorithm's efficiency.

2. Select a random sample (subsample) of size 'm' from the dataset, where 'm' is typically set to a fraction of the total number of instances in the dataset. A common value for 'm' is around 256.

Build Isolation Forest:

1. Choose a feature at random from the dataset.

2. Within the subsample range, choose a split point at random for the specified feature.

3. Based on the selected feature and split point, divide the subsample into two halves.

4. Repeat the process until all data points have been isolated (each data point is a separate leaf node).

5. Each tree's depth is limited to a number similar to the average depth of a binary tree (about 2 * log2(m)).

Path Length Computation: Determine the length of each decision tree's path to its root for each data point in the dataset. The number of edges that were traveled from the root node to a certain data point is represented by the path length. Anomalies are more likely to occur in data points that can be isolated with shorter path lengths.

Analysis of Anomaly Scores:

1. The path lengths of all the forest's trees are averaged to determine each data point's anomaly score.

2. Normal data points are anticipated to have shorter average paths, whereas anomalies are anticipated to have longer average paths.

Threshold and Identification of Anomalies:

1. To identify data points as anomalies or regular occurrences, choose an appropriate threshold for the anomaly scores. Techniques like percentile-based or domain-specific knowledge can be used to set the threshold.

2. Outliers are data points with anomaly scores above the threshold, whereas normal data points are those with scores below the threshold.

5.4 Evaluate the model:

In this section, we go into a detailed analysis of the important performance metrics and a thorough investigation of the confusion matrix to assess the success of the machine learning model. We can learn a lot about the model's precision, recall, and ability to forecast across different classes by analyzing these findings. This thorough analysis not only shows the model's advantages but also points out potential flaws, enabling future strategic planning and well-informed decisions. Considering 50% training and test ration we get the result of-

Accuracy: 0.966821551580382 The accuracy achieved by the model is an impressive 0.9668, indicating that approximately 96.68% of the model's predictions align with the actual outcomes. High accuracy suggests that the model's classifications are in good agreement with the true labels, demonstrating its overall reliability.

Precision: 0.9533070104482407 The precision of 0.9533 signifies that when the model predicts an instance to belong to a particular class, it is correct about 95.33% of the time. This high precision is particularly valuable when the cost of false positives is significant, and accurate identification of positive instances is crucial.

Recall: 0.966821551580382 The recall score, also known as sensitivity or true positive rate, shares the same value as accuracy at 0.9668. This indicates that the model is capable of capturing around 96.68% of all actual positive instances. In other words, the model effectively identifies positive cases, which is essential when the cost of false negatives is high.

F1 Score: 0.9598382891040446 The F1 score, a harmonic mean of precision and recall, is calculated as 0.9598. This score provides a balanced assessment of the model's performance, considering both false positives and false negatives. It serves as a valuable indicator when aiming for a balance between precision and recall.

Confusion Matrix: The confusion matrix provides a detailed breakdown of the model's predictions across different classes:

- True Negatives: 469,360 instances were correctly predicted as negative.
- True Positives: 261 instances were correctly predicted as positive.
- False Positives: 4,494 instances were falsely predicted as positive.
- False Negatives: 11,622 instances were incorrectly classified as negative.

The number of false negatives shows a potential area for improvement, despite the model having a high true positive rate and a relatively low false positive rate. These situations ought to have been rated as good but were instead incorrectly labeled as negative. Further investigation is necessary to comprehend the effects of false negatives and to consider possible solutions for this imbalance, depending on the context of the problem.

Chapter 6

Result Analysis

In this article, all of the tests are performed on a computer that has an Intel(R) Core (TM) i7-8550U CPU that operates at 1.80 GHz or 1.99 GHz and 8 GB of RAM, and the Jupyter Notebook is used for all of the programming. The precision, false positive rate (FPR), recall, and F1 score are the assessment metrics that are used in order to judge how efficient the recommended approach is. The parameters of the Isolation Forest Algorithm are adjusted in accordance with the requirements.

DTC: [Breiman et al., 2017] A classification issue that can be solved using supervised machine learning is referred to as a decision tree classifier. It accomplishes its goals by partitioning the data into subgroups according to a set of rules that are derived from the features of the data. These guidelines are presented in the form of a tree structure, with each node standing in for a feature-based judgment and each leaf node illustrating a conclusion on the classification. In order to correctly classify new data on the basis of patterns found in the training data, a DTC is used. It has widespread use in a variety of fields, including banking, marketing, and healthcare, where it is used to identify fraudulent activity, categorize clients, and diagnose illnesses.

RFC: [Breiman, 2001] The machine learning algorithm known as RFC, which stands for Random Forest Classifier, belongs to the family of learning methods known as ensemble learning approaches. It does this by generating a large number of decision trees and integrating their predictions in order to improve accuracy and reduce the risk of overfitting. Each tree in the forest is trained using a different subset of the attributes and data from the training data in order to increase the diversity of the trees and lower the correlation between them. RFC is often applied in settings where accurate and trustworthy classification is essential, such as in the banking industry, the healthcare industry, and in picture recognition.

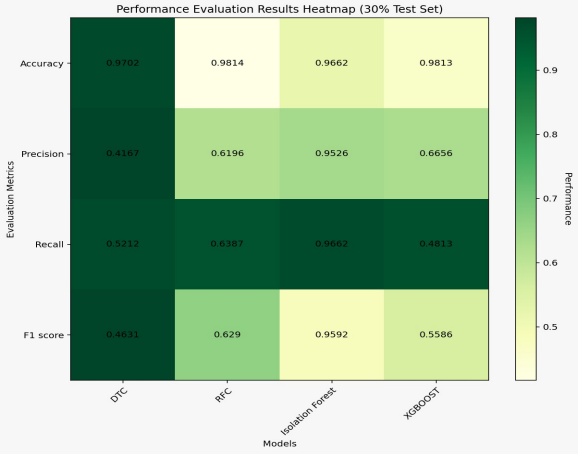
XGBOOST: [Chen and Guestrin, 2016] Extreme Gradient Boosting, sometimes referred to as XGBOOST, is a powerful method of machine learning that may be used to problems involving regression and classification. It is meant to be exceedingly scalable, efficient, and has gained favor in anomaly detection applications due to its potential to handle imbalanced datasets and strong accuracy performance. Additionally, it has a good accuracy performance. XGBOOST is used to boost the detection rate of irregularities in anomaly detection. This is accomplished by first constructing an

ensemble of decision trees, and then continually learning from the errors committed in previous rounds.

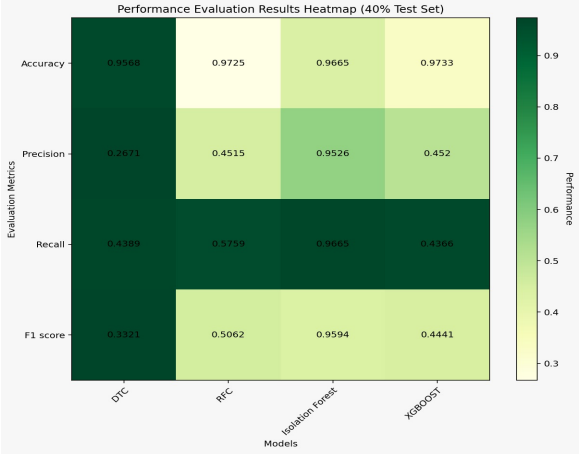
6.1 Metrics for Evaluating Performance

The objective of dividing the data on energy consumption into two groups—normal and abnormal (indicating that theft is likely to have occurred)—can be phrased as a binary classification problem in order to address the difficulty of identifying fraudulent activity with smart meters. Precision (Pre), Recall (Re), and F1 Scores (F1) are the metrics that are used to evaluate how well suggested solutions work. The effectiveness of anomaly detection has been assessed using the confusion matrix.

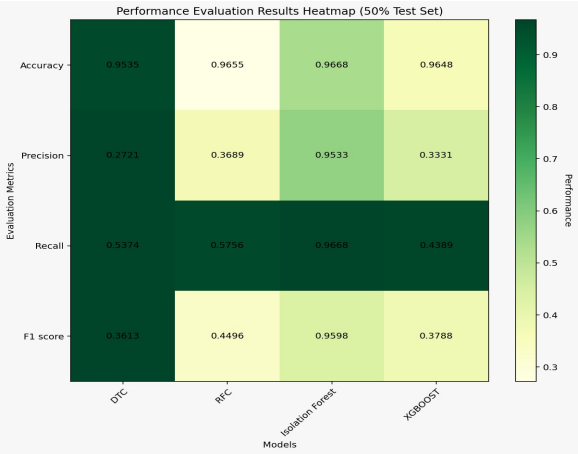
The following is the performance measure that we use, based on 6.1 and the concept of precision: It is possible to write the Recall and F1 Scores as,Precision is equal to the ratio of true positives to the sum of true positives and false positives [Depuru et al., 2011].



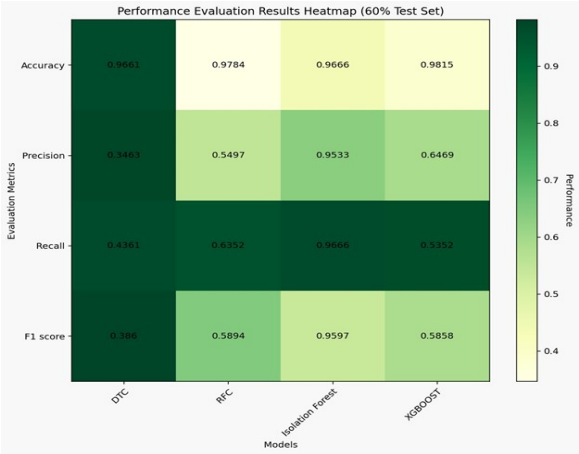
(a) 30% Test Set



(b) 40% Test Set



(c) 50% Test Set



(d) 60% Test Set

Figure 6.1: Performance Evaluation Results Heatmaps

Table 6.1: Confusion matrix applied in smart meter theft detection.

Users	Detected as a Theft User	Detected as a Normal Use
Theft users	TP (true positive)	FN (false negative)
Normal users	FP (false positive)	TN (true negative)

The formula for recall is true positives divided by the sum of true positives and false negatives [Jokar et al., 2015].

F1 Score = 2 * (Precision * Recall) / (Precision + Recall) [Goutte and Gaussier, 2005]

When compared to other approaches like DTC, RFC, and XGBOOST, the evaluation that was just presented leads us to the conclusion that Isolation Forest is the superior algorithm. When it comes to accuracy, other approaches may occasionally come out on top, but when it comes to anomaly detection, we need to be careful with a high false negative result. Because of this, we are able to identify with a high level of confidence that the isolation forest has a good precision on anomaly detection. The Recall % is the most critical factor to consider. It takes into account the capabilities of anomaly detection, and as a consequence, here in our findings, we can see that Isolation Forest has the greatest value in every performance in the Recall category.

6.2 Results obtained using the suggested Method

Our approach involves using the data from smart meters to train a model. Accuracy, precision, recall, and the F1 Score are the four main assessment metrics that are used in order to assess how well the suggested strategy performs. It is important to point out that we have picked data at random from our dataset in relation to 30%, 40%, 50%, and 60% of training ratios for the purpose of illustrating 6.4. After cleaning the data, we used about 971273 of them, and our method performed linearly 6.4a, whereas those of our competitors fluctuated often. Despite the fact that the accuracy is not the best, coming in at 96.65% in both the 40% and 60% evaluations. On the other hand, the other measure provides a score that is much higher than that of the other techniques. With the best possible score of 95.33 percent for accuracy, 96.65 percent for recall, and 95.97 percent for the F1 category.

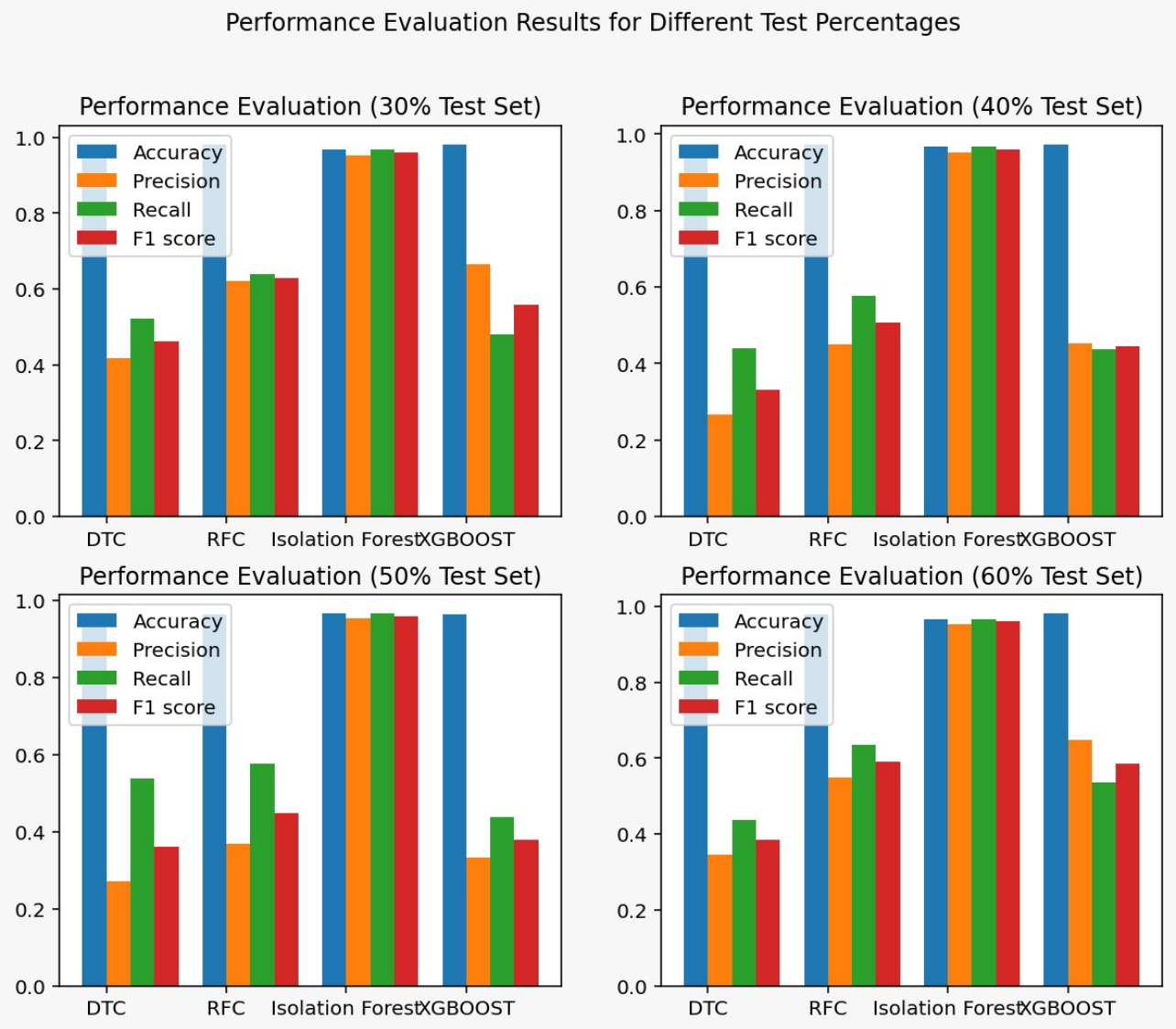


Figure 6.2: Performance Evaluation Bar Graphs for Different Test Percentages

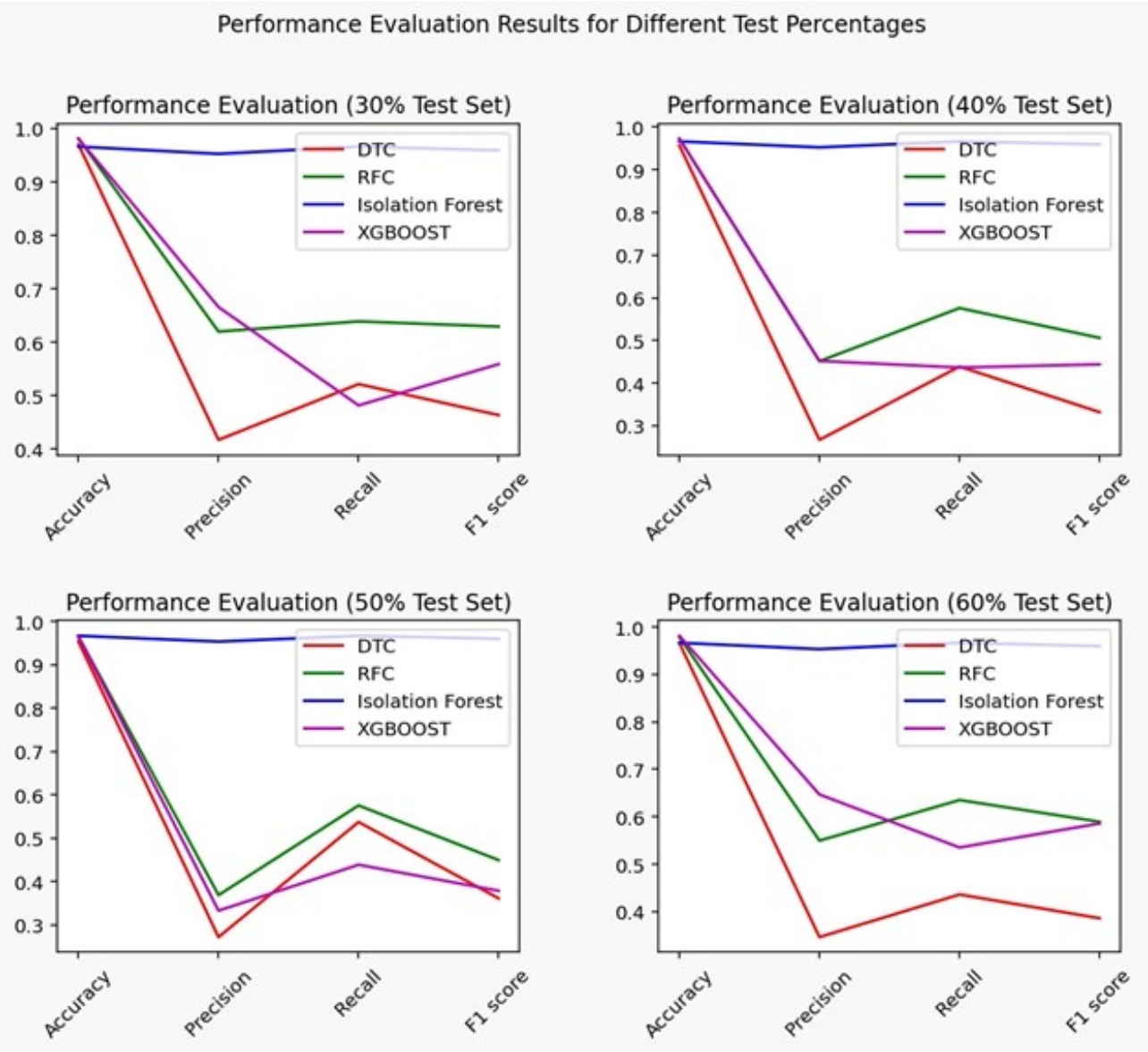
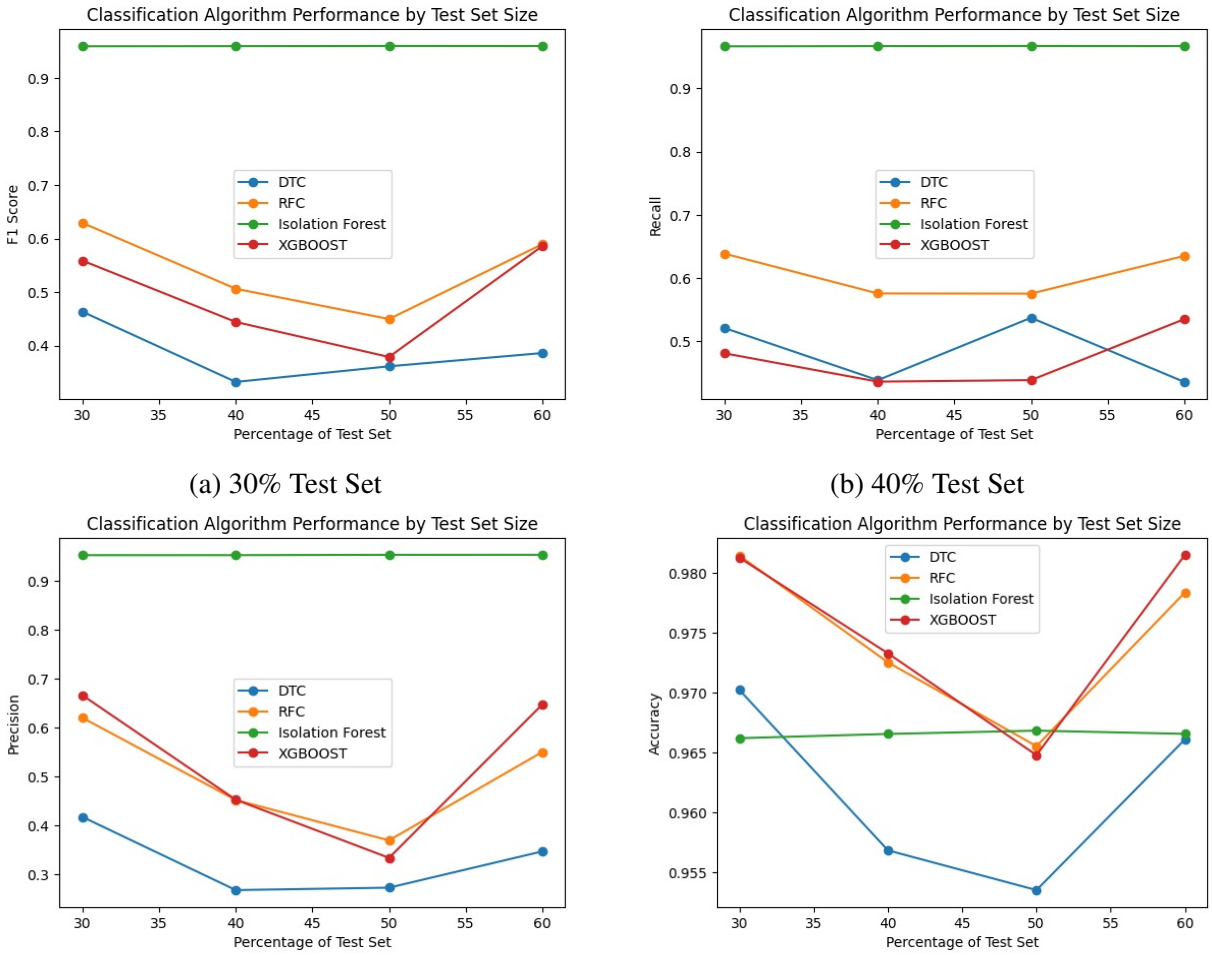


Figure 6.3: Performance Evaluation Line Charts for Different Test Sets



(a) 30% Test Set

(b) 40% Test Set

(c) 50% Test Set

(d) 60% Test Set

Figure 6.4: Classification Algorithm Performance By Test Set Size

When the assessment is conducted taking the training ratio of 60% into consideration 6.1, our technique has the best performance. The more information it is trained on, the greater performance we will be able to achieve.

		Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Considering 30% test set	Our Model	96.61	95.25	96.61	95.92
	DTC	97.02	41.67	52.12	46.31
	RFC	98.14	98.14	61.96	62.90
	XGBOOST	98.12	66.56	48.12	55.86
Considering 40% test set	Our Model	**96.65**	95.26	**96.65**	95.94
	DTC	95.68	26.71	43.88	33.11
	RFC	97.25	45.15	57.58	50.61
	XGBOOST	97.32	45.19	43.65	44.41
Considering 50% test set	Our Model	96.61	95.25	96.61	95.92
	DTC	95.35	27.20	53.74	36.12
	RFC	96.55	36.88	57.56	44.96
	XGBOOST	96.47	33.30	43.89	37.87
Considering 60% test set	Our Model	**96.65**	**95.33**	**96.65**	**95.97**
	DTC	96.60	34.63	43.60	38.60
	RFC	97.83	54.97	63.51	58.93
	XGBOOST	98.14	64.68	53.52	58.57

Table 6.2: Results of considering different performance evaluations with different test sets.

6.3 A Comparison of the Performance

Therefore, by using the equation that was just discussed as well as the data shown in table 6.2 (although taking into account just fifty percent as the foundation for comparison), we are able to observe that precision, which is also referred to as predictive value. Due to the fact that FN and TN are not accounted for in the method, the outcomes of our imbalanced courses may be affected negatively. Whereas our method produces much superior results than those obtained by the other in % of cases. The rate at which a fraud detection leads to a positive test result (also known as the true-positive rate) is referred to as recall. As we can see, this technique does not incorporate FP or TN; hence, sensitivity may result in an answer that is skewed. When reporting 50% of the performance assessment data set, the classifier has a sensitivity of 96.61%. In compared to the previous two measures, the F1 score incorporates both Recall and Precision and provides a more balanced view; yet, due to the fact that it does not include TN, it has the potential to generate biased findings in some circumstances. However, the F1 score that our technique achieves for every assessment test set is much higher than that of the other methods.

Chapter 7

Conclusion

Advanced Metering Infrastructure (AMI) and smart meters have emerged as crucial components in the quest for enhanced energy efficiency and waste reduction. These technologies provide real-time data on energy consumption, enabling utilities to remotely monitor and manage energy flow. Moreover, smart meters empower customers with the knowledge needed to make informed decisions about their energy usage, leading to conservation efforts and lower energy bills. Security issues, such as the possibility of cyber attacks and illegal access to private information, have been raised in response to the increasing implementation of smart meters. Strong security measures, such as encryption and access restrictions, are required to deal with these dangers. In this context, the isolation forest algorithm, a machine learning technique, has proven to be a valuable tool for detecting anomalies in the data collected from smart meters. By dividing data points into subsets and isolating anomalies, the isolation forest algorithm facilitates the identification of irregularities and potential security breaches. To assess its performance, the algorithm was compared with three other popular machine learning algorithms: Classifiers like Random Forest, the Decision tree and XGBoost. Evaluation metrics such as precision, recall, and F1 score were utilized to gauge the algorithms' effectiveness. The results revealed that the isolation forest algorithm outperformed the other methods, particularly in terms of its ability to detect anomalies with a high recall value. Notably, the proposed method achieved its peak performance when the training ratio was set at 60 percent, striking a balance between precision and recall through the F1 score. By harnessing the power of the isolation forest algorithm, utilities can effectively detect energy theft and other fraudulent activities, optimizing their energy management systems and yielding additional cost savings and environmental benefits. The accurate identification of anomalies and security breaches not only ensures the integrity and reliability of the data collected from smart meters but also safeguards customer information and prevents disruptions in the energy supply. In conclusion, while AMI and smart meters offer significant advantages in improving energy efficiency and waste reduction, it is vital to prioritize the implementation of robust security measures to mitigate the risks associated with cyber threats. The isolation forest algorithm emerges as a promising solution for detecting anomalies and optimizing energy management systems. By combining robust security measures with advanced data analysis techniques, the energy sector can continue to harness

the benefits of AMI and smart meters while ensuring the protection of customer data and the integrity of energy distribution networks. This approach paves the way for a more sustainable and secure energy landscape, ultimately resulting in long-term benefits for both utilities and consumers alike.

Bibliography

[Abdalzaher et al., 2022] Abdalzaher, M. S., Fouda, M. M., and Ibrahem, M. I. (2022). Data privacy preservation and security in smart metering systems. *Energies*, 15(19):7419.

[Al-Fuqaha et al., 2015] Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., and Ayyash, M. (2015). Internet of things: A survey on enabling technologies, protocols, and applications. *IEEE Communications Surveys & Tutorials*, 17(4):2347–2376.

[Anand et al., 2021] Anand, A., Chaudhary, Y., Mukherjee, R., and Yadav, A. (2021). Gsm based smart energy meter with theft detection and load control. In *2021 7th International Conference on Signal Processing and Communication (ICSC)*, pages 59–62.

[Aziz et al., 2020] Aziz, S., Naqvi, S. Z. H., Khan, M. U., and Aslam, T. (2020). Electricity theft detection using empirical mode decomposition and k-nearest neighbors. In *2020 International Conference on Emerging Trends in Smart Technologies (ICETST)*, pages 1–5. IEEE.

[Bentéjac et al., 2019] Bentéjac, C., Csörgő, A., and Martínez-Muñoz, G. (2019). A comparative analysis of xgboost.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

[Breiman et al., 2017] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (2017). Regression trees. *Classification And Regression Trees*, page 216–265.

[Caropreso et al., 2018] Caropreso, R. d. T., Fernandes, R. A., Osorio, D. P., and Silva, I. N. (2018). An open-source framework for smart meters: Data communication and security traffic analysis. *IEEE Transactions on Industrial Electronics*, 66(2):1638–1647.

[Chen and Guestrin, 2016] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

[Depuru et al., 2011] Depuru, S. S. S. R., Wang, L., Devabhaktuni, V., and Nelapati, P. (2011). A hybrid neural network model and encoding technique for enhanced classification of energy consumption data. In *2011 IEEE power and energy society general meeting*, pages 1–8. IEEE.

[Eder-Neuhauser et al., 2016] Eder-Neuhauser, P., Zseby, T., and Fabini, J. (2016). Resilience and security: A qualitative survey of urban smart grid architectures. *IEEE Access*, 4:839–848.

[Ford et al., 2014] Ford, V., Siraj, A., and Eberle, W. (2014). Smart grid energy fraud detection using artificial neural networks. volume 2015.

[Goutte and Gaussier, 2005] Goutte, C. and Gaussier, E. (2005). A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *Advances in Information Retrieval: 27th European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain, March 21-23, 2005. Proceedings 27*, pages 345–359. Springer.

[Halle and Shiyamala, 2021] Halle, P. D. and Shiyamala, S. (2021). Ami and its wireless communication security aspects with qos: A review. *Advances in Smart System Technologies: Select Proceedings of ICFSST 2019*, pages 1–13.

[Iorga and Shorter, 2015] Iorga, M. and Shorter, S. (2015). Advanced metering infrastructure smart meter upgradeability test framework. *National Institute of Standards and Technology Interagency*, 7823:71–71.

[Jokar et al., 2015] Jokar, P., Arianpoo, N., and Leung, V. C. (2015). Electricity theft detection in ami using customers' consumption patterns. *IEEE Transactions on Smart Grid*, 7(1):216–226.

[Kalogridis et al., 2013] Kalogridis, G., Sooriyabandara, M., Fan, Z., and Mustafa, M. A. (2013). Toward unified security and privacy protection for smart meter networks. *IEEE Systems Journal*, 8(2):641–654.

[Khattak et al., 2019] Khattak, A. M., Khanji, S. I., and Khan, W. A. (2019). Smart meter security: Vulnerabilities, threat impacts, and countermeasures. In *Proceedings of the 13th International Conference on Ubiquitous Information Management and Communication (IMCOM) 2019 13*, pages 554–562. Springer.

[Kumar et al., 2019] Kumar, P., Lin, Y., Bai, G., Paverd, A., Dong, J. S., and Martin, A. (2019). Smart grid metering networks: A survey on security, privacy and open research issues. *IEEE Communications Surveys & Tutorials*, 21(3):2886–2927.

[Liu et al., 2008] Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422.

[Marah et al., 2020] Marah, R., El Gabassi, I., Larioui, S., and Yatimi, H. (2020). Security of smart grid management of smart meter protection. In *2020 1st international conference on innovative research in applied science, engineering and technology (IRASET)*, pages 1–5. IEEE.

[MarketScreener, 2023] MarketScreener (2023). Smart meter, llc - 51 million americans were affected by healthcare data breaches in 2022 due to weak security: Marketscreener.

[Mendel, 2017] Mendel, J. (2017). Smart grid cyber security challenges: Overview and classification. *e-mentor*, 2017:55–66.

[Mishra and Tiwari, 2020] Mishra, P. K. and Tiwari, P. (2020). Cyber security in smart grid. *International Research Journal on Advanced Science Hub*, 2(6):26–30.

[Patruni et al., 2022] Patruni, M., Deebak, D. B. D., and Al-Turjman, F. (2022). *Smart grid metering: security, privacy, and open challenges*, pages 255–272.

[Polčák, 2023] Polčák, L. (2023). Responsible and safe home metering: How to design a privacy-friendly metering system. In *Information Security and Privacy in Smart Devices: Tools, Methods, and Applications*, pages 1–40. IGI Global.

[Prabhakar et al., 2022] Prabhakar, P., Arora, S., Khosla, A., Beniwal, R. K., Arthur, M. N., Arias-Gonzáles, J. L., Areche, F. O., et al. (2022). Cyber security of smart metering infrastructure using median absolute deviation methodology. *Security and Communication Networks*, 2022.

[Raciti and Nadjm-Tehrani, 2013] Raciti, M. and Nadjm-Tehrani, S. (2013). Embedded cyber-physical anomaly detection in smart meters. In *Critical Information Infrastructures Security: 7th International Workshop, CRITIS 2012, Lillehammer, Norway, September 17-18, 2012, Revised Selected Papers*, pages 34–45. Springer.

[Shehzad et al., 2021] Shehzad, F., Javaid, N., Almogren, A., Ahmed, A., Gulfam, S. M., and Radwan, A. (2021). A robust hybrid deep learning model for detection of non-technical losses to secure smart grids. *IEEE Access*, 9:128663–128678.

[Toftegaard et al., 2022] Toftegaard, O., Hagen, J., and Hämmerli, B. (2022). Are european security policies ready for advanced metering systems with cloud back-ends? *Critical Infrastructure Protection XVI*, page 47–69.

End quote goes here.