Design And Development Of Energy Theft Detection Model For Smart Meters

Thesis Submitted for the Award of the Degree of

DOCTOR OF PHILOSOPHY

In

School of Computer Applications

By

Asif Iqbal Kawoosa

Registration Number: 41800712

Supervised By

Dr. Deepak Prashar

Professor, School of Computer Science & Engineering

Lovely Professional University, Phagwara, Punjab, India

LOVELY PROFESSIONAL UNIVERSITY, PUNJAB

2024

DECLARATION

I, hereby declared that the presented work in the thesis entitled "Design and Development of Energy Theft Detection Model For Smart Meters" in fulfilment of degree of **Doctor of Philosophy (Ph.D.)** is outcome of research work carried out by me under the supervision Dr. Deepak Prashar, working as Professor, in the School of Computer Science & Engineering of Lovely Professional University, Punjab, India. In keeping with general practice of reporting scientific observations, due acknowledgements have been made whenever work described here has been based on findings of other investigator. This work has not been submitted in part or full to any other University or Institute for the award of any degree.

(Signature of Scholar)

Name of the scholar: Asif Iqbal Kawoosa

Registration No.: 41800712

Department/school: School of Computer Applications

Lovely Professional University,

Punjab, India

CERTIFICATE

This is to certify that the work reported in the Ph.D. thesis entitled “Design and Development of Energy Theft Detection Model for Smart Meters”submitted in fulfillment of the requirement for the reward of degree of **Doctor of Philosophy (Ph.D.)** in the School of Computer Applications, is a research work carried out by Asif Iqbal Kawoosa, Registration No.: 41800712, is bonafide record of his/her original work carried out under my supervision and that no part of thesis has been submitted for any other degree, diploma or equivalent course.

(Signature of Supervisor)

Name of supervisor: Deepak Prashar

Designation: Professor

Department/school: School of Computer Science & Engineering

University: Lovely Professional University, Phagwara, Punjab, India University

ACKNOWLEDGEMENT

First and foremost, I am profoundly grateful to Almighty Allah, whose grace and blessings have guided me through this journey and made this accomplishment possible. His wisdom and mercy have been my constant companions.

I would like to extend my heartfelt gratitude to my mother, whose unwavering support, prayers, and encouragement have been the foundation of my success. I also wish to express my deepest appreciation to my late father, whose unconditional love and support in my early life continue to be my greatest source of strength and inspiration.

To my dear family, wife, and children, I express my deepest thanks for their patience, love, and understanding throughout this challenging endeavour. Their sacrifices and constant encouragement have been my motivation to persevere.

I am also immensely grateful to the School of Computer Applications at Lovely Professional University, Phagwara, Punjab, India, for providing an inspiring and conducive environment for my research. The resources, guidance, and support from the faculty and staff have been invaluable to my academic growth and success.

I am deeply thankful to my supervisor, **Dr. Deepak Prashar** of the School of Computer Science & Engineering, Lovely Professional University, Phagwara, Punjab, India. His support extended far beyond the boundaries of this thesis. His mentorship, guidance, and encouragement have been instrumental not only in shaping this research but also in my overall academic and personal development. His unwavering belief in my potential and his readiness to assist in times of need have profoundly impacted my journey.

Lastly, I want to acknowledge the support and camaraderie of some very special friends. Their friendship, insights, and encouragement have played a significant role in this journey, and for that, I am sincerely thankful.

This thesis is a testament to the collective support and blessings I have received from everyone around me. Thank you all.

ASIF IQBAL KAWOOSA

ABSTRACT

Electricity theft poses significant challenges to utility companies by causing substantial financial losses and compromising the efficiency of energy distribution systems. Despite the advancements in Automatic Meter Reading (AMR) and Advanced Metering Infrastructure (AMI), which facilitate accurate measurement and monitoring of electricity consumption, these technologies are vulnerable to manipulation and tampering. This necessitates the development of robust detection methods.

This study examines the complexities of detecting electricity theft using both conventional and smart energy meters. Conventional meters record total consumption over a billing period and rely on periodic manual readings, making theft detection difficult and dependent on visual inspections and historical data comparisons. In contrast, smart meters provide real-time, granular data, enabling the quick detection of anomalies in usage patterns and voltage fluctuations. However, distinguishing between legitimate variations in consumption patterns and fraudulent activities remains a significant challenge. Existing detection methods often suffer from high false positive rates and are ineffective in handling the dynamic nature of electricity consumption influenced by environmental factors and consumer behaviour.

To address these challenges, an ensemble KPLX model is proposed, integrating K-means clustering for initial data segmentation, Principal Component Analysis (PCA) for feature reduction, Long Short-Term Memory (LSTM) networks for capturing temporal patterns and eXtreme Gradient Boosting (XGBoost) algorithm for classification of theft and non-theft case. Various machine learning algorithms are explored, including Support Vector Machines (SVM), Decision Trees, Random Forest, Logistic Regression, K-Nearest Neighbors (K-NN), AdaBoost, CatBoost, LightGBM, and XGBoost. Hyperparameter tuning is meticulously performed using grid search and cross-validation techniques to optimize model performance.

The study utilizes extensive datasets from the State Grid Corporation of China (SGCC) and Kashmir Power Development Corporation Limited (KPDCL), comprising 42,372 and 1,048,576 records respectively, collected at 15-minute intervals. Data preprocessing steps include filling missing values, removing outliers, and normalizing the data to ensure integrity and reliability. Feature engineering enhances the model's capability to detect six predefined types of theft attacks.

Among the evaluated models, XGBoost demonstrates the highest efficacy in detecting electricity theft, achieving an Area under the Curve (AUC) score of 0.96, precision of 0.94, recall of 0.91, and F1-score of 0.92. The ensemble model shows superior results, with a notable reduction in false positive rates and enhanced accuracy across various theft types. Hyperparameter tuning of XGBoost, involving parameters such as learning rate, max depth, subsample, and colsample_bytree, significantly contributes to these improved outcomes.

The robustness of the proposed model (KPLX Integrated Detection Model) is validated against environmental variations and adversarial attacks, proving its reliability in real-world applications. The model's scalability and efficiency make it suitable for large-scale deployment in smart grid systems. The integration of diverse detection techniques ensures a comprehensive defence against different types of theft attacks, enhancing the overall security and integrity of electricity distribution networks.

In conclusion, this research offers a potent solution to electricity theft detection, combining advanced machine learning techniques with robust feature engineering and hyperparameter optimization. The proposed KPLX Integrated Detection Model outperforms traditional methods and sets a new benchmark in the field. It holds significant potential for applications in enhancing grid security, reducing economic losses, and ensuring reliable energy distribution. Future research directions include integrating the detection framework with emerging smart grid technologies, developing real-time detection systems, and validating the framework with broader datasets to ensure its generalizability and robustness across diverse contexts.

Keywords: Electricity Theft, Smart Meters, Machine Learning, K-means Clustering, Long Short-Term Memory, XGBoost, Anomaly Detection, Feature Engineering, Hyperparameter Tuning, Grid Security.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background and Motivation

In today's fast-moving world, we all seek convenient ways to manage our lives, especially when it comes to tasks like paying for electricity. Automatic Meter Reading (AMR) is a technology that assists with this by automatically tracking the electricity usage and sending that information to the electric company for billing [1]. AMR relies on various communication methods, such as wires or wireless signals. The communication setup in AMR can be configured in two ways: one-way (unidirectional), or two-way (bidirectional) communication [2]. In one-way communication, known as conventional AMR, the system reads the utility consumption of each user and reports to the central utility provider. This is unidirectional communication between each meter and utility provider. Conversely, the advent of two-way communication, now known as automatic metering infrastructure (AMI), facilitates seamless bidirectional interaction between users, meters, and the central utility. This transparent communication empowers users with real-time consumption data and billing insights, enabling them to proactively manage their energy usage and budget effectively [3]. However, these meters are susceptible to various forms of manipulation and tampering, which can result in energy theft like manipulating meter readings or tampering with metering equipment, making it unreliable for billing purposes and data analysis. This could potentially lead to further financial losses and operational inefficiencies [4]. Attacks on AMIs create uncertainty in determining the amount of power to generate and consume, hindering accurate energy demand predictions. Energy theft detection (ETD) methods require enhancements to improve efficiency and increase the detection rate of theft cases. This study examines the limitations of current ETD methods and proposes solutions to address these threats. It explores innovative detection techniques to tackle the challenges of false positives, bridging existing gaps in the literature, and striving towards more robust solutions.

1.2 Introduction

Electricity theft continues to remain a major global problem, with different levels of prevalence depending on economic, social, and regulatory influences. The extent and modes of electricity theft may differ, but it generally results in significant revenue losses to energy suppliers and weakens the power grid's reliability. Countries are progressively embracing advanced technologies like smart meters, data analytics, and machine learning to reduce these losses. In the context of electricity theft detection, unauthorized tapping or tampering with the system can disrupt legitimate consumption patterns, making it crucial to accurately monitor and detect anomalies in the usage of electric power [5].

A device with a power rating of one kilowatt runs for one hour and consumes one kilowatt-hour of energy. Units such as joules (J) or kilowatt-hours (kWh) typically measure electrical energy. This measurement is crucial in monitoring and quantifying the amount of electricity consumed by users, which is an essential factor in identifying unusual consumption patterns that may indicate electricity theft [6]. An electric energy meter is essential for identifying electricity theft as it quantifies the electric energy flowing across its circuits and records the usage for metering purposes. These meters are critical for accurately estimating and monitoring power usage in a readable format, typically in kilowatt-hours (kWh) [7]. This measurement allows consumers and utility providers to monitor energy usage patterns, identify inefficiencies, and make informed decisions about energy consumption and reductions. Nevertheless, these meters are vulnerable to a variety of manipulation and tampering, which may result in the theft of electricity [8]. Electricity theft, whether manipulating meter readings or tampering with metering equipment, distorts the data recorded by the energy meters, making it unreliable for billing purposes and data analysis, potentially leading to further financial losses and operational inefficiencies.

1.3 Categorization of Smart Meters

An advanced digital device known as a smart meter measures and monitors electricity, gas, or water consumption in homes, businesses, and other facilities [9]. Smart meters have several benefits compared to traditional energy meters. They offer real-time monitoring of electricity usage, enabling more precise and prompt metering. Smart meters can identify abnormal usage patterns and possible manipulation through their abundant data availability, unlike conventional meters. In addition, they provide remote data transmission, reducing the need for manual meter readings and improving operating efficiency for utility companies. Smart meters enable users to monitor and control their energy consumption more efficiently. [10]. Smart meters measure power usage in real time, at very close 15-minute intervals. Figure 1.1 illustrates the basic block diagram of a smart meter. There are various types of smart meters in the design; each has its own specific application and utility services. [11].

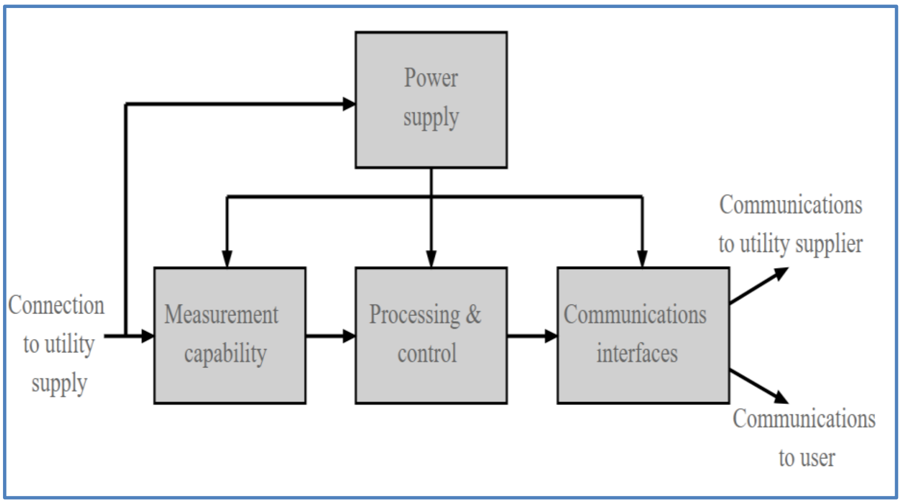
Power supply
Communications to utility supplier
Connection to utility supply
Measurement capability
Processing & control
Communications interfaces
Communications to user

Figure 1.1: Smart Meter Basic Block Diagram

- **Electricity Smart Meters:** Smart meters, which measure electrical energy consumption in kilowatt-hours (kWh), are an advanced type of energy meter. They provide real-time data without a need for manual readings, allowing for accurate billing based on actual consumption. They also offer features such as remote monitoring, which helps identify and address issues like energy theft and inefficiencies more effectively. Additionally, smart meters can provide consumers with detailed insights into their energy usage patterns, empowering them to make informed decisions about their consumption [12].
- **Gas Smart Meters:** Gas smart meters measure natural gas or propane consumption in cubic meters or cubic feet. They monitor gas usage, detect leaks, and facilitate remote monitoring and control of gas distribution networks [13].
- **Water Smart Meters:** Water smart meters measure water consumption in cubic meters or gallons. They track water usage, identify leaks, and help conserve water resources by providing accurate data on consumption [14].

- **Heat Smart Meters:** Heat smart meters measure the consumption of thermal energy for heating or cooling purposes. They are used in district heating and cooling systems to monitor energy usage and optimize efficiency [15].
- **Dual-Fuel Smart Meters:** Dual-fuel smart meters combine the functionality of electricity and gas meters into a single device. They provide integrated monitoring and billing for both energy sources, offering convenience for consumers and utilities [15].
- **Advanced Metering Infrastructure (AMI) Meters:** These meters are integral components of a comprehensive system that also includes smart meters, communication networks, and data management software. AMI meters enable remote monitoring, data analytics, and demand-response programs, allowing utilities to optimize grid operations and improve efficiency [16].
- **Time-of-Use (TOU) Smart Meters:** These meters measure energy consumption based on different periods, such as peak and off-peak hours. They allow utilities to implement time-based pricing strategies, encouraging consumers to shift their energy usage to off-peak hours and thereby reducing load on the grid [17].

1.4 Electricity Energy Theft in Smart Meters

Electrical infrastructures are categorized as critical infrastructures due to their essential role in maintaining the functionality and stability of a country. Disruptions in these systems can have severe consequences on the economy, development, and public services [18]. Electricity powers virtually all aspects of modern life, including transportation, agriculture, communication, healthcare, and financial services, making its security a top priority. The threat landscape for electrical infrastructures is vast, with cybercriminals, hackers, and terrorists constantly seeking to exploit vulnerabilities for malicious purposes. A successful attack on the electric grid could result in widespread blackouts affecting entire cities, states, or even countries. However, the increasing number of smart meters and the automation of the electricity infrastructure bring out new security concerns. Malicious individuals, including cybercriminals, hackers, and even terrorists, continuously attempt to take advantage of vulnerabilities in these systems, which might result in major black outs with the ability to impact whole cities, states, or even nations. Due to the vital importance of electrical infrastructure, governments place a high priority on its security regardless of political ideologies. Despite constant enhancements in security measures, there has been a rise in new challenges, including cyber-attacks, in the last ten years. This emphasizes the importance of always remaining vigilant and proactive in safeguarding smart meter networks and the broader electricity infrastructure [19]. Additionally, Electricity theft in smart meter systems refers to the unauthorized use or manipulation of electricity consumption data in devices equipped with smart meter technology. Smart meters are advanced devices that digitally measure and record electricity usage in real time, providing detailed insights into consumption patterns. However smart meters are susceptible to various forms of manipulation and tampering, which can result in energy theft [20]. Electricity theft through smart meters can occur in several forms, including bypassing the meter, hacking into the communication system, or physically tampering with the meter itself. Electricity energy thefts are of various types:

- **Meter Tampering:** Meter tampering involves physically altering or bypassing the smart meter's mechanisms to manipulate consumption readings. This can include methods such as short-circuiting, magnet tampering, or altering the meter's firmware [21].
- **Meter Bypassing:** Meter bypassing entails creating alternative wiring paths to divert electricity away from the meter, allowing unauthorized consumption without detection. Common techniques include tapping directly into power lines or installing unauthorized connections before the meter [21].
- **Illegal Connections:** Illegal connections involve tapping into the electrical grid without proper authorization, often through makeshift or hazardous wiring setups. These connections pose safety risks and can result in overloaded circuits, fires, and electrocution hazards [22].
- **Manipulation of Meter Readings:** A manipulating meter reading involves exploiting vulnerabilities in smart metering systems to falsify consumption data. This can be achieved through software hacks, device tampering, or exploiting communication protocols to intercept or alter data transmission [22].

1.5 Detection Methods for Power Theft

Detection methods for power theft encompass a range of techniques and strategies aimed at identifying and preventing unauthorized consumption of electricity. These methods can be broadly categorized into three main approaches: network-oriented, data-oriented, and hybrid approaches.

1.5.1 Network-Oriented Approaches

Network-oriented detection methods focus on analyzing the physical characteristics and behavior of the electrical grid to identify anomalies indicative of power theft. These methods leverage data collected from sensors, meters, and other monitoring devices deployed throughout the distribution network [23]. Key techniques used in network-oriented approaches include:

Load Profiling: Load profiling involves analyzing the consumption patterns of individual customers or groups of customers over time. Deviations from expected load profiles may indicate abnormal or unauthorized usage, signalling potential instances of power theft [24].

Power Quality Monitoring: Power quality monitoring involves assessing the quality and stability of electrical signals within the distribution network. Irregularities in voltage, frequency, or waveform characteristics may suggest the presence of unauthorized connections or tampering with metering equipment.

Line Loss Analysis: Line loss analysis involves measuring the discrepancy between the amount of electricity supplied to a distribution network and the amount consumed by customers. Unexplained losses or discrepancies beyond typical levels may indicate losses due to theft or technical inefficiencies.

Geographic Information Systems (GIS): GIS technology enables utilities to visualize and analyze spatial data related to electricity consumption, infrastructure, and demographics. GIS-based analysis can help identify areas with unusually high or low consumption relative to neighbouring regions, providing insights into potential theft hotspots [25].

1.5.2 Data-Oriented Approaches

Data-oriented detection methods focus on analyzing consumption data collected from smart meters, billing records, and other sources to identify patterns or anomalies indicative of power theft. These methods use statistical analysis, machine learning algorithms, and data mining techniques to identify irregularities and flag suspicious activities [26]. Key techniques used in data-oriented approaches include:

Anomaly Detection: Anomaly detection involves identifying deviations from normal consumption patterns or statistical norms. Machine learning algorithms trained on historical consumption data can detect anomalies indicative of power theft, such as sudden spikes or consistent underreporting of usage [26].

Pattern Recognition: Pattern recognition techniques involve identifying recurring patterns or signatures associated with power theft. Statistical analysis of consumption data can reveal distinctive patterns characteristic of theft behavior, allowing utilities to identify and flag suspicious accounts for further investigation.

Load Correlation Analysis: Load correlation analysis involves comparing the consumption patterns of interconnected customers or groups of customers to identify correlations or discrepancies in usage. Unusual correlations or inconsistencies may indicate unauthorized connections or collusion among customers to steal electricity [27].

1.5.3 Hybrid Approaches

Hybrid approaches combine elements of both network-oriented and data-oriented detection methods to enhance the effectiveness and accuracy of power theft detection. These approaches leverage the complementary strengths of network-level data and customer-level consumption data to identify and mitigate theft-related risks [28]. Hybrid approaches may involve:

Integrating Network and Customer Data: Hybrid approaches integrate data from multiple sources, including network-level sensors, smart meters, billing records, and historical consumption data. By combining network-level insights with customer-level consumption patterns, utilities can identify theft-related anomalies more effectively [28].

Multi-Stage Analysis: Hybrid approaches often employ multi-stage analysis workflows that incorporate both network-oriented and data-oriented techniques. These workflows may involve pre-processing network data to identify potential theft indicators, followed by detailed analysis of customer-level consumption data to confirm suspicions and identify specific theft patterns [29].

1.6 Strategies to Combat Smart Meter Energy Theft

Effective strategies to combat electricity energy theft in smart meters involve a combination of technological solutions, regulatory measures, and collaborative efforts between utility providers, law enforcement agencies, and consumers [30]. Here are some key strategies

Advanced Metering Technology: Implementing advanced smart metering technology with robust tamper detection mechanisms, encryption protocols, and real-time monitoring capabilities can enhance the detection and prevention of electricity theft. Smart meters equipped with advanced analytics algorithms can identify suspicious consumption patterns and alert utility providers to potential instances of theft [30].

Remote Monitoring and Management: Leveraging remote monitoring and management capabilities of smart meters allows utility providers to track electricity usage in real-time and detects irregularities indicative of theft. Remote access to metering data enables proactive intervention and response to suspicious activities, minimizing the impact of theft on revenue and grid reliability [31].

Data Analytics and Machine Learning: Utilizing data analytics and ML improve the detection of energy theft by analyzing large volumes of consumption data and identifying the patterns suggestive of fraudulent activities in electricity consumption. Machine learning algorithms can be trained on previous data to detect subtle deviations from normal usage patterns and flag potential instances of theft for further investigation [31].

Tamper-Proof Design and Security Measures: Designing smart meters with tamper-proof enclosures, secure communication protocols, and embedded security features can deter tampering and unauthorized access. Using robust security protocols, including encryption, access control, and authentication helps in protecting metering data and prevent unauthorized manipulation or tampering [32].

Public Awareness and Education: Educating consumers about the risks and consequences of electricity theft, as well as the importance of reporting suspicious activities, can help deter theft and foster a culture of compliance. Public awareness campaigns, outreach initiatives, and consumer education programs can empower individuals to play an active role in combating electricity theft and promoting a culture of integrity and accountability [32].

Regulatory Enforcement and Legal Measures: Enforcing strict regulatory standards and penalties for electricity theft, including fines, penalties, and legal sanctions, may send a strict warning to the fraudsters and offenders involved in electricity theft. Collaborating with law enforcement agencies to investigate and prosecute cases of electricity theft helps deter criminal activity and uphold the integrity of the energy grid [33].

Collaboration and Information Sharing: Fostering collaboration and information sharing among utility providers, industry stakeholders, and law enforcement agencies facilitates the exchange of best practices, insights, and intelligence related to electricity theft detection and prevention. Collaborative initiatives, task forces, and partnerships can enhance the effectiveness of anti-theft efforts and promote a coordinated response to emerging threats [33].

1.7 Challenges in Addressing Electricity Theft in Smart Meters

Electricity energy theft in smart meters not only results in substantial financial losses for utility providers but also poses operational challenges that affect the reliability and efficiency of electricity distribution networks. Electricity theft in emerging economies represents a significant challenge, leading to substantial financial losses and disrupting the balance between supply and demand. Globally, utility companies lose billions annually, impacting both developed and developing countries [34]. The problem extends to rich nations like the United States and the UK, where losses due to illegal consumption reach billions of dollars annually. These behaviors not only result in financial losses but also affect the reliability of power systems by overloading transformers and causing voltage imbalances [35]. In countries like India, where non-technical loss (NTL) due to electricity theft amounts to approximately $17 billion annually, utilities face considerable hurdles in improving power networks and achieving financial stability [35]. Below give figure 1.2 shows the Electricity energy theft rates worldwide (Reproduced with permission from Fehrenbacher 2013).

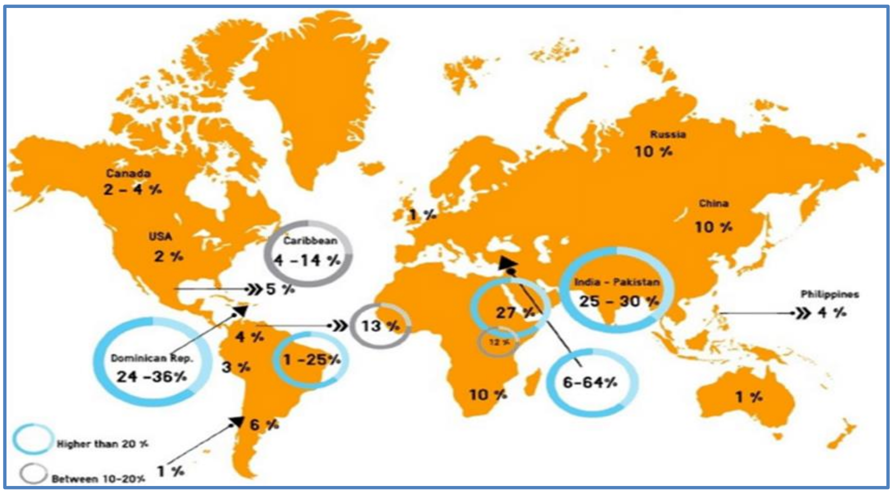

Figure 1.2: Global Electricity Theft Rates (Adapted with Permission from Fehrenbacher, 2013)

Addressing electricity theft in smart meters presents multifaceted challenges. False alarms due to consumption variations, technical vulnerabilities, resource limitations, and privacy concerns complicate detection efforts [36].

False Alarms: Smart meters may generate false alarms due to legitimate variations in consumption patterns, leading to unnecessary investigations and resource allocation [36].

Technical Constraints: Smart meters may have limitations in detecting sophisticated theft techniques or bypass methods, requiring ongoing research and development to stay ahead of evolving threats [37].

Resource Limitations: Utility providers may face challenges in dedicating sufficient resources to monitor and investigate potential instances of electricity theft, including manpower, time, and financial constraints [37].

Privacy Concerns: Smart meters raise privacy concerns regarding the collection and use of consumer data for theft detection purposes. Balancing the need for effective theft prevention with consumer privacy rights requires clear policies and robust safeguards to protect sensitive information [37].

1.8 Exploring Diverse Types of Machine Learning: An Overview

Machine learning (ML) is considered as a branch of artificial intelligence (AI) that primarily is used to create algorithms and statistical methods to enable computers to compute without explicit instructions. Instead, these systems rely on patterns formed derived from historical data. ML algorithms are trained on historical stored databases to recognize or construct the patterns in the data, and based on that make decisions [38]. Machine learning can be broadly categorized into three types (Figure 1.3): supervised learning (SL), unsupervised learning (USL), and reinforcement learning (RL). Additionally, there are specialized forms like semi-supervised learning and self-supervised learning [39].

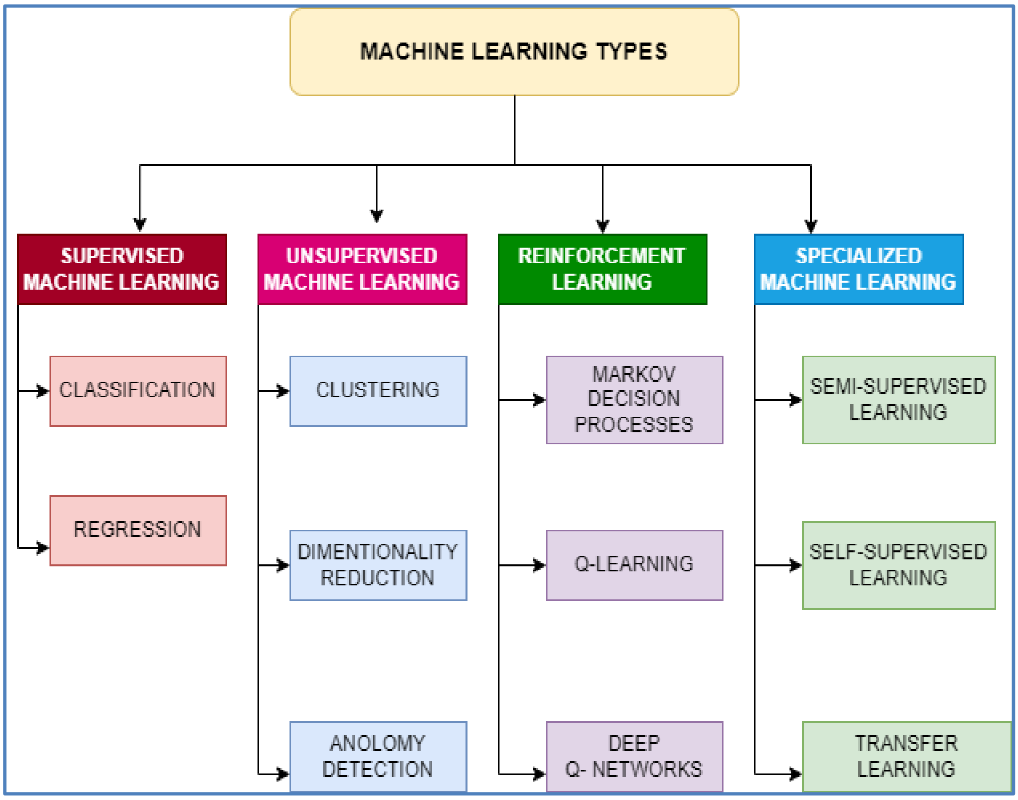

Figure 1.3: Machine Learning and Its Types

1.8.1 Supervised Learning

In supervised learning, we train a model using a labelled dataset, where each training example is associated with an output label. The objective is for the model to learn how to map inputs to outputs [40] as shown in Figure 1.4. This study uses XGBoost as an example of a supervised learning algorithm. During the classification phase, the XGBoost classifier uses labeled data to achieve higher accuracy in theft detection. XGBoost categorizes instances of theft and non-theft using features obtained from LSTM results. Research on electricity theft detection suggests that unbalanced datasets, such as those with significantly fewer theft instances than non-theft cases, may lead to overfitting in supervised algorithms[40].

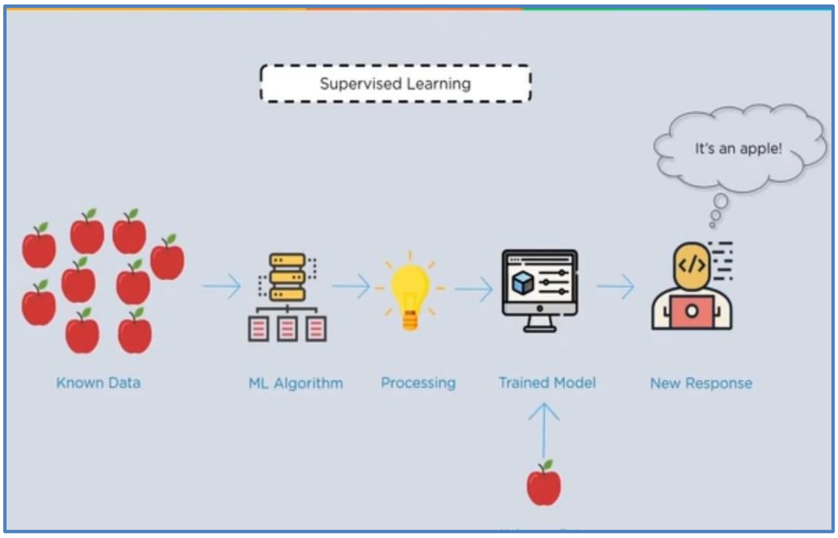

Figure 1.4: Working of Supervised Machine Learning Algorithms

1.8.2 Unsupervised Learning

Unsupervised learning explores hidden patterns or basic patterns in unlabeled data as shown in Figure 1.5. This study employs an unsupervised learning algorithm for clustering or anomaly detection using LSTM. It detects irregularities in the consumption pattern. An unsupervised learning algorithm like k-means clustering and LSTM is used to retrieve patterns from raw data during the pre-processing and anomaly detection phases, respectively.

On smart meters, the unsupervised methods may identify anomalies in power. This allows algorithms like isolation forests to discover suspicious consumption patterns that may indicate theft. [41].

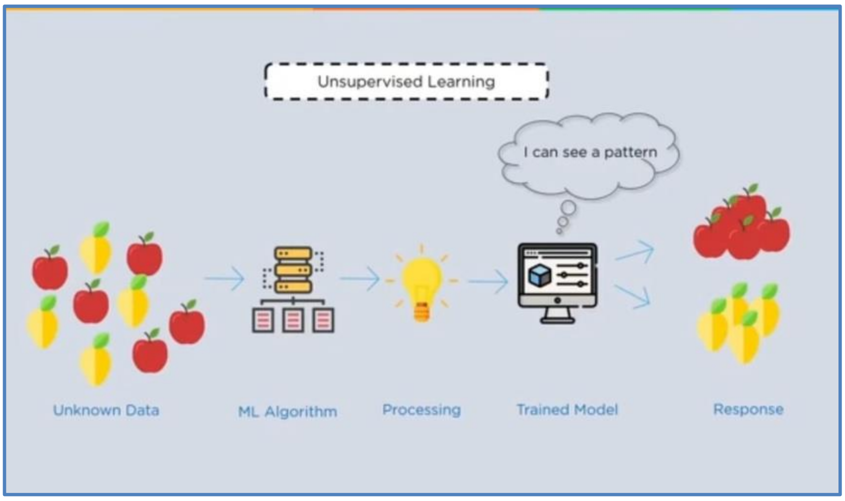

Figure 1.5: Working of Unsupervised Machine Learning Algorithms

1.8.3 Reinforcement Learning

In Reinforcement learning (RL), an agent learns by interacting with an environment, by receiving rewards or gets penalties on performed actions. The aim is to attain maximum cumulative rewards and minimises cumulative penalties over time [42] as shown in Figure 1.6. Reinforcement learning (RL) is less applicable in the proposed model, as the problem does not involve sequential decision-making or dynamic interaction with an environment. As it requires a clear reward signal, which is difficult to define for static datasets or theft detection tasks. It is computationally expensive and complex to implement it for this study. RL could theoretically be used in broader energy distribution optimization but is not well-suited for theft detection, as theft labelling is static and does not involve dynamic exploration [42].

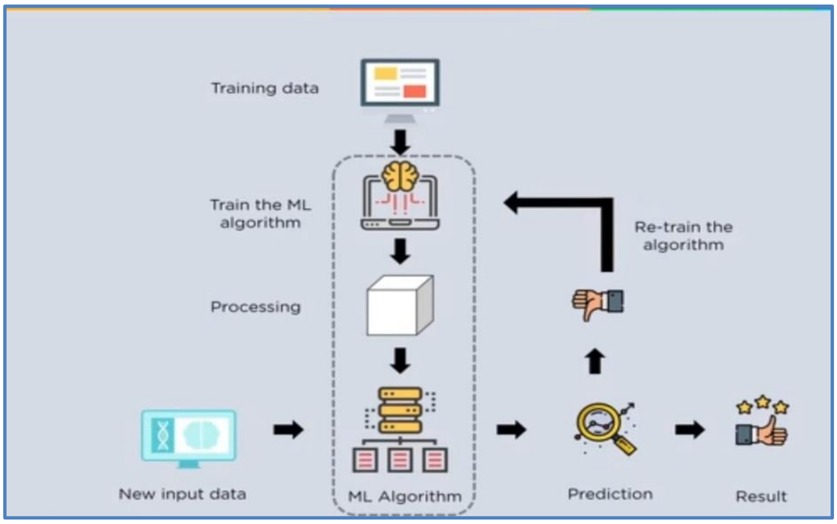

Figure 1.6: Working of Reinforcement Learning

1.8.4 Specialized Forms of Machine Learning

Specialized machine learning algorithms like semi-supervised learning mix less labeled data with more unlabeled data during training. This approach proves advantageous when the new task involves a limited number of data points [43]. This research includes an explanation of the applications of algorithms used within the context of energy theft detection.

1.9 Challenges of Machine Learning in Electricity Energy Theft

The deployment of machine learning (ML) to address electricity energy theft presents a promising avenue for utility companies, yet it is not without significant hurdles. Despite the potential for enhanced detection and prevention capabilities, several challenges must be overcome to ensure effective and practical implementation [44]. Key challenges include data quality and availability, the complexity of theft patterns, scalability, interpretability, and regulatory and privacy concerns.

- **Data Quality and Availability:** One of the primary challenges in applying ML to detect electricity theft is the quality and availability of data. Accurate and comprehensive datasets are crucial for training effective ML models. However, utility companies often face issues such as incomplete, noisy, or biased data. Smart meters, which provide detailed consumption data, may not be installed universally, leading to gaps in data coverage. Moreover, historical data on theft may be scarce or not well-documented, making it difficult to train supervised learning models effectively [44].
- **Complexity of Theft Patterns:** Electricity theft can occur in various forms, ranging from meter tampering and bypassing to sophisticated cyber attacks on the grid. These diverse methods create complex and subtle patterns that are challenging for ML models to detect. Traditional algorithms might struggle with the non-linear and high-dimensional nature of these patterns. Advanced techniques like deep learning can capture such complexities but require large amounts of labelled data and substantial computational resources [45].
- **Scalability:** As utility companies expand their infrastructure and increase the number of smart meters and sensors, the volume of data generated grows exponentially. ML models must scale efficiently to handle this vast amount of data in real time. Ensuring that ML systems can process and analyze data from millions of devices without significant delays or performance degradation is a significant technical challenge [45].
- **Interpretability:** Many advanced ML models, particularly deep learning models, are often considered "black boxes" due to their complexity and lack of transparency. Utility companies and regulators require interpretable and explainable models to trust and act on the predictions made by these systems. Interpreting why a model flagged certain activities as suspicious is essential for decision-making and for addressing customer queries and legal requirements.
- **Regulatory and Privacy Concerns:** The deployment of ML in detecting electricity theft raises important regulatory and privacy issues. Collecting and analyzing detailed consumption data can lead to concerns about user privacy and data security. Utility companies must navigate strict regulatory frameworks that govern data usage and ensure compliance with privacy laws. Balancing the need

for effective theft detection with the protection of customer privacy is a delicate task [45].

- **Adaptability to Evolving Tactics:** As ML models become more effective at detecting theft; perpetrators may develop new and more sophisticated methods to evade detection. ML systems need to be adaptable and continuously updated to keep up with evolving theft tactics. This requires ongoing investment in research and development, as well as a robust mechanism for integrating new knowledge and insights into existing models [46].
- **Integration with Existing Systems:** Integrating ML-based theft detection systems with existing utility infrastructure can be challenging. Legacy systems may not be designed to handle the real-time data processing and analytics required by ML models. Ensuring seamless integration and interoperability with current operational systems, while maintaining reliability and performance, is a significant hurdle [46].
- **Cost and Resource Constraints:** Integrating machine learning solutions for detecting electricity theft can be expensive, demanding substantial investments in technology, infrastructure, and trained personnel. For many utility companies, especially smaller ones, financial and resource constraints can be a major barrier. Additionally, the maintenance and continuous improvement of ML models necessitate ongoing expenditures [46].

1.10 Statement of problem

Electricity theft poses a significant challenge in utility management due to its financial implications and operational complexities. High rates of false positives, primarily due to insufficient input features and environmental variations, hinder existing methodologies for detecting electricity theft using machine learning models. Furthermore, manual feature engineering in current approaches often leads to poor generalization outcomes. This research proposes a novel approach called the KPLX integrated detection model using machine learning models to address the challenge of false positives in electricity theft detection, integrating advanced techniques such as K-means clustering and anomaly detection using LSTM and XGBoost for classification of theft and non-theft cases. By focusing on multidimensional datasets and consumption patterns, the proposed model seeks to improve detection accuracy and outperform existing methodologies. Through experimentation and validation using real-world datasets, the research aims to demonstrate the efficacy of the proposed approach in detecting electricity theft in challenging environments. By enhancing the accuracy of theft detection and reducing false positives, the research aims to provide utility providers with a more robust solution for managing electricity theft and ensuring the integrity of their systems.

1.11 Research Objectives

The research aims to address the issue of electricity theft in smart meters which presents significant challenges for utility providers worldwide. To achieve this goal effectively, the following objectives have been defined:

- To study and analyze the existing energy theft detection techniques for the conventional and smart meters.
- To develop an ensemble learning-based model for energy theft detection for mitigating the anomalies of false positives.
- To compare and validate the proposed model with the conventional energy theft detection techniques.

1.12 Scope of the Study

The scope of this research is to develop a machine-learning model specifically tailored for detecting instances of electricity theft. We designed the model to take into account environmental variations and consumption patterns, which are known to influence detection accuracy. We validate the model's effectiveness using real-world datasets from reputable sources like the State Grid Corporation of China and KPDCL. We apply various data preprocessing techniques to enhance the model's performance. However, this study focuses solely on the development of a machine learning-based detection model, excluding other methodologies such as rule-based systems or state-based anomaly detection algorithms.

1.13 Organization of the Thesis

This thesis is divided into five chapters.

Chapter 1 Introduction provides an overview of the importance of energy sustainability and introduces the concept of smart meters. It also outlines the significance of addressing energy theft in smart metering systems and presents the research objectives, statement of the problem, and scope of the study.

Chapter 2 Literature Review examines existing approaches to energy theft detection, focusing on their strengths, limitations, and current challenges. It also discusses the need for energy theft detection models and explores the potential benefits of implementing effective detection systems.

Chapter 3 Methodology outlines the research methodology, including data collection procedures, algorithm selection criteria, and model development techniques. It describes the process of designing and developing the energy theft detection model for smart meters.

Chapter 4 Model Development, Results, and Discussions presents the design and implementation of the energy theft detection model. It discusses the selection of algorithms, feature engineering techniques, and model validation procedures. In this chapter research results are presented presents including findings, data analysis, and interpretations. The discussion section interprets the findings of the study and provides insights into the implications for smart metering systems. It also addresses the limitations of the research and suggests avenues for future research.

Chapter 5 Conclusion summarizes the key findings of the study, discussing implications for theory and practice, and suggesting areas for future research.

CHAPTER II

LITERATURE REVIEW

2.1 Introduction

Energy theft poses a significant concern in Advanced Metering Infrastructure (AMI), leading to substantial financial losses annually in both developed and developing countries. A range of studies have proposed methods for detecting energy theft in smart meters and focused on improving energy theft detection within smart metering systems. Researchers explored various methodologies, including the integration of advanced technologies like convolutional neural networks (CNNs) and encryption algorithms. Additionally, researchers investigate the functionality of smart meters beyond traditional energy monitoring, examining their potential for identifying unauthorized activities like tapping on distribution lines. Some studies propose adaptive systems capable of continuous learning to differentiate between legitimate and fraudulent energy usage patterns. This chapter delves into existing literature concerning the methodologies and technologies employed in identifying and preventing electricity theft. Furthermore, it examines the challenges and limitations faced by current systems in accurately detecting fraudulent activities. Additionally, the chapter explores the potential benefits of implementing machine learning algorithms and artificial intelligence to enhance the accuracy and efficiency of energy theft detection. By analyzing the current state of research and technology in this field, we can gain a deeper understanding of the advancements and opportunities for improving energy management and security within smart metering systems. Ultimately, the goal is to create a more resilient and secure energy infrastructure that benefits both consumers and utility providers alike.

2.2 The Conventional Approach to Theft Detection

The conventional approach to energy theft detection for smart meters involves a range of methods. We have used traditional methods like support vector machines (SVM), decision trees (DT), fuzzy C-means clustering, K-nearest neighbor (KNN), fuzzy logic, hierarchical clustering, user profiling, and genetic algorithms, but they often require a smart energy meter and may have accuracy issues [47]. Recent studies have concentrated on data analytics techniques, combining the maximum achievable coefficient of information and clustering techniques to enhance detection accuracy. [48]. These methods are particularly effective in detecting abnormal electrical behaviors and thefts with arbitrary shapes. Building on these improvements, researchers like [49] looked at EDA methods for finding power theft in UK home networks and found that fuzzy C-means clustering (FCM) was the best at finding customers who seemed sketchy. Sensitivity analysis confirmed FCM's superior performance, using metrics like accuracy, geometric truth rate, and AUC..

Researchers such as [50] proposed a combination of a self-organizing map (SOM) neural network with K-means clustering to further enhance detection methods. They introduced an improved algorithm for load curve similarity, demonstrating its effectiveness through simulations. This study highlighted high accuracy in theft detection and recommended further research on integrating various distance metrics and advanced machine learning techniques. Similarly, [51] developed a model to detect electricity theft among consumers, even without historical data. Using GriFTable dLab-D simulations, they generated and shared a Github dataset of theft scenarios, clustered users, and applied machine learning for classification. The model achieved high accuracy, effectively identified new users, and varied theft percentages within clusters. The study noted challenges with contextual data and privacy concerns, suggesting future exploration of privacy-enhanced machine learning approaches.

A range of advanced methods using smart meters have been proposed for detecting electricity theft. Researchers like [52] and researchers like [53] both highlight the effectiveness of anomaly detection techniques, with Barzamini specifically noting the superiority of the PCA method over the peer-to-peer (P2P) method. Similarly, researchers [54] introduced a consumption pattern-based detector, which predicts the customers' normal and fraudulent electricity usage patterns to identify suspicious activity, and [55] in their study provides an overview of machine learning techniques for energy theft detection, emphasizing the need to address the challenges in this field. These studies collectively underscore the potential of advanced methods using smart meters for detecting electricity theft, with anomaly detection and consumption pattern-based techniques showing particular promise.

Researchers like [56] investigated the feasibility of using outlier detection algorithms to enhance the security of AMI and identify electricity theft in consumer energy usage datasets. These algorithms showed effectiveness, though their performance varied with dataset characteristics and scalability concerns. Similarly [57] proposed a novel approach combining the highest achievable information coefficient and Clustering techniques to detect electricity theft. The information coefficient identified correlations between non-technical power losses due to theft and electricity usage patterns, while clustering detected abnormal users among load profiles. Despite data availability and clustering complexity challenges, the method proved effective in numerical experiments on an Irish smart meter dataset. Researchers like [58] developed a feature-engineering framework utilizing a Gradient Boosting Machine (GBM) algorithm for electricity theft detection, achieving superior performance. Meanwhile, researchers like [59] proposed a comprehensive approach within Advanced Metering Infrastructure (AMI), introducing a Modular Detection and Tampering Algorithm (MDTA) for physical tampering and a Unique String Authentication Procedure (USAP) for data hacking detection. These studies collectively aim to enhance the security and reliability of smart grid systems by addressing energy theft challenges through advanced detection and mitigation techniques.

Researchers such as [60], [61], and [62] investigated the use of specific external hardware devices and designs, specialized metering devices, distribution transformers, sensors, and various types of metering devices for energy theft detection (ETD). Additionally, [63] discussed a method utilizing an adapted ammeter device for theft detection on the low-voltage (LV) side of the power network. This approach compares differences in electrical parameters between local and remote devices to identify theft, employing state estimation at the substation level to detect anomalies and electricity theft within a cluster. However, this method faces drawbacks, primarily the high cost of implementing additional devices and the challenges of integrating these devices into the existing system.

Various researchers, including [64], [65], and [66], utilized a game theory approach, which involves strategic interactions between dishonest consumers (electricity thieves) and utilities to achieve a Nash equilibrium, thereby deterring electricity theft.

While the game theory approach is cost-effective, it presents challenges in defining precise functions for each customer and utility company for theft detection.

Researchers such as [67] employed a hybrid methodology that uses network-oriented measurements, such as power flow and voltage measurements, to estimate the state variables of the power system at the low voltage or distribution level using machine learning techniques. These techniques leverage historical data and relevant features to train models capable of detecting theft patterns based on consumer behavior and consumption patterns.

In study [68] researchers introduce a privacy-preserving method using peer-to-peer computing to identify fraudulent users, underscoring the significance of privacy, real-time monitoring, and data analytics in energy theft detection for smart meters. Similarly, Researchers like [69] emphasize the importance of privacy in detection methods, presenting a privacy-preserving approach to identifying manipulated energy generation measurements. Similarly, in a study [70] authors proposed the Unique String Authentication Procedure (USAP) to detect and mitigate energy theft in Smart Metering and Advanced Metering Infrastructure, addressing vulnerabilities by inserting a one-way function into the meter. Meanwhile, authors in their study [71] proposed a Principal Component-based Theft Detection scheme for addressing energy theft in Advanced Metering Infrastructure (AMI), achieving a high detection rate for energy theft attacks but with limitations in applicability to AMI and potential representation issues with real data. Further research is needed to enhance the scalability and robustness of the proposed scheme for implementation in larger utility networks.

In contrast, Researchers like [72] proposed a smart prepaid energy metering system for detecting meter bypassing and tampering, complementing a combination of data mining techniques to detect various types of electricity theft. Additionally, Researchers like [73] present a CNN-LSTM-based system for electricity theft detection, addressing class imbalance with synthetic data but acknowledging limitations related to dataset imbalance and the absence of a time attribute. These studies highlight the importance of using a combination of innovative approaches to effectively combat energy theft in AMI systems. Furthermore, the incorporation of advanced machine learning algorithms and real-time data analysis can enhance the accuracy and efficiency of detection methods. Similarly, Researchers like [89] propose a scheme for energy theft detection with energy privacy preservation in the smart grid, utilizing CNNs and the Paillier algorithm, highlighting the effectiveness of the proposed scheme while emphasizing the need for further evaluation of data privacy measures against cyber threats.

Furthermore, in a study [74] authors introduced a novel feature-engineering framework for theft detection in smart grids, employing clustering and a Genetic Programming algorithm to enhance accuracy. Meanwhile, in the study [75] authors present an efficient algorithm for detecting non-technical loss (NTL) in power distribution networks, achieving optimal detection accuracy while maintaining time efficiency, albeit with limitations in validation and scalability. Similarly, in a study [76] researchers present a novel method for detecting electricity theft in smart meter data streams through anomaly pattern detection, applicable in real situations without relying on previous customer usage records, but facing challenges in normalizing data for certain attack types and collecting illegal power consumption data.

2.3 Critical Review of Machine Learning-Based Approaches

We conduct a thorough examination of machine learning methodologies employed in detecting energy theft within smart metering systems. Our analysis encompasses various machine learning algorithms and their applications in energy theft detection, offering insights into their performance, potential challenges, and implications for enhancing smart metering infrastructure security and efficiency. We also consider the influence of data quality, feature selection, and model interpretability on the overall effectiveness of machine learning-based detection methods. This collaborative strategy ensures the security, reliability, and resilience of smart metering systems against potential attacks, thereby safeguarding the energy grid's integrity and shielding consumers from fraudulent activities. Researchers like [77] scrutinize contemporary energy-theft detection strategies based on smart metering, highlighting their implications for distribution networks' efficiency and security, particularly non-technical losses (NTL). While delving into advancements in artificial intelligence-based detection techniques, it highlights certain constraints, such as proposals focusing solely on partial aspects of the electricity theft issue, and advocates for future methodologies that amalgamate diverse detection techniques for greater efficacy. constraints, Also, in study [78], researchers carefully look at machine learning applications for finding energy theft through smart meter data. They show that smart meters are vulnerable to targeted attacks and stress the problems that still need to be solved in this area. Meanwhile, in a study [79], authors evaluate electricity consumption data for unearthing electricity thefts, lauding their prowess in attaining top-tier performance but recognizing constraints such as the emphasis on specific metering systems and the call for further exploration in detector design and privacy-preserving techniques. Collectively, these studies underscore the imperative of harnessing machine learning methodologies to combat energy theft in smart grids while also advocating for sustained research to tackle the evolving nature of fraudulent activities and augment detection capabilities in this pivotal domain. Additionally, [80] furnishes a panoramic view of modeling techniques for spotting electricity theft within smart grid systems, spotlighting innovative and cost-efficient approaches to curtail non-technical losses. Researchers like [81] assert that machine learning is the quintessential tool for detecting electricity theft, delineating its methodology alongside other detection techniques and proposing a systematic framework for evaluating and juxtaposing detection techniques. Moreover, collaborative research endeavors can culminate in implementing more robust security measures in smart grid systems. By combining machine learning with other detection methods, such as data analytics and anomaly detection, a complete plan can be made to effectively stop electricity theft in smart grid systems. This plan would reveal patterns and outliers in data about energy use that show improper use or tampering. Similarly, researchers like [82] compare different AI-based fraud detection methods and point out their pros and cons. [83] does the same thing but in a more organized way, looking at deep learning methods in smart meter data analytics and planning how to solve problems in the energy supply field and where future research should go. Meanwhile, in a study [84], researchers compare various machine learning methods for unearthing electricity fraud, underscoring the necessity of robust anti-power-theft algorithms and noting the differing efficacy of various method combinations in spotting abnormal power usage. On the other hand, [85] suggests a new way to find energy theft in distribution systems by using a three-phase state estimator based on phasor measurement units. This method works well and is reliable at finding and identifying energy theft, but it needs more testing and only uses data from 5000 consumers. Similarly, in studies [86] and [87], authors introduce a data-driven approach using machine learning, specifically deep neural networks, to detect energy theft in smart grids, overcoming data challenges while acknowledging limitations related to data reliability.

A range of supervised machine-learning algorithms have been proposed for electricity theft detection. In Study [88], the authors introduced a feature-engineered CatBoost model, achieving high accuracy and detection rates. Similarly, in research [89], researchers used a machine learning algorithm to identify suspect customers based on power usage patterns, while [90] compared the performance of decision trees, artificial neural networks, deep artificial neural networks, and AdaBoost, with the latter outperforming the others. Researchers like [91] proposed a system using optical character recognition and the SARIMAX algorithm to monitor electricity consumption and detect theft. These studies collectively demonstrate the potential of supervised machine learning for addressing electricity theft.

Researchers like [92] propose a machine-learning approach to address electricity theft in smart grids, focusing on classifying users into legitimate customers and potential thieves. Various algorithms are employed, aiming to develop a robust classifier to differentiate between genuine and fraudulent activities. Similarly, researchers like [93] investigate the use of ensemble machine learning models for energy theft detection in smart grids. By analyzing consumption patterns, ensemble models are utilized to create predictive models, aiming to reduce both bias and variance. Bagging models, particularly random forests and extra trees, demonstrate superior performance in detecting energy theft, achieving high precision and accuracy.

A multitude of studies have delved into the realm of machine learning for electricity theft detection. Notably, researchers like [94] and [95] showcased remarkable accuracy and detection rates through supervised learning methods. [94]'s feature-engineered CatBoost model notably outshone traditional gradient boosting algorithms. Similarly, [96] and [97] explored ensemble machine learning models, with [96] highlighting the superiority of bagging models, particularly random forest and extra trees. Meanwhile, [97]’s comprehensive approach, utilizing various machine learning techniques, aimed to pinpoint the most effective model for conserving electricity and preventing economic loss. These collective efforts underscore the immense potential of machine learning in tackling the complexities of electricity theft detection.

A smart meter-based solution was suggested in the study [98] that uses an artificial neural network (ANN) classification model to detect suspicious customers with high accuracy and distinguish between real and fraudulent electricity usage. This offers a promising solution for efficient power theft detection similarly, similarly, [99] detects electricity theft in low-voltage networks by using a cubic support vector machine classification algorithm; the model achieves average accuracy and an optimal detection rate. These results demonstrate the effectiveness of analytics and machine learning techniques for enhancing security in low-voltage networks. Researchers like [100] present a system utilizing Smart Meter data and the SARIMAX algorithm. By leveraging OCR, the system extracts data from Smart Meter images for analysis, offering seasonal analysis capabilities and automation of the entire process from image entry to bill generation. The proposed system enhances efficiency, reduces manpower requirements, and offers the potential for further automation with smart meters. A range of studies have explored the use of ensemble and hybrid techniques involving machine learning algorithms for the detection of electricity theft in smart grids. It was backed by researchers like [101], who found that the Leveraging Bagging algorithm and the Adaptive Random Forest base classifier were better than other algorithms in terms of accuracy, precision, AUC, ROC, recall, F-1 score, and kappa statistic. Similarly, researchers like [102] proposed an adaptive stacking ensemble algorithm that combined long short-term memory, a Convolutional neuron network, and a hybrid multi-head attention Convolutional network and used a genetic algorithm to optimize hyperparameters. This algorithm demonstrated superior performance in terms of precision-recall area under the curve and F1-score.

Table 2.1 comprehensively summarizes various research contributions, findings, limitations, improvements, and performance metrics associated with different methods and techniques for detecting electricity theft. Spanning references from 47 to 102, it highlights a broad spectrum of approaches, including machine learning, deep learning, anomaly detection, feature engineering, and privacy-preserving methods.

This comprehensive review indicates significant advancements while also identifying areas needing further research and optimization.

Table 2.1: Comprehensive Summary of Electricity Theft Detection Methods

Ref.	Contribution	Findings	Limitations	Improvements
[47]	Use of SVM, Fuzzy C-means, Fuzzy logic, User profiling, and genetic algorithms	Traditional methods were identified but with low accuracy .	Requires smart energy meters; low accuracy	Combination of MIC and CFSFDP for better accuracy
[48]	Combination of MIC and CFSFDP for detection	Effective in detecting abnormal electrical behaviors.	Data availability and clustering complexity	Integration with advanced machine learning techniques
[49]	Assessment of EDA techniques	FCM is most effective for identifying suspicious consumers	Limited to specific metrics and datasets	Sensitivity analysis confirmed superior performance
[50]	SOM neural network with K-means clustering	High accuracy in theft detection	Challenges with integrating various distance metrics	Further research on distance metrics and machine learning integration
[51]	Model to detect theft without historical data	High accuracy in identifying new users and thief percentages	Contextual data and privacy concerns	Exploration of privacy-enhanced machine learning
[52]	Anomaly detection techniques	PCA method superior to the P2P method	Limited scalability, requires extensive historical data	Enhancements for real-time analysis and reduced computational load
[53]	Anomaly detection techniques	Highlighted effectiveness	High computational cost	Real-time detection and reduced computational requirements
[54]	Consumption pattern-based detector	Uses predictability of normal and malicious consumption	Dependence on historical data	Incorporation of real-time analytics

		patterns		
[55]	Overview of machine learning techniques for energy theft detection	Emphasizes the need to address challenges in this field	Not specified	Not specified
[56]	Outlier detection algorithms	Effective in enhancing the security of AMI	Performance varied with dataset characteristics and scalability concerns	Scalability improvements and real-time processing
[57]	MIC and CFSFDP combination for detection	Effective in detecting electricity theft	Data availability and clustering complexity challenges	Integration with more advanced machine learning models
[58]	Feature-engineering framework with Gradient Boosting Machine (GBM)	Achieved superior performance	Complexity in feature engineering	Automated feature selection and real-time analysis
[59]	Modular Detection and Tampering Algorithm (MDTA) and Unique String Authentication Procedure (USAP)	Enhances security and reliability of smart grid systems	High implementation cost	Cost reduction and ease of deployment
[60] **[61]** **[62]**	Utilization of specific external hardware devices and designs	Effective for ETD	High cost and installation challenges	Cost-effective solutions and simplified installation
[63]	Adapted ammeter device for theft detection	Focuses on comparing electrical parameters between local and remote devices	High cost and installation challenges	Simplification and cost reduction of devices
[64] **[65]** **[66]**	Game theory approach	Cost-effective in deterring theft	Challenges in establishing precise functions	Enhanced function precision and real-time capabilities
[67]	Network-oriented	Estimates state	Complexity	Simplified models

	measurements using machine learning techniques	variables of the power system	in implementation	and integration with existing systems
[68]	Privacy-preserving method using peer-to-peer computing	Effective in identifying fraudulent users	Computationally intensive	Optimization for computational efficiency
[69]	Privacy-preserving approach for detecting manipulated energy generation measurements	Emphasizes the importance of privacy	Requires complex cryptographic methods	Simplification and efficiency improvements
[70]	Unique String Authentication Procedure (USAP)	Addresses vulnerabilities in Smart Metering and AMI	Implementation complexity	Simplified implementation and cost reduction
[71]	Principal Component-based Theft Detection Scheme	High detection rate for energy theft attacks	Applicability to AMI and representation issues with real data	Enhancing scalability and robustness
[72]	Smart prepaid energy metering system	Detects meter bypassing and tampering	Implementation complexity	Simplified design and integration with existing systems
[73]	CNN-LSTM-based system	Addresses class imbalance with synthetic data	Dataset imbalance and absence of a time attribute	Real-time analysis and handling of time-series data
[74]	Feature-engineering framework using Genetic Programming algorithm	Enhances accuracy	Complexity in feature engineering	Automation and real-time capabilities
[75]	Efficient algorithm for detecting non-technical loss (NTL)	Optimal detection accuracy while maintaining time efficiency	Validation and scalability limitations	Validation of diverse datasets and scalability improvements
[76]	Anomaly pattern detection method	Applicable in real situations without relying on previous	Normalizing data for certain attack types and	Enhanced normalization techniques

		customer usage records	collecting illegal power consumption data	
[77]	Contemporary energy-theft detection strategies	Highlights advancements and constraints in AI-based detection techniques	Focus on partial aspects of the electricity theft issue	Advocates for methodologies amalgamating diverse detection techniques
[78]	Machine learning applications in detecting energy theft	Flags the susceptibility of smart meters to targeted assaults	Unresolved challenges	Enhanced security measures and real-time capabilities
[79]	Data-driven methods for electricity fraud detection	Lauds top-tier performance	Emphasis on specific metering systems	Calls for further exploration in detector design and privacy-preserving techniques
[80]	Modeling techniques for spotting electricity theft	Innovative and cost-efficient approaches	Not specified	Not specified
[81]	Machine learning for detecting electricity theft	Provides a systematic framework for evaluating and juxtaposing detection techniques	Not specified	Collaborative research for more robust security measures
[82]	Comparative scrutiny of various fraud detection methods leveraging AI	Highlights strengths and weaknesses	Varying efficacy of different methods	Hybrid approaches combine best practices
[83]	Systematic exploration of deep learning methods in smart meter data analytics	Tackles challenges in the energy supply domain	Computational complexity	Optimization for efficiency and real-time analysis
[84]	Comparison of various machine learning methods for electricity fraud detection	Underscores the necessity of robust anti-power-theft algorithms	Differing efficacy of various method combinations	Not specified

[85]	Three-phase state estimator using phasor measurement units	Effective and robust in detecting and identifying energy theft	Limitations in testing and reliance on data from 5000 consumers	Broader validation and scalability improvements
[86] **[87]**	Data-driven approaches using deep neural networks	Overcomes data challenges	Data reliability limitations	Improved data validation and noise reduction
[88]	Feature-engineered CatBoost model	High accuracy and detection rates	Requires extensive feature engineering	Automated feature selection
[89]	A machine learning algorithm for identifying suspect customers	Effective in identifying suspect customers	Requires comprehensive data collection	Improved data integration and real-time analysis
[90]	Comparison of decision tree, ANN, deep ANN, and AdaBoost	AdaBoost outperforms others	Computational complexity for deep models	Optimization for efficiency and scalability
[91]	System using OCR and SARIMAX algorithm	Monitors electricity consumption and detects theft	Not specified	Real-time analysis and integration with other data sources
[92]	machine learning approach for classifying users	Classifies users into legitimate customers and potential thieves	Requires extensive historical data	Incorporation of real-time data and enhanced features
[93]	Ensemble machine learning models for energy theft detection	High precision and accuracy using bagging models	Computational complexity	Simplified ensemble techniques
[94] **[95]**	Supervised learning methods for electricity theft detection	High accuracy and detection rates	Data dependency and computational cost	Optimization for real-time detection
[96]	Ensemble machine learning model	High detection rates	Not specified	Integration with other data sources for enhanced detection
[97]	classifier system	High detection	Not specified	Real-time analysis

	using machine learning for theft detection	rates		and better scalability
[98]	CNN-LSTM deep learning-based model	Addresses imbalances with synthetic data	Dataset imbalance, lack of time attribute	Real-time capabilities and handling time-series data
[99]	Comparison of machine learning methods	Highlights the effectiveness of certain methods	Applicability to different datasets	Exploration of hybrid methods for better detection
[100]	machine learning approach for energy fraud detection	High detection rates	Not specified	Not specified
[101]	Supervised learning techniques for energy theft detection	High accuracy in identifying fraudulent activities	Data availability and model complexity	Optimization and integration with real-time data sources
[102]	Machine-learning-based classifiers	Effective in distinguishing between normal and fraudulent consumption	Not specified	Real-time analysis and integration with various data sources

In summary, recent research offers valuable insights into energy theft detection using data-driven techniques. It encompasses a comprehensive review of methodologies, technologies, and challenges encountered in combating electricity theft in smart metering systems. Researchers explore diverse approaches, including hardware devices, game theory, and machine learning techniques like convolutional neural networks (CNNs) and genetic programming algorithms. Researchers also examine the efficacy of privacy-preserving methods and anomaly detection schemes in identifying fraudulent activities while safeguarding consumer privacy. Furthermore, the study scrutinizes machine learning-based approaches, assessing their effectiveness, limitations, and implications for enhancing smart metering infrastructure security and efficiency. We advocate for collaborative efforts and continuous research to address evolving threats and enhance detection capabilities, ensuring the integrity of smart grid systems and safeguarding consumers from fraudulent activities.

2.4 Research Gaps and Limitations in Current Approaches

This section examines the existing research gaps and limitations present in the methodologies and technologies currently employed for identifying and preventing electricity theft. This section will delve into various aspects where existing approaches fall short, providing a critical analysis of their limitations and areas for improvement.

Scalability and Applicability: While several methods have been proposed for detecting energy theft, many of them face challenges in scaling up for implementation in larger utility networks. Researchers [12], for example, introduced a Principal Component-based Theft Detection scheme with a high detection rate but noted limitations in its scalability to larger systems. There's a need for approaches that can be easily implemented across diverse utility networks without sacrificing accuracy or efficiency.

Data Privacy Concerns: Several studies emphasize the importance of privacy-preserving methods in energy theft detection, such as those proposed by Researchers [9] and [10]. However, there's a lack of comprehensive evaluation of these privacy-preserving techniques against potential cyber threats. Future research should focus on developing robust privacy-preserving methods while considering potential vulnerabilities to ensure the security of sensitive consumer data.

Dataset Imbalance and Generalization: Some approaches, like the CNN-LSTM-based system proposed by researcher [14], address class imbalance with synthetic data but acknowledge limitations related to dataset imbalance and the absence of a time attribute. This highlights the challenge of generalizing detection methods across different datasets and environments. Future research should aim to develop approaches that are robust to dataset variations and can generalize well across different scenarios.

Model Interpretability: The effectiveness of machine learning-based approaches for energy theft detection relies heavily on the interpretability of the models used. While these models may achieve high detection rates, understanding how they make decisions is crucial for trust and transparency. Many studies do not thoroughly address the interpretability of their models, indicating a gap in understanding the inner workings of these detection systems.

Integration of Diverse Detection Techniques: While machine learning methods show promise in detecting energy theft, there's a need to integrate diverse detection techniques for heightened efficacy, as emphasized by [19]. Combining machine learning with other approaches, such as data analytics and anomaly detection, can provide a more holistic approach to combating energy theft. Future research should focus on developing frameworks that seamlessly integrate these techniques to enhance detection capabilities.

Robustness against Adversarial Attacks: With the increasing sophistication of cyber threats, there's a need to ensure that energy theft detection systems are robust against adversarial attacks. While some studies mention the susceptibility of smart meters to targeted assaults, there's a limited exploration of their robustness against such attacks. Future research should explore techniques for detecting and mitigating adversarial attacks in smart metering systems.

By addressing these gaps, researchers can pave the way for the development of more robust strategies to combat electricity theft and enhance the overall security of energy systems.

2.5 Proposed Research Approach and Contributions

Addressing these research gaps and limitations is pivotal for advancing energy theft detection and bolstering the security and reliability of smart metering systems. Future studies should prioritize resolving challenges such as dataset imbalance, data privacy concerns, and cyber threats to augment the effectiveness of detection systems. Additionally, leveraging advanced machine learning techniques like reinforcement learning and short term usage prediction could significantly improve the accuracy and efficiency of energy theft detection within smart grid environments.

In our research, we address the identified limitations and research gaps in current approaches to energy theft detection. Specifically, we designed a novel model for energy theft detection using machine learning techniques that have not been extensively explored in existing literature. By leveraging innovative machine learning algorithms and methodologies that are not currently utilized, we overcome scalability challenges, reduction in false positive observations, and develop robust privacy-preserving methods. Our model offers comprehensive coverage of energy theft aspects, ensuring effectiveness across different metering systems and real-world utility networks.

CHAPTER III

NOVEL APPROACH FOR DETECTING ELECTRICITY THEFT

3.1 Introduction

Electricity theft poses a significant challenge for utilities worldwide, leading to revenue losses and potential safety hazards. Traditional methods of detecting theft often struggle to distinguish between genuine anomalies in consumption patterns and fraudulent activities, resulting in high false positive rates and inefficient resource allocation. In this chapter, we present a novel approach for detecting electricity theft using machine learning models tailored to mitigate the limitations of existing methodologies. Our approach integrates clustering, feature extraction, and ensemble classification techniques to achieve robust and accurate detection results.

3.2 Challenges in Electricity Theft Detection in Smart Meters

Detecting electricity theft is a multifaceted task with complexities and nuances that pose significant challenges for utilities and researchers alike. The prevalence of false positives, which mistakenly flag legitimate variations in consumption patterns as instances of theft, is one of the primary challenges. These false alarms not only undermine the credibility of the detection system but also impose unnecessary burdens on utility resources, leading to inefficient investigations and potential customer dissatisfaction [103].

The dynamic nature of consumer behavior and the influence of environmental factors on electricity consumption patterns exacerbate the issue of false positives. For instance, fluctuations in weather conditions, seasonal variations, and cultural events such as holidays and festivals can lead to temporary spikes or dips in energy usage that may resemble anomalous behaviour [104]. Moreover, changes in household demographics, appliance usage, and lifestyle habits further contribute to the complexity of distinguishing between genuine anomalies and fraudulent activities.

Another significant challenge in electricity theft detection is the influence of environmental factors on consumption patterns. Environmental variations such as temperature extremes, humidity levels, and daylight hours can significantly impact energy demand and usage patterns [105]. Moreover, changes in household demographics, appliance usage, and lifestyle habits further contribute to the complexity of distinguishing between genuine anomalies and fraudulent activities.

Another significant challenge in electricity theft detection is the influence of environmental factors on consumption patterns. Environmental variations such as temperature extremes, humidity levels, and daylight hours can significantly impact energy demand and usage patterns. For example, during periods of extreme heat or cold, consumers may resort to energy-intensive heating, ventilation, and air conditioning (HVAC) systems, leading to higher-than-normal electricity consumption. Similarly, fluctuations in natural lighting conditions can affect the use of lighting appliances, further increasing consumption behavior.

Furthermore, the addition of new energy sources, such as renewable energy based on solar energy and other supplementary electricity sources, adds to the detection methods' complexity. These distributed generation technologies not only alter the traditional flow of electricity within the grid but also create new opportunities for fraudulent activities, such as reverse power flow manipulation and unauthorized grid connections. Detecting theft in the presence of distributed generation requires innovative approaches that leverage advanced machine learning techniques, robust feature engineering, and comprehensive data analysis to achieve accurate and reliable detection results. [106].

3.3 Methodological Framework

The methodology proposed for detecting electricity theft in smart meters comprises a structured framework that integrates clustering, feature extraction, and ensemble classification techniques. This section provides an in-depth overview of the framework, elucidating its key components and their interplay in the detection process. The workflow is divided into four main phases: Data Pre-processing, Anomaly Detection, Feature Engineering, and Model Training and Evaluation [107]. The details of each phase are described as follows:

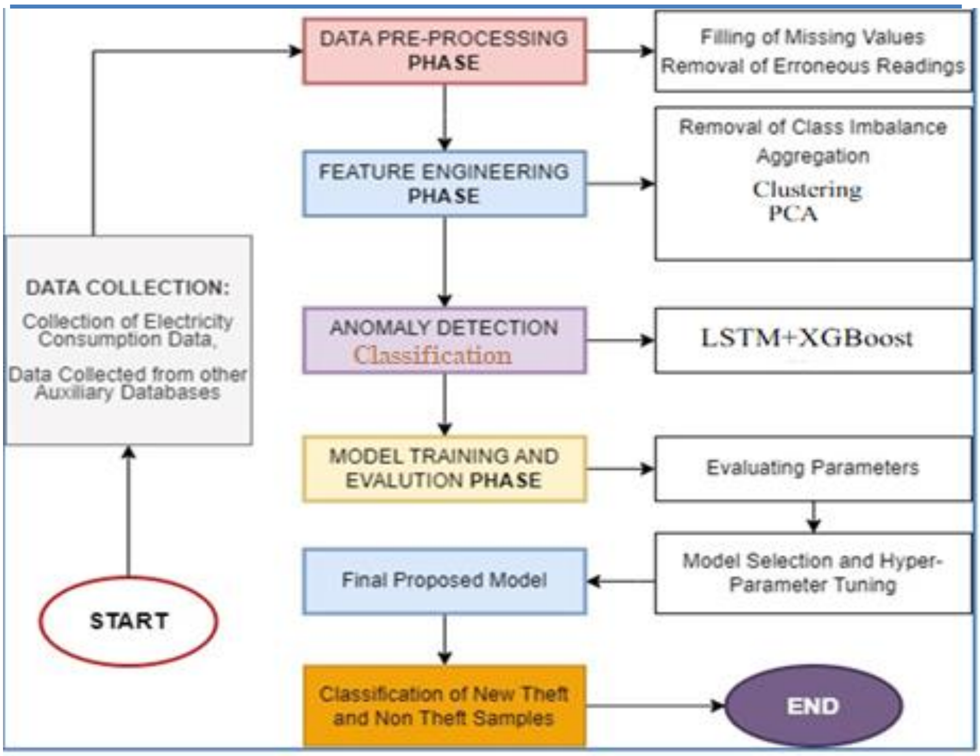
DATA PRE-PROCESSING PHASE
Filling of Missing Values
Removal of Erroneous Readings
FEATURE ENGINEERING PHASE
Removal of Class Imbalance
Aggregation
Clustering
PCA
DATA COLLECTION:
Collection of Electricity Consumption Data,
Data Collected from other Auxiliary Databases
ANOMALY DETECTION
Classification
LSTM+XGBoost
MODEL TRAINING AND EVALUTION PHASE
Evaluating Parameters
Model Selection and Hyper-Parameter Tuning
Final Proposed Model
START
Classification of New Theft and Non Theft Samples
END

Figure 3.1: "Framework of proposed model for the detection of electricity theft in Smart Meters"

3.3.1 Data Collection and Data Pre-Processing Phase

In this research, we utilized two distinct datasets. The SGCC (State Grid Corporation of China) dataset was employed to train, test, and develop the model named as KPLX integrated detection model. The KPDCL (Kashmir Power Distribution Corporation Limited) dataset was used to evaluate and test the performance and generalizability of the model [108]. Auxiliary weather data is also utilized from local MET weather databases.

SGCC Dataset: The SGCC dataset provides comprehensive information on electricity consumption, power parameters, and consumer profiles across various regions in China. It spans from January 2014 to October 2016 and includes 42,372 records, with 3,615 instances indicating abnormal consumption (potential electricity theft) and 38,757 instances of normal consumption. This dataset, collected at 30-minute intervals, features attributes such as electricity consumption patterns, power parameters, consumer profiles, tariff agreements, and weather conditions as indicated in Table 3.1

Table 3.1: Insights from SGCC Dataset

Total time of study	**January2014toOctober2016**
Total No. of consumers	42,372
Total No. electricity stealers	3615
Total No. of genuine consumers	38,757

To ensure precision in identifying electricity theft, factors such as adverse temperatures and power availability are considered. The dataset is categorized into "benign" and "malicious" sets. The "benign" set embodies instances of normal electricity consumption behavior, while the "malicious" set comprises potentially fraudulent activities indicating electricity theft. Given the rarity of instances depicting electricity theft, synthetic addition of malicious data is crucial to ensure a balanced dataset for robust model training and evaluation.

- **KPDCL (Kashmir Power Distribution Corporation Limited) Dataset**

The KPDCL dataset was meticulously curated to ensure the anonymity and confidentiality of consumer information. It contains raw electricity consumption data, meter specifications, and anomalies that may suggest electricity theft. By removing private and sensitive information, the dataset was prepared for research purposes, enabling us to assess how effectively our model could detect electricity theft—a complex challenge for power distribution systems globally. Additionally, we supplemented the KPDCL dataset with weather data and information on scheduled power curtailments collected from various online sources.

In this study, we utilized the SGCC dataset to train, test, and validate our KPLX integrated detection model. The comprehensive and detailed records in the SGCC dataset laid a strong foundation for developing a KPLX-integrated detection model. After creating the KPLX integrated detection model on the SGCC dataset, we tested its performance on the new KPDCL dataset to evaluate how effectively the KPLX integrated detection model generalizes to different contexts and identify potential areas for improvement.

By integrating these datasets, we aimed to build a robust model capable of accurately detecting electricity theft in diverse contexts. The combination of real-world data from KPDCL and the extensive, detailed records from SGCC provided a solid foundation for evaluating and enhancing the detection capabilities of our proposed model.

Let the data be represented as matrix as X and Let X $\in$ $R^{n X d}$ be the dataset, where n is the number of consumers, and d is the number of features (e.g., consumption data points over time). Each row $x_i \in R^d$ represents the consumption pattern of the i-th consumer.

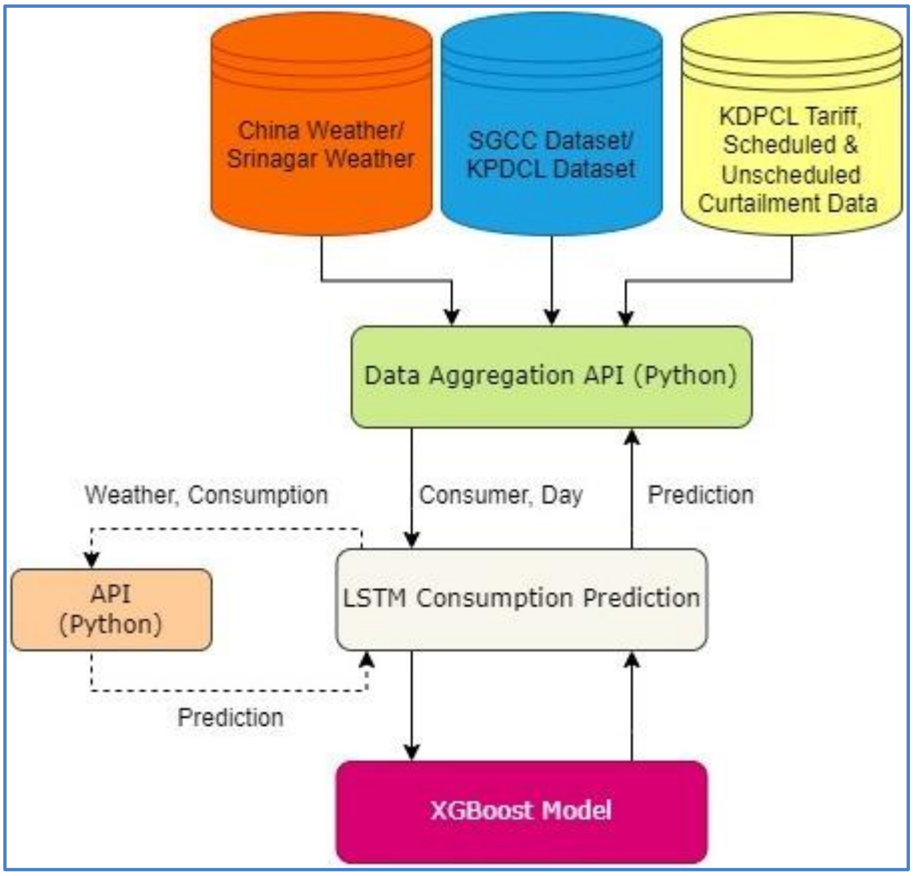
China Weather/ Srinagar Weather
SGCC Dataset/ KPDCL Dataset
KDPCL Tariff, Scheduled & Unscheduled Curtailment Data
Data Aggregation API (Python)
Weather, Consumption
Consumer, Day
Prediction
API (Python)
LSTM Consumption Prediction
Prediction
XGBoost Model

Figure 3.2: Dataflow Diagram of the KPLX Integrated Detection Model

3.3.2. Filling of Missing Values

It is important to address missing values in a dataset as they can cause issues during analysis. This involves filling in the data with appropriate values where data is missing [1]. The SGCC electricity consumption dataset contains missing and incorrect values due to smart meter failures, storage problems, measurement errors, or unreliable transmission. By analyzing and cleaning the dataset, these errors and missing values can be identified and removed. Common techniques for filling in missing values include using the mean, median, and mode. Sometimes more complex methods like interpolation and imputation are used based on other data points in the dataset [110]. In this study, the technique of linear interpolation was utilized to identify and recover the missing values in the dataset [111]. Consequently, the missing values were filled by using Equation (3.1).

$$f\ (x)\ = \begin{cases} \frac{E}{2}, & where\ E = (x_{i-1} + x_{i+1}), x_i \in NaN, but\ x_{i-1}, x_{i+1} \nexists\ NaN. \\ 0, & x_i\ \in NaN, and\ \ x_{i-1}, or\ \ x_{i+1} \in NaN. \\ \boldsymbol{x_i}\ , & x_i \nexists\ \ NaN, \end{cases} \qquad \text{Equation (3.1)}$$

Where x_i represents the value in the electricity usage data over a given time period (such as a day). If x_i is either missing or contains non-numeric characters, we denote it as NaN (NaN indicates a set).

Similarly, we have detected outliers that distort the data, complicating the training process and negatively affecting the final ETD performance due to overfitting. We applied here the "three-sigma rule of thumb" as used in equation[2] to identify and correct these outliers.

the data values are normalized using min-max normalization as shown in equation 3.1a.

$$\mathrm{N}\big(\mathrm{v_i(t)}\big) = \frac{\mathrm{v_i(t)-min(\quad)}}{\mathrm{imax(\quad)-min}(\bar{v})} \qquad \text{Equation (3.1a)}$$

vi (t) is the usage of electricity at time say t , min (v), the usage of minimum electricity, and max(v) is the usage of electricity at the time (t).

Though XGBoost used in our model does not require normalized data for classification and can handle real data without normalization, the study normalized the data to facilitate comparison with other models. Min-max normalization was specifically used, as it ensures effective handling of different features, unlike some machine learning algorithms which are sensitive to data scale.

To normalize the electricity consumption of consumer A in kWh with values ranging from 50 to 300 kWh. The minimum value (X_min) is 50 kWh, and the maximum value (X_max) is 300 kWh. we calculated the normalized data of this feature:

Feature: Daily consumption

Minimum value (X_min): 50 kWh

Maximum value (X_max): 300 kWh

For a specific observation with a daily consumption of 150 kWh, the normalized value would be calculated as:

$$X_{norm} = \frac{X - \mu_x}{\sigma_x}$$

where μ_x are mean and σ_x the standard deviation of the data matrix X, respectively.

For Consumer A,
We have
$X_{norm} = \frac{150-50}{300-50}$ = 0.4. Thus, the normalized daily consumption is 0.4

3.3.3 Feature selection

The feature extraction technique enhances the identification of anomalies in electricity consumption data that may indicate outliers of anomalies related to electricity theft. Analyzing raw data directly can be difficult because of noise and irrelevant information. By utilizing or inputting fewer features into the model, we can concentrate on specific features, reduce noise, eliminate irrelevant features for our research, and concentrate on the most informative aspects of the consumption patterns [112].. In the first step, the electricity and weather data were imported and concatenated in Python.

Subsequently, three different time features were subtracted from the dates of the time series: hour of the day, day of the week and month of the year. The electricity consumption in the timeline, electricity consumption as a function of the outside temperature, electricity consumption as a function of the hour of the day, and electricity consumption as a function of the day of the week.

a. Removal of Class Imbalance

We have implemented six distinct types of synthetic attacks designed to match real-world theft scenarios. These attacks are categorized as Type 1 through Type 6, with each type representing a unique method of electricity theft. By diversifying the attack strategies, we aim to encompass a broader range of theft techniques, ensuring comprehensive coverage and robust detection capabilities across all possible electricity theft scenarios.

Six types of attacks defined as type-1 to type-6 are synthetically added to generate theft data for balancing the dataset and later for identifying the theft:

Type 1: A scaling attack where a smart meter's reading is reduced by constant value i.e. multiplied by a constant factor ($\boldsymbol{\alpha}$t) where alpha is between 0.1 to 0.9.

Type 2: A type of random attack, where in the reading is multiplied by a distinct random value ($\boldsymbol{\alpha}$t) at different intervals.

Type 3: A load-shifting attack, where only half of the actual readings during peak load times are recorded and the full readings during off-peak hours are recorded.

Type 4: A random offset attack where the average electricity consumption readings are multiplied by a random factor (αt).

Type 5: A baseline theft attack where the mean value of the energy usage is recorded only and sent to the utility's control center for billing.

Type 6: A reverse order attack where fraudulent consumers send readings in reverse order, with high readings during off-peak hours and low readings during peak hours.

These attack types aim to simulate various theft scenarios, making it challenging for the machine learning model to identify the theft.

Including the theft attack data in the dataset, enables us to train the model on various theft attacks leading to accurate, reliable, and optimized results.

b. Aggregation

The high-frequency electricity consumption data collected from smart meters is aggregated and organized into hourly, daily, weekly, monthly, and seasonal readings. Figures 3.3 (L) and 3.3 (R) shows the daily load profiles of customers on working days and holidays (scale used 10 :1 kWhr). Similarly, figure 3.4 shows load profiles for residential and non-residential users, and figure 3.5 shows the load profile for customers across the four seasons. This aggregation allowed for a comprehensive analysis of energy consumption patterns over different periods, enabling the identification of anomalies like abnormal increases or decreases in energy usage, that could indicate potential theft.

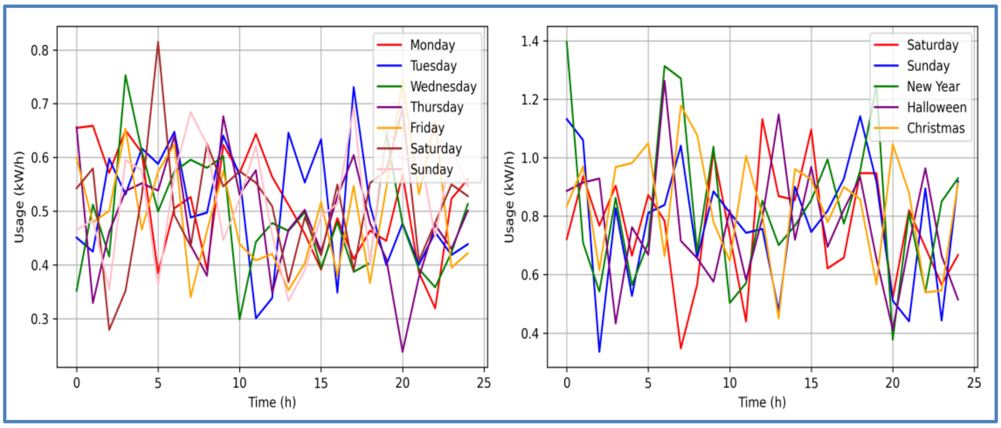

Figure 3.3 (L): Load Profiles (daily) for a Customer in Working Days

Figure 3.3 (R): Load Profiles (daily) for a Customer in Holidays

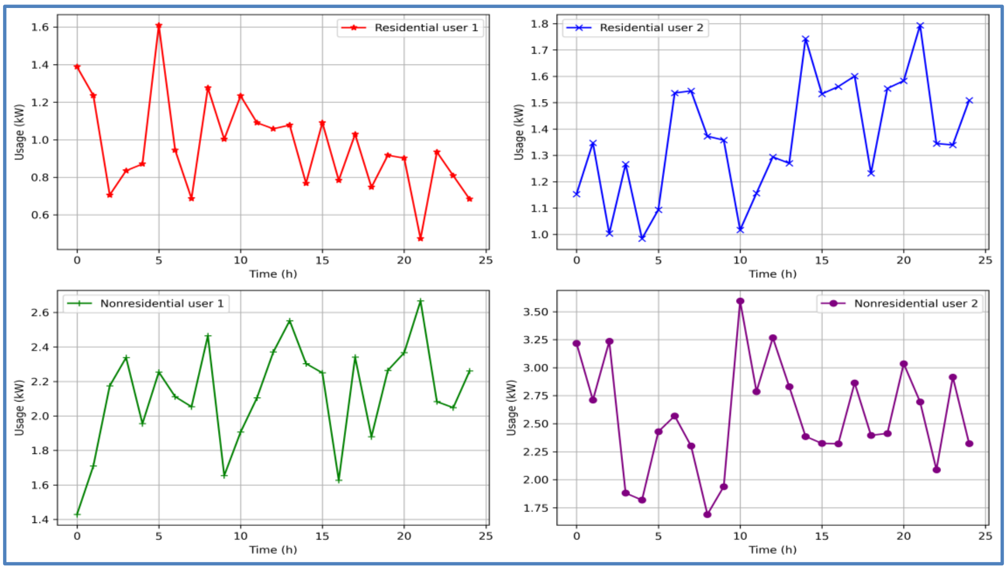

Figure 3.4: Load Profiles (daily) for Residential and Non-residential Users

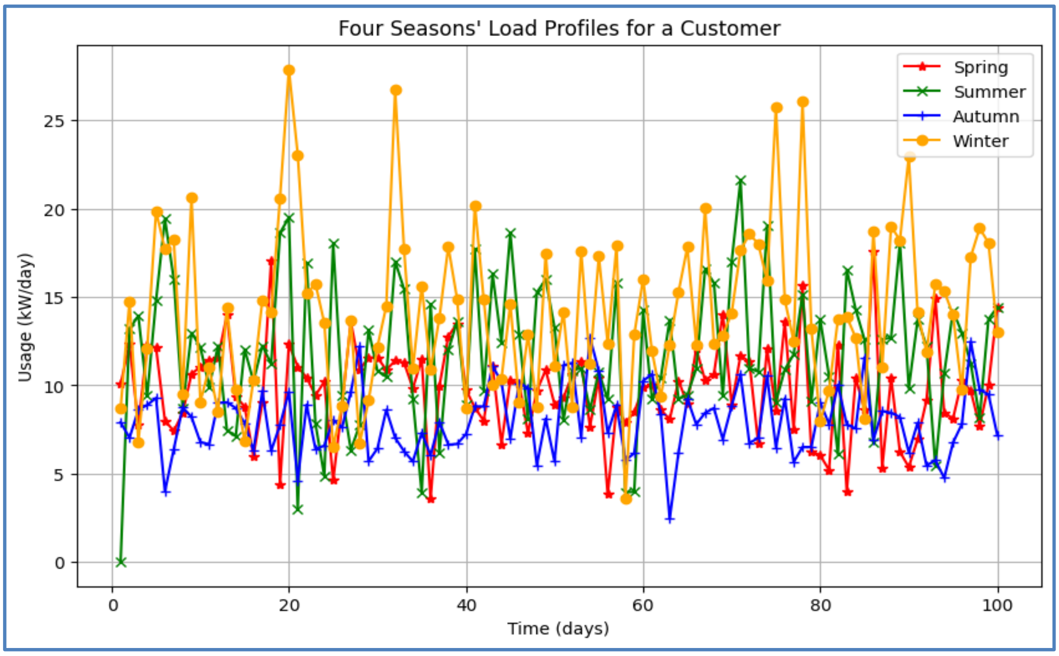

Figure 3.5: Four Season's Load Profile for a Customer

c. Resampling

In order to handle high-frequency data and effects of random varaiation, we resampled the data from 15-minute readings to hourly readings. This allowed us to remove out short-term fluctuations and concentrate on longer-term consumption trends. Converting the data to a lower frequency enabled us to identify significant deviations that could indicate irregularities or tampering [113].

d. Differencing

The data was adjusted using differencing to the aggregated or resampled data to remove trends (seasonal) and make the time series stationary. This technique involves subtracting the previous observation from the current one to emphasize changes in consumption that deviate from expected patterns. By differencing the data, we identify unusual consumption behaviors, such as abrupt increases or decreases in energy usage that do not align with typical household patterns [114].

Through the systematic application of aggregation, resampling, and differencing techniques, we improved our ability to analyze energy consumption data and identify potential anomalies.

e. Principal Component Analysis (PCA)

In this research, we applied Principal Component Analysis (PCA) for dimension reduction in the electricity consumption dataset. PCA is a type of statistical technique applied to decrease the total number of dimensions in a dataset while retaining the majority of the variance in the data. The main objective of this technique is to transform high dimensionality data into less dimensional space while retaining the most of the original information. PCA accomplishes this by identifying the primary constituents. The primary constituents, known as principal components are orthogonal vectors that represent the directions of maximum variance in the dataset [115, 116]. By using PCA, we enhanced the detection accuracy in our proposed KPLX model. Figure 3.6 illustrates the graphical depiction of PCA applied to electricity theft detection model.

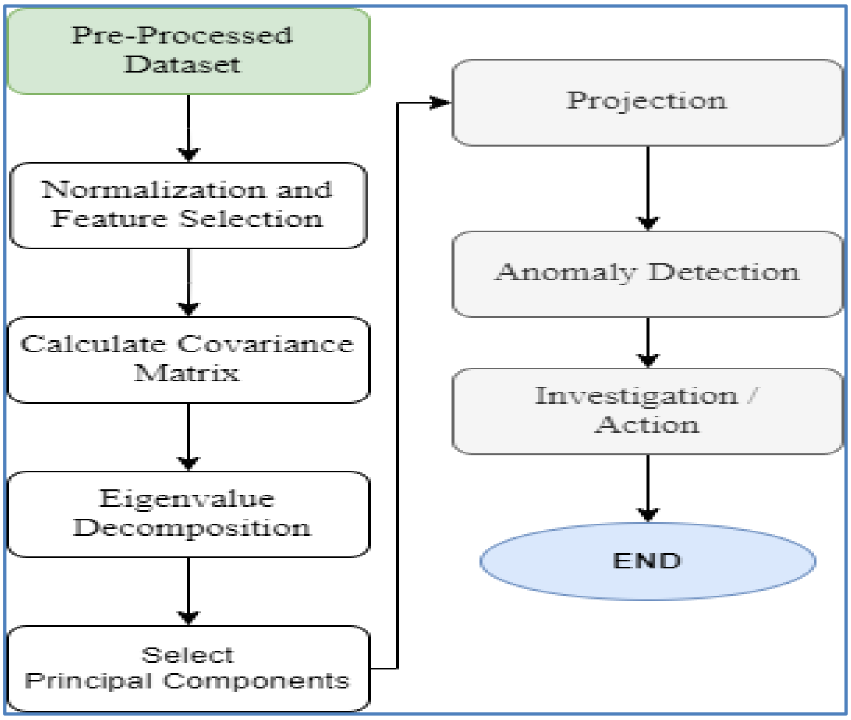

Figure 3.6: PCA for energy theft detection in smart meters.

This dataset has high-dimensions consisting of 24 hourly readings for each monitored consumer. By using PCA, we were able to create a feature set with fewer dimensions. Some features used to understand the consumption pattern are average, standard deviation, kurtosis, energy, chaos, skewness, and periodicity. After initially resampling the samples on an hourly basis, we further resampled on a daily, weekly, monthly, and seasonal basis. The daily segment was divided into four specific periods night time, morning time, day time, and evening time. Clusters discussed in next segment were created on the relative average power for each period. The characteristics were assessed using non-linearity, skewness, trend, serial correlation, kurtosis, self-similarity, seasonality, chaotic behavior, and periodicity

3.3.4 Clustering and Anomaly Detection

In this phase, the identification of an optimal number of clusters (k) using the K-means clustering to group consumers on similar consumption patterns. The K-Means algorithm groups consumers with comparable consumption behaviors to differentiate normal variations and theft anomalies in consumption. The goal is to differentiate normal variations from anomalies, to minimize false positives.

Consumers fed from the same electric substation experience similar electricity availability, interruptions due to feeder faults and due to scheduled feeder curtailments (in the case of a demand-supply gap), and unscheduled curtailments (forced curtailment due to unforeseen conditions).

This clustering method involves training the classifier separately on each cluster and then generalizing it on a larger cluster. This helps the classifier to learn the differences between normal and abnormal variations in electricity consumption behavior. The model also recognizes the variation of high consumption followed by electricity restoration after both scheduled and unscheduled outages. This variation in usage pattern is considered normal by the classifier within the cluster and improves in reducing the false positives. Consequently, the relative power consumption of each cluster remains consistent compared to other clusters. The K-Means algorithm maximizes similarity within each cluster, ensuring that the relative power usage within each cluster remains consistent compared to others, further reducing false positives.

a. K-Means Clustering for Consumer Segmentation

The methodology begins with applying clustering algorithms to group consumers based on their electricity consumption patterns. In this study, K-means clustering was chosen for its simplicity and efficiency. K-means divides consumers into clusters, each characterized by a centroid representing the average consumption pattern and each group can be analyzed for irregularities that could indicate electricity theft. By iteratively optimizing centroids to minimize variance within clusters, K-means aims to create internally homogeneous and externally heterogeneous clusters, enabling meaningful segmentation [116]. Figure 3.7 illustrates the flowchart of the K-means clustering algorithm for energy theft detection in smart meters.

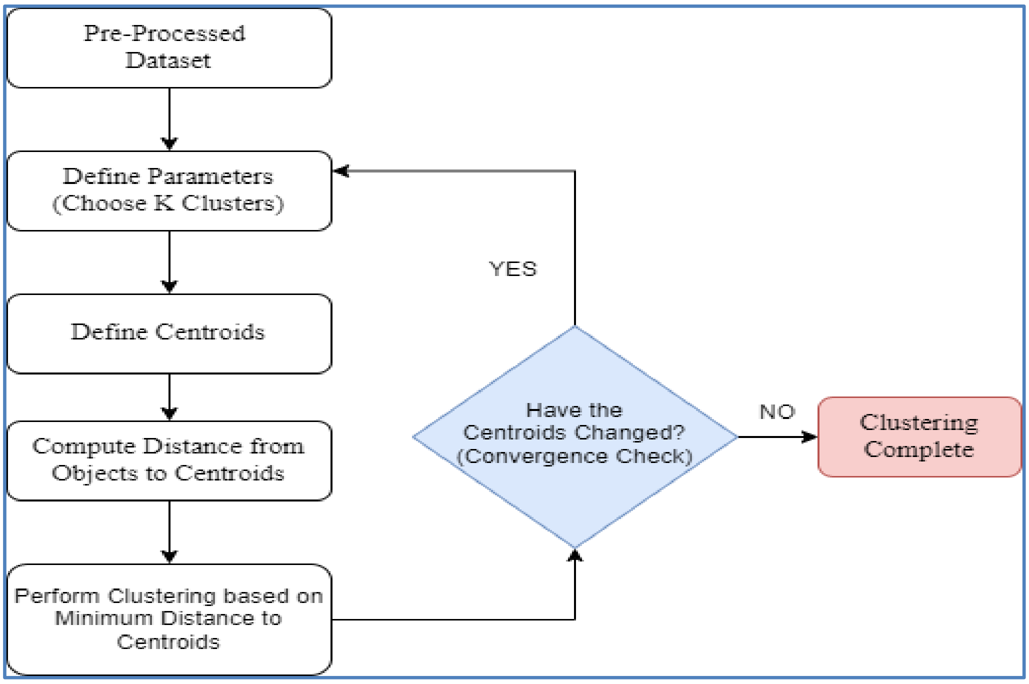
Pre-Processed Dataset
Define Parameters (Choose K Clusters)
Define Centroids
Compute Distance from Objects to Centroids
Perform Clustering based on Minimum Distance to Centroids
Have the Centroids Changed? (Convergence Check)
YES
NO
Clustering Complete

Figure 3.7: Visual representation of the K-means clustering algorithm's flowchart

Daily electricity consumption data from households with smart meters between 2014 and 2016 underwent K-means clustering. Testing clusters from 3 to 12 revealed 10 optimal clusters with distinct consumption patterns, informing targeted policies and energy reduction strategies. Households were assigned to clusters based on location, dwelling type, occupancy, job, income, and appliance details including heating/cooling wattage, energy efficiency ratings, and contracted power [117].

Analysis of 500 households indicated higher winter consumption and lower summer consumption, inversely correlated with temperatures (r ≈ -0.80). Most households (87%) had mean daily consumption below 12 kWh. Below is the summary using k-means clustering to group electricity households.

i. Network Daily Load (NDL): It averages all consumers' daily electricity use over a year with 365 daily load profiles. It helps to find a general pattern of the consumer’s usage each day.

ii. Consumer’s Daily Load (CDL):It averages one consumer's daily electricity use over a year. It uses 500 load profiles (one for each consumer’s household). It provides the shape of the daily electricity usage for each consumer.

iii. Consumer Week-daily Load (CWDL): It is same like NDL, but also separates data for each day of a week. It uses 3,500 load profiles (500 consumer’s × 7 days). It smooths out the shape of the daily electricity use for each household by day of the week.

iv. Consumer Weekday and Seasonal Load (CWDSL): It is same Like CDL but also separates data for all the 4 seasons. It uses 14,000 load profiles (500 consumer’s × 7 days × 4 seasons). It smooth’s out the shape of the daily electricity use for each household by day of the week and season.

v. Raw Daily Load (RDL): It uses the raw, un-averaged daily electricity use data. It uses 1,82,500 load profiles (500 household’s × 365 days). The purpose is to show the actual daily variations in electricity use for each household.

K-means clustering is used through the python’s scikit-learn package. The silhouette score helped determine the best "k" value. Silhouette scores range from -1 to 1, where closer to +1 indicates better clustering performance. The clustering algorithm with "k =n $_{opt}$" (optimal number of clusters) was chosen where intra-cluster similarity was high and inter-cluster similarity is low. The approach used focused on normalizing daily profiles and dividing them into four time periods. The time periods chosen were:

1. Night Time: From 11:00 PM to 7:00 AM.
2. Morning: From 7:00 AM to 10:00 AM.
3. Day Time: From 10:00 AM to 4:00 PM.
4. Evening: From 4:00 PM to 11:00 PM.

Each of these periods represents a different part of the day with varying electricity usage patterns. These periods were chosen to capture key segments during the day that are important for anomaly detection, such as evening and morning peaks in electricity usage. Sample aggregation approaches for 500 households included raw daily profiles and average profiles across households for each monitored day, useful for understanding distinct patterns and identifying the anomaly in the daily profiles. The k in k-means was calculated using the silhouette score which measures a similarity of a data point in a cluster compared to its neighbours:

$$\mathrm{s(i)} = \frac{\mathrm{b(i)-a(i)}}{\mathrm{max\{a(i),b(i)\}}} \quad \text{Equation (3.2)}$$

Where ‘a(i) is the average distance from the point iii to the other points in the same cluster, and b(i) is the minimum average distance from the point iii to points in a different cluster’.By using these mathematical formulations, the paper clusters household electricity consumption patterns effectively, enabling the identification of distinct groups based on their usage behaviors.

The **k-means algorithm** works iteratively as follows:

1. **Initialization:** Randomly initialize **k** centroids μ1,μ2,...,μk.
2. **Assignment Step:** Assign each consumer xi to the nearest cluster based on the minimum distance:
3. $\boldsymbol{C_j} = \{x_i : d(x_i, \mu_j) \leq d(x_i, \mu_m), \nabla m \ \in \{ 1,2, \dots., k\}\}$
4. Update the centroid μ_j for each cluster

$$\frac{1}{|C_j|} \sum_{x_i \in C_j} x_i$$

5. **Convergence:** Repeat the assignment and update steps until the centroids stabilize (i.e., there is little or no change in cluster assignments).

b. Anomaly Detection

Anomaly detection techniques are essential for identifying observations that deviate from expected normal behavior. These techniques use a statistical method of normal behavior and flag any divergence as a potential anomaly, especially in malicious scenarios. We applied the Long Short-Term Memory (LSTM) technique to predict future energy usage and identify anomalies based on these predictions. This approach helps distinguish between typical energy demand patterns and anomalous instances. Public datasets are crucial for evaluating energy demand characteristics, offering essential measurements necessary for research, including (Real power (W), Reactive power (VAR), Apparent power (VA), Phase voltage (V), Current (A), Mains frequency (Hz)). These measurements are also vital for fault detection or power line failures. Different homes exhibit distinct load shapes based on the use of appliances, which can be categorized into four types:Type I: frequent switching devices, Type II: State Machines, Type III: variable power usage devices, and Type IV: Constant power usage devices.

Understanding these types is vital for modeling electric load curves and identifying residential energy demand patterns, which are essential for distinguishing legitimate from illegitimate load curve.

c. LSTM for Anomaly Detection

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) specifically utilized to address the vanishing gradient problem that frequently arises in conventional RNNs. LSTM networks are ideal for detecting electricity theft in smart meter data due to their ability to model temporal dependencies in sequential data [118]. Using daily electricity consumption data from 500 households collected between 2014 and 2016, we train an LSTM model to identify anomalies indicative of theft. Below given figure 3.8 shows the LSTM architecture for anomaly detection in the given dataset. The dataset includes daily electricity consumption values ranging from 3.99 kWh to 28 kWh, along with household characteristics and appliance details. Key features include location (urban/rural), dwelling type, number of rooms, number of families, type of feeder (fully/partially metered), number of occupants, job, income, employment status, heating/cooling wattage, energy efficiency ratings, rated power, and contracted power. We first normalize the electricity consumption data to a scale between 0 and 1 to facilitate better learning. We also include household characteristics and appliance details as additional features [119].

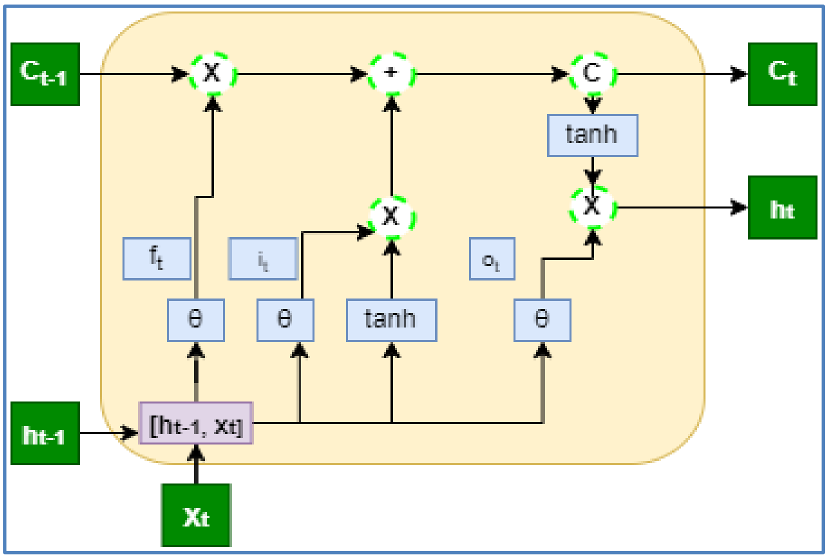

Figure 3.8: Architecture of LSTM model for Anomaly Detection

The model architecture consists of an input layer that takes in sequential daily consumption data along with household and appliance features. This is followed by two LSTM layers with 50 units each to capture complex temporal dependencies. Finally, a fully connected dense layer with a single neuron outputs the predicted next day's consumption. The proposed KPLX framework is developed utilizing 'Mean Squared Error' (MSE) as the loss metric and the Adam optimization algorithm, for 100 iterations with a batch size of 32. After training, the model predicts daily consumption values, and we identified anomalies by comparing these predictions with actual consumption.

In many practical scenarios, normal behavior instances significantly outnumber unexpected cases. The main concept behind using Long Short-Term Memory (LSTM) for anomaly detection is to model normal data samples by adjusting the network's weights to accurately represent the training data. Anomalies are then identified based on deviations or errors in the predictions.

Consider a time series X={x1, x2, …, xn+1}, where each point xi is a multi-dimensional sequence of vector representing a normal sequence. A subset of X, Y= {y1,y2,…,ym} (where m≤n+1) is used as a normal validation set. The LSTM network is trained on X and its corresponding labels Y. After training, the model predicts values and computes prediction errors from the time series.

For instance, given the input xi−1, the model predicts the next value x^i. This process results a set vectors of errors {e1,e2,…,en}, where ei=|x^i+1−xi+1|. Anomalies are detected based on these prediction errors.

The model works by assuming that normal data falls within a specific range and setting a threshold to differentiate between normal and abnormal data. The effectiveness of this approach depends on how much the data overlap. To make a decision, electricity theft detection model assign a probability or theft score to each observation, indicating its likelihood of being theft or abnormal. A binary result, where 0 represents normal and 1 represents theft, is then calculated based on whether the observation exceeds the predefined threshold.

$$r = \begin{cases} 1 & if\ s > \in \\ 0 & if\ s \leq \in \end{cases} \qquad \text{Equation (3.3)}$$

Result r **ϵ** {0,1}, where 0 signifies normal and 1 signifies anomalous or theft data and ∈denotes the threshold.

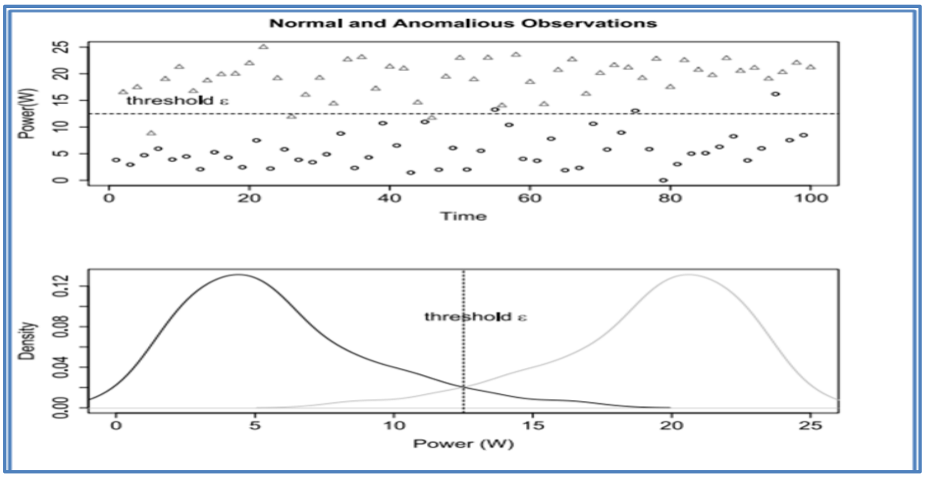

Figure 3.9: Visualisation of the classification problem

We use power (W) as the anomaly score, denoted as ’s’. The example comprises two distinct sets of data: genuine and fraudulent data represented by points and triangles, respectively. In our study, normal consumers were simulated with a training dataset of 2810 normal samples and a test dataset of 148 normal samples. After training the LSTM network, the model's performance was evaluated using a testing dataset of 400 normal and abnormal samples as shown in figure 3.9.

Using the quantiles=0.99, we obtained threshold of ∈ s=0.448. The LSTM-based anomaly detection enhanced the accuracy of our model to 94%, a precision of 96%, an F-score of 86%, and a recall of 98%, outperforming the existing methods in the literature. Quantiles are values that split a dataset into equal parts, each containing the same percentage of the data. Each quantile shows where a specific portion of the data falls within the entire set. Quantiles = 0.99 refers to the number below which 99% of data points fall. This allows us to discover outliers by creating a cut-off point based on the distribution of the data. In the visual representation, the upper plot distinguishes normal data (points) from fraudulent data (triangles), with the dotted line indicating the threshold. An observation assigned a score's' above the threshold, but actually exhibiting legitimate behavior (dot), is classified as a False Positive (FP). Conversely, an event assigned a score's' below the threshold, but actually involving fraudulent behavior (triangle), is classified as a False Negative (FN). Correctly assigned results are referred to as True Positive (TP) for anomalies and True Negative (TN) for genuine behavior.

3.3.5 Model Training and Validation Phase

a. XGBoost Ensemble Method

XGBoost (Extreme Gradient Boosting) is a powerful ensemble learning algorithm, ideal for electricity energy theft detection in smart meters. By integrating predictions from multiple base learners (decision trees), XGBoostcreates a robust and accurate theft detection model. It effectively addresses imbalanced data by assigning higher weights to minority class samples during training, enhancing theft detection capability. Additionally, XGBoost feature importance analysis identifies key factors indicative of suspicious consumption patterns associated with theft. Its ability to capture complex non-linear relationships enables detection of subtle anomalies, crucial for identifying potential instances of theft. With efficient handling of massive datasets, XGBoost is employed for timely and effective energy theft detection across utility networks [120].

XGBoost also incorporates several engineering optimizations that enhance its performance. It employs a sophisticated algorithm for parallel tree construction, which makes it significantly faster than traditional gradient boosting. Additionally, XGBoost supports sparsity-aware learning, which efficiently handles missing values in the data, and it includes mechanisms to prune trees, removing branches that do not contribute to improving the model's performance.The combination of these features makes XGBoost a highly efficient and accurate algorithm to be used for classification. It is capable of handling large datasets and complex models with ease, making it a popular choice in data science competitions and real-world applications. XGBoost's ability to build models that are both powerful and interpretable has contributed to its widespread adoption and success in various predictive modeling tasks [121].The figure 3.10 illustrates the working of KPLX Model for Electricity Theft Detection in Smart Meters, which is created using the XGBoost machine learning ensemble model.

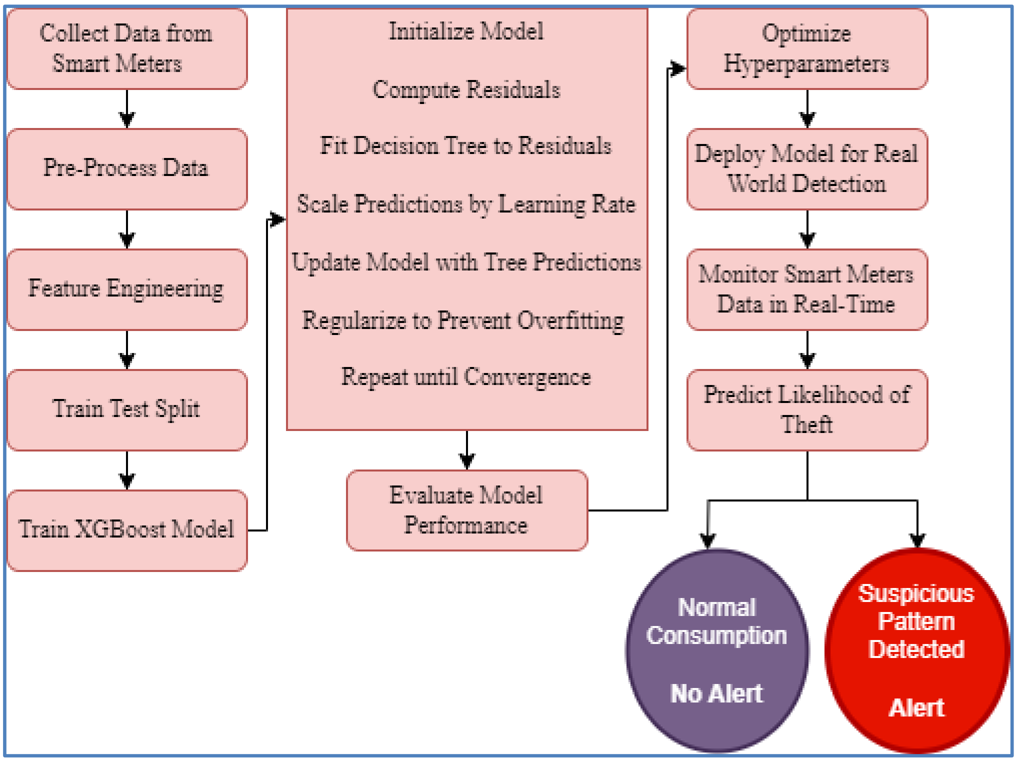
Collect Data from Smart Meters
Pre-Process Data
Feature Engineering
Train Test Split
Train XGBoost Model
Initialize Model
Compute Residuals
Fit Decision Tree to Residuals
Scale Predictions by Learning Rate
Update Model with Tree Predictions
Regularize to Prevent Overfitting
Repeat until Convergence
Evaluate Model Performance
Optimize Hyperparameters
Deploy Model for Real World Detection
Monitor Smart Meters Data in Real-Time
Predict Likelihood of Theft
Normal Consumption
No Alert
Suspicious Pattern Detected
Alert

Figure 3.10: Working of KPLX Model for Electricity Theft Detection in Smart Meters

a. Model Evaluation Techniques

This section focuses on assessing the effectiveness of the model in detecting theft in electricity energy in smart meters to ensure that the developed models are accurate, reliable, and efficient. In this research fundamental and significant evaluation techniques are used which are as [122-124]:

Confusion Matrix:A table used to visualize the performance of an algorithm. It displays the true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). From the confusion matrix, several important metrics can be derived:

Accuracy: The proportion of correctly detected theft cases or true results (both true theft results true positives and true non-theft results or true negatives) in a number of total examined cases.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad \text{Equation (3.4)}$$

Precision: The proportion of positive identifications those were actually correct.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad \text{Equation (3.5)}$$

Recall (Sensitivity): The proportion of actual positives that were identified correctly.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+ FN}} \text{Equation (3.6)}$$

F1 Score: The harmonic mean of precision and recall, providing a single metric to balance the two.

$$\text{F1} = 2 * \frac{\text{Precision}*\text{Recall}}{\text{Precsion}+\text{Recall}} \quad \text{Equation (3.7)}$$

ROC Curve: A graph that shows how well a binary classifier system can diagnose as its discrimination threshold changes. It displays the true positive (TP) rate (recall) versus the false positive (FP) rate.

AUC: The area under the ROC curve; It offers a comprehensive performance assessment across various classification thresholds. A perfect model is indicated by an AUC of 1, whereas an AUC of 0.5 signifies a model lacking discriminative ability.

Precision-Recall (PR) Curve: This curve is particularly useful for imbalanced datasets like energy theft detection, where the number of theft cases is much smaller compared to non-theft cases. It plots precision against recall for different thresholds. The area under the PR curve (AUPRC) is a valuable metric in such scenarios.

Cross-Validation: Cross-validation is a robust method for evaluating ML model’s performance. It involves dividing the dataset into multiple folds and ensuring that each fold gets a chance to be the validation set. Common techniques include:

K-Fold Cross-Validation: The dataset is split into K subsets and the model is trained K times with a different subset as the test set or the validation set and rest K-1 subsets are used for the training of the model.

Stratified K-Fold: It is a variation of K-Fold cross-validation that ensures each fold has the same class distribution as the entire dataset. This technique is especially useful for imbalanced datasets, as it maintains the proportion of each class in every fold. By doing so, it provides more reliable performance estimates and reduces bias, resulting in a better assessment of a model's ability to generalize to unseen data.

Mean Absolute Error (MAE):MEA measures the average magnitude of errors in number of predictions or forecasting made without considering the direction.

Root Mean Squared Error (RMSE): The square root of the average (mean value) of the squared differences (squaring the result of subtracting one number from another) between a prediction and the observed result**.**

3.4 Chapter Summary

This chapter introduces the KPLX Integrated Detection Model, which utilizes advanced machine learning techniques for predictive analysis and anomaly detection in the context of electricity consumption data. The model integrates Long Short-Term Memory (LSTM) for predictive analysis and anomaly detection, as well as eXtreme Gradient Boosting (XGBoost) for the classification of theft and non-theft data points. By leveraging significant features beyond electricity consumption, this approach clusters consumers with similar usage patterns, aiming to address the limitations of conventional ETD techniques, which often suffer from high false positive rates and inefficient use of human resources. The chapter details the approach, which combines clustering, feature extraction, and ensemble classification techniques to improve detection accuracy and robustness.

The methodology is structured into four phases, starting with data collection and pre-processing. The model employs a real-world electricity consumption dataset from China, spanning from January 2014 to October 2016. This dataset consists of 42,372 records, including 3,615 instances flagged as abnormal (indicative of potential theft) and 38,757 instances of normal consumption. Data is collected at 30-minute intervals, encompassing detailed electricity consumption metrics, power parameters, consumer profiles, and weather conditions. Pre-processing steps include handling missing values and ensuring consistency across the dataset. Feature engineering follows, extracting relevant features from the raw data. This includes capturing daily, weekly, and monthly consumption trends, temporal features such as peak usage times and duration of high consumption, and environmental factors like temperature and daylight hours. These features are crucial for improving the input to machine learning models, enhancing their ability to classify theft accurately. The second phase is anomaly detection and classification, which starts by grouping consumers using k-means clustering with similar consumption patterns, followed by anomaly detection in each group using LSTM. The predicted consumption of the consumer is compared with the actual consumption with a set standard threshold of the cluster to detect the anomaly. This stage helps to identify potential anomalies from normal consumption variations.

In the model training and evaluation phase, various ETD models, including the proposed model, are trained on the common features and validated using cross-validation techniques to ensure robustness and generalizability. Performance metrics such as F-score, Recall precision, and accuracy determine the models' effectiveness. By addressing key challenges such as high false positive rates and dynamic consumer behavior based on environmental influences, this methodology offers a significant improvement over traditional detection methods. This comprehensive approach enables utility companies to reduce financial losses and improve operational efficiency by effectively distinguishing between legitimate consumption variations and fraudulent activities.

CHAPTER IV

TECHNIQUES, RESULTS, AND INTERPRETIVE DISCUSSIONS

4.1 Introduction

In this chapter, we delve into the methodologies employed for detecting electricity theft, the results obtained from these methods, and a comprehensive discussion of these findings. The primary focus of our research is the implementation of machine learning techniques to identify electricity energy theft in smart meters by analyzing historical consumption patterns and other relevant features. In this research, we used a raw dataset made publicly available by the SGCC. To ensure accuracy and reliability, the dataset underwent extensive preprocessing methods. The pre-processing included filling in missing values, removing outliers, and standardizing the data. Additionally, we artificially constructed theft instances to balance the dataset, thereby enhancing the robustness of the developed model. We employed Principal Component Analysis (PCA) for the reduction in the dimensionality of the data, incorporating various parameters of electricity usage alongside features derived from statistical techniques and auxiliary databases. Multiple models were trained on given dataset, with XGBoost emerging as the most accurate. The KPLX integrated detection model demonstrated accurate detection with a low rate of false indications, showcasing its effectiveness in discerning patterns in electricity consumption. To evaluate the performance and generalizability of KPLX integrated detection model, we then tested it on the new KPDCL dataset.

4.2 Dataset Description

The dataset of electricity consumers of SGCC covers a vast consumer base, providing a rich source for analyzing electricity usage patterns. It consists of a total of 42,372 records, including 3,615 instances of abnormal consumer data (indicative of potential electricity theft) and 38,757 instances of normal consumer data. The data, collected at 15-minute intervals (though only hourly data is considered), spans from 2014 to 2016.This dataset includes various attributes related to electricity consumption, power parameters, and consumer profiles. Attributes encompass details such as tariff agreements, types of residential houses, lists of registered gadgets, household population counts, and the occupations of family members. In addition to data from smart meters, which includes both genuine and fraudulent consumption patterns, auxiliary datasets like weather data and GIS data are also utilized. The dataset covers a diverse range of residential and industrial customers over different periods (e.g., daily, weekly, monthly). Notably, the dataset is well-labelled, distinguishing between normal and energy theft cases.

The novel KPDCL dataset (used for performance and generalizability evaluation of the developed KPLX integrated detection model) provides detailed records of electricity consumption, containing a total of 2,075,259 observations from December 16, 2006, to November 26, 2010. The data is captured at minute-level intervals, offering high-resolution insights into household power consumption patterns.

We utilized weather data retrieved from the online MET portal.

In the first step, the electricity and weather data were imported and concatenated using Python libraries. Subsequently, three different time features were subtracted from the dates of the time series: hour of the day, day of the week and month of the year.

4.2.1 Data Pre-processing

These datasets consist of various missing values because of problems with the energy meters, cyber-attacks, servicing, data transfer, and storage issues. To address missing values, we employed the linear interpolation method, which is an reliable technique for filling missing values in electricity consumption data because it assumes a steady, continuous change in consumption between known data points, which often reflects real-world electricity usage patterns. This approach maintains the pattern of the data without creating false fluctuations, ensuring uniformity. Furthermore, linear interpolation is computationally efficient, which makes it suitable for handling large EC datasets. It also maintains temporal relationships, ensuring that the interpolated values remain realistic for time-series data. The data collection frequency of smart meters is typically at 15-minute intervals. However, for our study, we adjusted the observation interval to hourly intervals due to the absence of significant changes. This modification resulted in a total of 24 data points per day. The dataset was found to be incomplete in terms of frequency information, as evidenced by the absence of a frequency value (freq = 'None'). The frequency was modified to take place at intervals of 60 minutes. Pandas provides a range of frequency options for calculating frequencies, such as hourly (‘H’), daily (‘D’), weekly (‘W’), monthly (‘M’), annual (‘A’), and additional options.

The electricity consumption data have outlier values recorded immediately after electricity curtailment or fault in the electricity network. These outliers are removed using the “three-sigma rule (TSR) of thumb”. These outliers are removed using the three-sigma rule as mentioned in below equations 4.1 and 4.2

$$\mathrm{f\,(v)} \;= \left\{ \frac{\frac{N}{2}, \qquad v_i \in NaN, v_i(m-1), v_i(m+1) \in NaN}{0, \qquad v_i \in NaN, v_i(m-1) or\ v_i\ (m+1) \in NaN} \right\} \forall \mathrm{v_i \text{ and } v_i} \notin \mathrm{NaN} \quad \text{Equation (4.1)}$$

$$\mathrm{O(v_i, t) = \;\; wi\,f\,v_i\overline{(t)} > v_i(\bar{t})otherwise}$$

$$\mathrm{where\ w = avg(v_i)) + 2}s\mathrm{(v_i(t))} \quad \text{Equation (4.2)}$$

Normalization of Data : After filling in the missing values and removing outlier values, Min-max normalization is utilized for preserving the relationships between values by scaling them to a fixed range, usually [0, 1], without distorting differences. Min-max normalization method ensures all features, regardless of scale, are within a uniform range, which improves the performance of machine learning models like neural networks and XGBoost. Additionally, since electricity consumption is non-negative, this method naturally fits by keeping values positive and within a consistent range. This approach is equally easy and effective in handling varying scales of consumption data. The data values have been normalized using min-max normalization according to equation 4.3.

$$\mathrm{N\big(v_i(t)\big) = \frac{v_i(t) - min(\quad)}{imax(\quad) - min(\bar{v})}} \qquad \text{Equation (4.3)}$$

vi (t) is the usage of electricity at time say t , min (v), the usage of minimum electricity, and max(v) is the usage of electricity at the time (t).

4.2.2 Imputing False Theft Data

An electrical thief plans to manipulate the energy meter to reflect a lower consumption than the actual quantity utilized or to strategically steal energy during high usage periods. We design the six types of theft cases based on benign scenarios. We assume that no fraudulent users have altered any of the historical data. The consumers' daily metering data are denoted by the notation $X = (x_1 + x_2 + x_3 + \cdots x_{24}$)reading after every 60 minutes in 24 hours. Smart meters communicate metering data (in kilowatts) to the data management system every 15 minutes, but we resampled the data to record data at an interval of 60 minutes.

We assume $\boldsymbol{x} = (x_{1,} x_2, x_{3,\ldots\ldots,} x_n)$ a vector of genuine consumption values for 24 hours with n samples, and x $\boldsymbol{\epsilon}$ X, in which X is a random vector and having P0 distribution. The utility will compute the electricity consumption on the values say $\boldsymbol{y} = (y_{1,} y_2, y_{3,\ldots\ldots,} y_n)$ from the meter readings. In the case of honest customers, we have y = x, but for fraudsters, y = h(x), where y $\boldsymbol{\epsilon}$ Y again here Y is a random vector and has a P1 distribution, so that$E[Y] \leq \ E[X]$. It is possible to figure out h (x) by studying various energy theft scenarios and their effect on values that have been measured. For example, if h(x) = $\boldsymbol{\alpha}$x, then 0 <= $\boldsymbol{\alpha}$<= 1 is a possibility. So, using the benign dataset to make malicious samples is a smart option. Even though it might not be possible to define all functions that lead to$E[Y] \leq E[X]$, a complete set of attack samples can be made by looking at different situations and using the generalization property of the classifier.

The various scenarios of theft attacks to alter metering data produce malicious samples and are defined below. Data balancing using imitated real attacks is implemented to balance the theft and non-theft samples in the SGCC dataset.

In all the six types of attacks which are described in chapter 3 we opt for attacks more similar to the modernized theft attacks mentioned and formulated more pragmatic and real abnormal consumption patterns. The genuine consumption of a user is denoted by $\mathrm{E_t}$, where, (t $\boldsymbol{\epsilon}$ [0, 1034]). In this study, the SGCC dataset contains the total of 1035 days' consumption data. The six types of attacks($\mathrm{t_1}$ to $\mathrm{t_6}$) used here are presented in mathematical formats. These attacks are used to balance the dataset and test the model :

t1(x_t) = $x_t *$ random (0.1, 0.9),

t2(x_t) = $x_t * r_t$, r_t = random (0.1, 1), t3(x_t) = $x_t *$ random[0, 1],

t4(x_t) = mean(X) $*$ random(0.1, 1),

t5(x_t) = mean(X),

t6(x_t) = $x_{t1034} - t$,

Where, X is a set comprising elements $\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3, \ldots\ldots \mathrm{x}_{1034}$.

- In theft attack 1, the function t1(x_t) scales each entry in a row (representing actual readings) by a randomly generated value between 0.1 and 0.9. This behavior indicates that consumers may have altered the current transformer (CT) by manipulating the smart meter or that the smart meter itself has been tampered with using hardware.
- In theft attack 2, the function t2(x_t) applies a unique randomly generated multiplier to each timestamp within a row. These multipliers are drawn from a range between 0.1 and 1, inclusive. For instance, if the consumption values at three timestamps are 50, 75, and 100, and the random multipliers are 0.3, 0.7, and 1.0 respectively, the resulting values will be: 50 *0.3 = 15, 75*0.7 = 52.5, and 100* 1.0 = 100. We have taken the type 2 theft attack when a consumer is involved in theft at different periods of the day either by hooking or by inserting some external hardware in between energy meter and load.
- In theft attack 3, the behavior of the consumer alternates between reporting the actual electricity consumption (EC) values and zero values. This means that at some timestamps, the reported consumption is accurate, reflecting the true energy usage, while at other timestamps, the reported consumption is falsely recorded as zero.For example, consider the actual consumption values over a sequence of five timestamps: [30, 45, 60, 75, 90]. Under theft attack 3, the reported values might look like this: [30, 0, 60, 0, 90]. Here, the first, third, and fifth timestamps reflect the true EC values, while the second and fourth timestamps report zero consumption, creating a pattern of intermittent fraudulent data. In this scenario, the theft is identified when the consumer uses bypassed lines specifically during the operation of high power-consuming devices.
- The 4th theft attack mimics a consumer bypassing the energy meter but

maintaining a small constant load during this period. Only a single computed value is recorded as the actual reading. In this attack, a random multiplier is applied to the mean value of consumption recorded during other times to generate synthetic theft readings.

- For the 5th attack, using illegal hardware to record less consumption aligns with this theft attack. If the hardware consistently records less consumption, it resembles the fourth theft attack with a random multiplier applied to the mean value. However, if the hardware reports the mean consumption value directly, it is the same as the fifth theft attack.
- The 6th theft attack simulates the behavior of a consumer engaged in theft by sending the actual readings in reverse order. This attack is particularly beneficial in scenarios where Time of Use (TOU) rates differ. We applied these six synthetic theft attacks to genuine consumers' consumption data to strike a balance between the theft and non-theft data samples.

All six attacks are tested separately as well as in combination to test the performance of the proposed model taking into consideration the weather and erratic power supply conditions.

4.3 PCA for Dimensionality Reduction

In the context of electricity consumption datasets, the data is organized into a matrix format, where each row represents a sample (e.g., a day or an hour) and each column represents a feature (e.g., different electricity consumption attributes e.g. MeterID, Timestamp, electricity consumption EC in KWh, peak load, average load, minimum load, load profile, power factor(PF), Time-of-use, average load factor, historical consumption, tariff details, socio-demographic data, load agreement details, GIS data, weather data, etc.). Standardization for EC features is performed by subtracting the mean and dividing by the standard deviation of each feature which is crucial to ensure that features are on a similar scale, as PCA is sensitive to the variances of the features, the covariance matrix is computed to analyze relationships between features [125-126]. The eigenvectors and eigenvalues are computed from the calculated covariance by performing an Eigen-decomposition on the matrix. The primary components are the new orthogonal axes in the feature space, and they are represented by the eigenvectors. The eigenvalues reveal how much variation is explained by each principal component. For instance, if you started with 50 features per time step and reduced it to 10 using PCA, then k=10. Let's suppose we take 100 consumers (n=100). We analyse daily electricity consumption over 30 days (t=30) and we reduced the features to 10 principal components (k=10). The tensor $\mathrm{X}_{k,Seq} \in R^{100\,X\,30\,X\,10}$ would represent 100 rows (one for each consumer). 30 columns representing the 30 days of consumption data. Each entry in the tensor for a particular consumer and day would be a vector of 10 features (the principal components).

PCA enables dimensionality reduction while retaining important information for tasks such as visualization, anomaly detection, or theft detection. As described in Table 4.1, the performance metrics of the proposed model is compared to other models trained using same technique. While using dimensionality reduction the standardization of data matrix is done by subtracting the mean and scaling to unit variance. The normalized value is calculated as below:

$$X_{Std} = \frac{X - \mu_x}{\sigma_x} - - - - - - - - Equation\ (4.4)$$

where μ_x are mean and σ_x the standard deviation of the data matrix X, respectively.

Mathematically the application of PCA for dimensionality reduction is discussed below:

We compute the covariance matrix Σ of the standardized data below:

$$\Sigma = \frac{1}{n-1} {X_{std}}^T X_{std} - - - - - - - - Equation\ (4.5)$$

We perform eigenvalue decomposition on the covariance matrix here:

$$\Sigma \mathrm{v}_i = \lambda_i v_i \mathrm{i} - - - - - - - - Equation\ (4.6)$$

We then select principal components by selecting the top k eigenvectors corresponding to the largest eigenvalues to form the principal components matrix:

$$\mathrm{v}_k \in R^{n\,X\,K} \lambda_i v_i \mathrm{i} - - - - - - - - Equation\ (4.7)$$
$$\mathrm{X}_k = \mathrm{X}_{std} V_k - - - - - - - - Equation\ (4.8)$$

After that the projection onto Principal Components is performed and we project the original data X onto the principal components to obtain the reduced-dimensionality data

$$\mathrm{X}_k \in R^{n\,X\,k} --------- Equation\ (4.9)\mathrm{X}_k$$
$$= \mathrm{X}_{std}V_k --------- Equation\ (4.10)$$

X_k now contains the most informative features extracted from the original dataset. The notation $\mathrm{X}_{k,Seq} \in R^{n\,X\,t\,X\,k}$ refers to a tensor that represents the input data arranged in sequences for each consumer, where: n is the number of consumers. t is the number of time steps or the length of the sequence (e.g., daily, weekly, monthly consumption data). k is the number of features per time step after applying PCA (Principal Component Analysis). The evaluated metrics include F-score, recall, precision, AUC-ROC and accuracy after applying PCA is given in Table 4.1 below:

Table 4.1: XGBoost Theft Detector Metrics (With and Without PCA)

Metrics	Without PCA	With PCA
Accuracy	90%	95%
Precision	85%	90%
Recall	95%	98%
F1 score	90%	94%
AUC-ROC	0.95	0.98

- Accuracy: The XGBoost model without PCA achieved 90% accuracy, while the model with PCA reached 95%, showing a notable improvement.
- Precision: The model without PCA had a precision of 85%, compared to 90% with PCA, indicating better precision with dimensionality reduction.
- Recall: The recall improved from 95% without PCA to 98% with PCA, suggesting the model with PCA detected more actual theft cases.
- F1 Score: The F1 score increased from 90% to 94% with PCA, highlighting an overall performance boost.
- AUC-ROC: The AUC-ROC improved from 0.95 to 0.98 with PCA, demonstrating a higher discriminatory power.

In summary, as described in Table 4.1, the proposed model trained on the PCA-reduced dataset outperformed the model trained on the original dataset across all evaluated metrics, suggesting that PCA effectively enhances the model's performance.

4.4 Exploratory Data Analysis in Feature Engineering

Features are extracted to add additional parameters from the existing to capture relevant patterns or relationships. Relevant features can enhance the efficiency and reliability of the electricity theft-detecting model [127]. The various features collected directly and extracted using statistical functions include Consumer specific unique ID, electricity usage time and date (Timestamp), electricity consumption (kWh), active power, reactive power, average voltage, global intensity, power factor, max_load, min_load, average_load, load dispersion, peak demand, total load profile, seasonal variation, time-of-use, historical consumption, socio-demographic data, billing information, geographic information, time of day usage, weekday/weekend, and state holiday. In addition to the electricity consumption and electric power parameters, the utility has other features available, like load agreement details of the customer, tariff information, meter location, and customer’s CIBIL score (which is similar to credit scoring methods to evaluate a consumer's payment history, reliability, consistency and stability in electricity consumption), curtailment schedule (if any due to demand-supply gap).Apart from that, data collection is done using GIS location details. Further, the weather database available (max. temp., min. temp., precipitation, etc.), and the technical details (meter type, meter location, etc.) are also included in the processed dataset. Hourly data is calculated as summing every four intervals (15-minute data) to get hourly values.

- **Daily Data**: Sum all ninety six intervals for a full day's consumption.
- **Weekly Data**: Sum the daily data for seven days.
- **Monthly Data**: Sum daily data for all days in the month.
- **Seasonal Data**: Sum the monthly data for the three months corresponding to a season.

This derived data will help in identifying consumption trends and anomalies over various time periods.

The extracted features also include the categorical variable to store values for the time of day (M: morning, A: afternoon, E: evening, N: night). Binary variable for weekday/weekend (0 for weekday, 1 for weekend), and binary variable for state holidays (Holiday as 1 and rest as 0).Each consumer's electricity consumption is analyzed over a period of time. The focus is on understanding the consumption patterns, identifying anomalies or deviations specific to each consumer, and detecting any unusual behavior like electricity theft within the consumption data. This approach allows analysis for finding the unique characteristics and consumption patterns of each consumer [3]–[5][3]–[5][3]–[5][3]–[5][3]–[5][2]–[4]. Data aggregation is performed on data consumption over different periods (daily, weekly, monthly). It involves combining the individual consumption readings within each period and calculating statistical measures or creating lag variables to capture temporal patterns. The following features are extracted using statistical functions:

$$\text{data\_columns } [{}'\text{said}, {}'\text{energyConsumption}/\text{hh}', {}'\text{Total KWhr}'] \quad \text{Equation (4.4)}$$

a) Hourly ans Daily Aggregation

To aggregate hourly and daily data, all the consumption readings are collected to produce statistical metrics i.e. mean, variance, minimum, maximum, or sum of electricity consumption during a day for a single consumer. These parameters reveal daily average, spikes of low and high usage (as shown in figure 4.1).

Let $x_{i,j}$ represent the energy consumption at the j-th 15-minute interval of the i-th day. Here i is the day index and j=1,2,…,96 represents the 15-minute intervals as originally recorded by smart meter within each day. To derive hourly data from 15-minute intervals, we take sum of every four consecutive 15-minute readings. For a given hour h on day i: $X_{i,h}^{hourly} = \sum_{j=4(h-1)+1}^{4h} x_{i,j},$ h = 1,2,3…24 for each day and summing four intervals gives total consumption for each hour. Likewise daily consumption is calculated as : $X_{i}^{daily} = \sum_{j=1}^{96} x_{i,j}.$

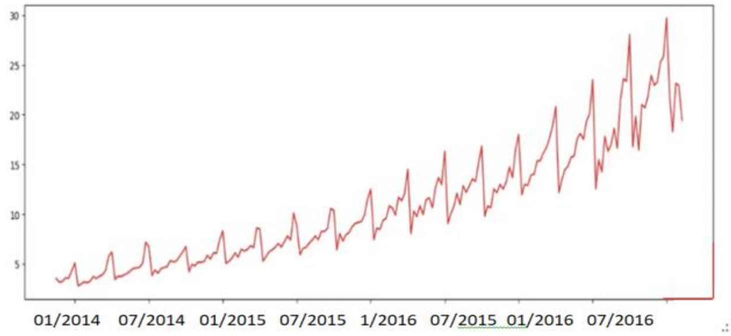

Figure 4.1**:** Trend in electricity usage of a consumer

b) Weekly Aggregation

For weekly aggregation, weekly data is combined. The statistical functions reveal the weekly average, weekly consumption patterns, and high and low usage days in a week (as shown in Figure 4.2).

To derive weekly data, sum the daily totals over 7 consecutive days. For week w, the total energy consumption is calculated as: $X_w^{weekly} = \sum_{i=7(w-1)=1}^{7} x_i^{daily}$.

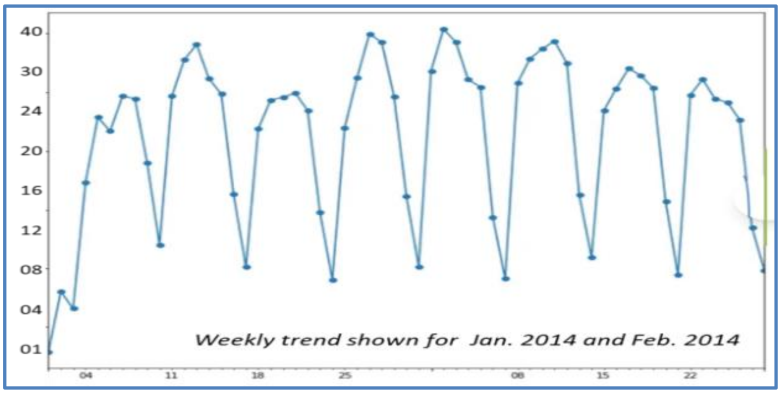

Figure 4.2: Weekly trend in electricity usage of a consumer (ID: 15167)

c) Monthly Aggregation

The monthly aggregation groups the consumption data by month. Monthly aggregation helps in uncovering the long-term consumption trends such as seasonal fluctuations in this study. For monthly data, we take the sum of the daily consumption totals across all days in a month. For month m, the total energy consumption is:

$X_m^{monthly} = \sum_{i=dm1}^{dm2} x_i^{daily}$. Where dm1 and dm2 are the first and last days of month m. Likewise for seasonality, seasonal data sum of the daily data across a full season (e.g., summer, Autumn/fall, winter, spring). Let's say for season $\boldsymbol{s}$, the total energy consumption is: $X_s^{Seasonal} = \sum_{i \in season\ s} x_i^{daily}$. For example we consider summer months as June, July, and August. Seasons generally follow a 3-month period (e.g., winter, spring, summer, Autumn)

d) Lag Variables

By following the approach, we derived aggregated consumption data from 15-minute intervals to various timeframes such as hourly, daily, weekly, monthly, and seasonal levels. This gave us the flexibility to analyze consumption patterns at different scales, which proves to be crucial for anomaly detection and identifying seasonal patterns in the electricity usage.

Past consumption builds lag variables and captures temporal patterns. Lag variables represent the difference in consumption of the preceding day, preceding week, or preceding month on comparison. The lag variables assist us in identifying data increases in usage over time, dependencies, and consumption patterns. This study utilizes the Resampling which is a statistical technique that involves the consolidation of data within a specified timeframe. The data is initially divided into time bins, and subsequent computations are carried out on each bin. Resampling is done on an hourly, daily, monthly, and yearly basis to provide relevant statistics such as minimum, maximum, and mean values in consumption[6], [7][6], [7][6], [7][6], [7][6], [7][6], [7][5], [6]. We compute hourly mean values for each consumer’s electricity consumption:

Electricity consumption data_columns = [‘smID’, 'energyConsumption/hh', 'Total KWhr']

data_hourly_mean = data[data_columns].resample ('H').mean ()

H stands for hourly data_hourly_mean. Likewise, weekly and monthly mean is calculated by using weekly ('W') and Monthly ('M') mean.

e) A Rolling Window Technique for Weekly Trends

The distinction between the rolling and hourly/weekly/ monthly lies in the overlapping nature of the bins[3]–[7][3]–[7][3]–[7][3]–[7][3]–[7][2]–[6].The bins for weekly rolling resampling are organized as follows: Jan.1st to Jan.7th, Jan.8th to Jan.14th, and Jan.15th to Jan. 21st, and so on. The bins are organized on a weekly rolling basis, with each bin representing seven days. For example, the first bin spans from January 1st to January 7th, the second bin spans from January 2nd to January 8th, the third bin spans from January 3rd to January 9th, and so on. In order to calculate a 7-day rolling mean, we follow the mathematical procedure below

$$data_{7d_{rol}} = data[data_{columns}].rolling(window = 7, center\ True).mean(X)data_{7d_{rol}}$$

In the above-mentioned code, the parameter center = True indicates that when calculating the rolling mean for a given time bin, such as from Jan 1 to Jan 8, the resulting value will be positioned adjacent to the middle of the bin, specifically on Jan 4. Figure 4.3 shows rolling trends on a daily, weekly, and yearly basis.

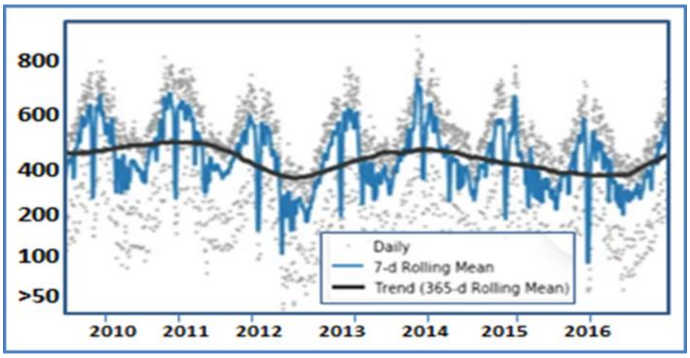

Figure 4.3: Daily, weekly, and 365-day rolling trends

f) Visualizing Trends in Data using Rolling Means

Trend is the smooth long-term tendency of a time series. It might change direction (increase or decrease) as time progresses.

g) Seasonal Trends

One effective method for visualizing the trends is through the utilization of rolling means at various time scales as shown in figure 4.4.Upon analyzing Figure 4.4for the 365-day rolling mean time series, it becomes evident that the general annual trend in electricity consumption of a genuine consumer exhibits a considerable level of consistency. For training a machine learning model trends need to be removed as trends can obscure the true underlying patterns in the data and can lead to spurious correlations and incorrect conclusions in statistical analyses.

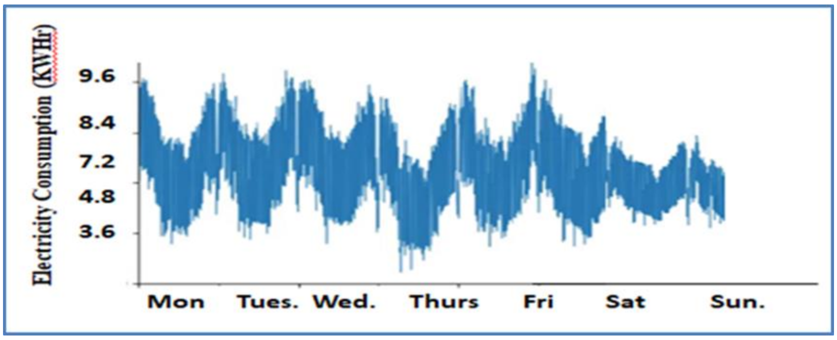

Figure 4.4: Consumer's consumption pattern peaks in the evenings and is low during the day Also, Lower spikes during weekends

By removing the trend, the data is transformed into a stationary series, making it more amenable to the problem of theft detection on historical data. Removal of trends improves efficiency and proves advantageous, especially when the trend is prominently observable as depicted in figure 4.2. In this study, the trend is eliminated using a differencing technique (figure 4.5). Differencing involves the creation of a data value in which the value at a given time (t) is calculated by subtracting the actual recorded reading at that time (t) from the actual recorded reading at the preceding time (t-1).

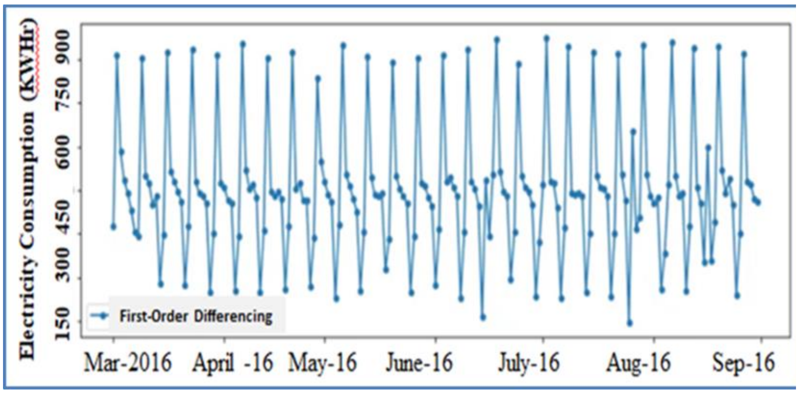

Figure 4.5: First order differencing of consumption

Differencing transforms non-stationary data into stationary data. This facilitates the precise assessment of the seasonal fluctuations or random fluctuations observed in the electricity consumption time series data. Now, the values at this differenced column are a subtraction of two consecutive values recorded by the smart meter. In general, the information conveyed by differenced readings are not about the specific value at a given point in time, but rather the magnitude of its deviation from the previous point in time. The graph is a plot of differenced values, the preponderance of the values will be distributed along both sides of the x-axis (where y=0). This is due to the likelihood that most consumption values will either be higher or lower than the previous day and fewer instances of values where the difference is zero between two consecutive days. The experiments show that the dataset has performed well on first-order differencing as demonstrated in figure 4.3 and figure 4.4. A research study was undertaken to ascertain the existence of any weekly patterns within the dataset. Based on the depicted plot in figure 4.4 the observed practice displays a prominent peak during the evening and night hours while diminishing during the daytime. Moreover, it is evident that the consumption on weekdays surpasses that on weekends, as indicated by the lower spikes observed on Saturdays and Sundays. In this study, we analyze a dataset that encompasses two months, specifically January 2014 and Feb. 2014 as seen in figure 4.3. The presence of distinct weekly variations is readily apparent in the observed data. The analysis of power consumption patterns reveals a notable disparity between weekends and weekdays, with the former exhibiting a lower level of energy usage and the latter characterized by significantly higher consumption rates. Monthly aggregation helps in uncovering the long-term consumption trends such as seasonal fluctuations in this study as seen in Figure 4.6 and Figure 4.7.

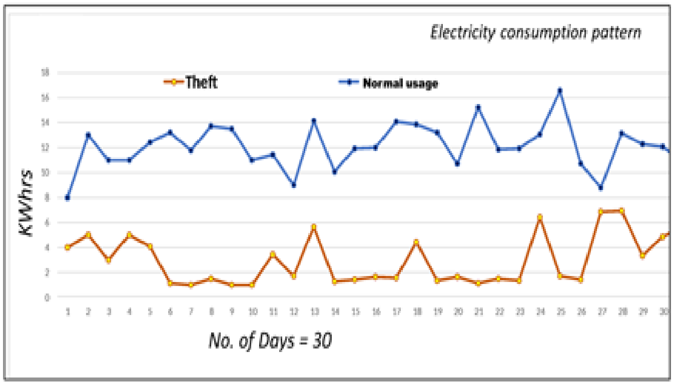

Figure 4.6: Comparison in usage trend of 60 consumers' normal vs. fraud consumer

h) Autocorrelation

Autocorrelation can find seasonal trends in time series data. The autocorrelation function (ACF) is highly useful in analyzing historical electricity to reveal seasonal patterns. High autocorrelation values at various lags suggest a strong link between past and future values on the daily, weekly, seasonal, or monthly consumption patterns. Aberrant usage, spot anomalies, or deviations indicate electricity. Autocorrelation analysis can also help in forecasting which is beyond this study. Based on the above figure 4.6 and figure 4.7; it is evident that there exists a notable peak in correlation at the seventh-day lag. Subsequently, a similar peak is observed on the fourteenth day, followed by subsequent occurrences. The observed phenomenon exhibits a repeating pattern over 7 days, indicating a weekly time series. The observed pattern exhibits a gradual decline in effectiveness over approximately three months or approximately 300 days. As the time increases, the degree of correlation between them diminishes. Figure 4.7 reveals that the consumption series is genuinely auto-correlated with a lag of 1 week for a specific normal consumer.

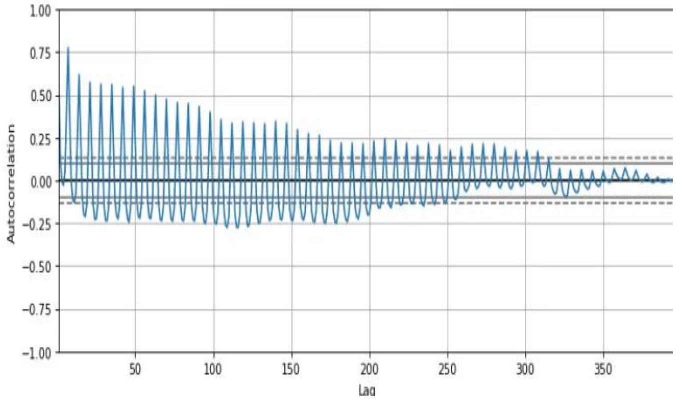

Figure 4.7: Shows the trend vanishing after about 300 days

4.5 Anomaly Detection using LSTM

The LSTM is used **t**o capture temporal dependencies. LSTM, a type of recurrent neural network (RNN) specifically is utilized for detecting anomaly in electricity consumption due to their ability to model temporal dependencies in sequential data. The LSTM model is trained to learn temporal patterns in the data. If a time series of electricity consumption data X={x1,x2,…,xT} is taken where T is the total number of time steps (e.g., 15-minute intervals). The LSTM learns a mapping from a sequence of previous observations to a future value like:

$\hat{x}_t = LSTM(x_{t-1}, +x_{t-2}, + x_{t-3}, \dots., x_{t-k})$ Where $\hat{x}_t$is the predicted consumption at time t1. $(x_{t-1}, +x_{t-2}, + x_{t-3}, \dots., x_{t-k})$ are the past k observations used as input to predict x_t. And LSTM is the trained model. The prediction error, also called the residual, is computed as the difference between the actual and predicted values:

$e_t = x_t - \hat{x}_t$ where e_t is the error at time t, x_t is the actual consumption value $\hat{x}_t$ is the predicted electricity consumption. A simple approach is used to define a fixed threshold ϵ. If the absolute prediction error ∣et∣ exceeds this threshold, the data point at time t is flagged as an anomaly i.e if ∣et∣ > ϵ indicates if the absolute error ∣et∣ exceeds the threshold ϵ epsilon, the data point at time t is considered an anomaly. Say, we set a threshold ϵ=10, and say if get the values like below:

Table 4.2: Threshold to distinguish normal variations from anomalous behavior

Time	**Actual Consumption (kWh)**	**Predicted Consumption (kWh)**	**Error (Residual)**
20:00	10	8	2
20:15	15	12	3
20:30	17	4	-13
20:45	20	5	-15
21:00	22	5	-17

In Table 4.2, the time point at 20:30 where the error is -13 will be flagged as an anomaly because ∣−13∣>10. Likewise at 20:45 and 21:00 the error is more than threshold so after 20:30 the consumer is seen to be consume very less energy and hence is flagged as anomaly.

The working of LSTM cell at each time step t is governed by:

$$f_t = \sigma(W_f.[h_{t-1}, x_{k,t}] + b_f) --------\text{Equation (4.11)}$$
$$i_t = \sigma(W_i.[h_{t-1}, x_{k,t}] + b_i) --------\text{Equation (4.12)}$$
$$\tilde{C}_t = tanh(W_C.[h_{t-1}, x_{k,t}] + b_C) --------\text{Equation (4.13)}$$
$$f_t = \sigma(W_f.[h_{t-1}, x_{k,t}] + b_f) --------\text{Equation (4.14)}$$

$$C_t = f_t \,\Theta\, C_{t-1} + i_t \,\Theta\, \tilde{C}_t \text{ ---------Equation (4.15)}$$

$$i_t = \sigma\left(W_o.\left[\, h_{t-1}, x_{k,t}\right] + b_o\right) \text{ ---------Equation (4.16)}$$

$$h_t = o_t \,\Theta\, \tanh(C_t) \text{ ---------Equation (4.17)}$$

where h_t represents the hidden state (temporal feature) at time step t. Use the hidden state h_T at the final time step T as a feature vector for anomaly detection. Anomalies are identified by setting thresholds on the LSTM output. By setting appropriate thresholds on the prediction errors from the LSTM model, anomalies or unexpected deviations in energy consumption are identified. Choosing the right threshold is critical for detecting true anomalies while minimizing false positives.

4.6 Classification using XGBoost

The hidden states $\boldsymbol{h_T}$ from the LSTM are used as features for the XGBoost classifier as shown below :

H $\in R^{n\,X\,h}$

Objective Function of XGBoost is used to minimize the following objective:

$$D = \{(x_i, y_i)\}(|D| = n, x \in R^m, y_i \in R)$$

$$£(\Theta) = \sum_{i=1}^{n} l(\,\hat{y_i} - y_i) + \sum_{j=1}^{T} \Omega\,(f_j)$$

where:

- $l(\,\hat{y_i} - y_i)$ is the loss function (e.g., log loss for classification).
- $\Omega(f_j)$ is the regularization term for the j-th tree.
- T is the number of trees in the ensemble.
- $\hat{y_i}$ is the predicted label.

Prediction: The prediction for a new sample is given by:

$$\hat{y_i} = \sum_{j=1}^{T} f_j\,(h_{T,i})$$

Where $\left(h_{T,i}\right)$ is the LSTM-derived feature vector for the i-th consumer.

4.7 Comparative Analysis of Machine Learning Models for Electricity Energy Theft Detection

Electricity theft is a major challenge for utility companies, causing significant financial losses and operational inefficiencies. Traditional methods of detecting theft, such as manual inspections and rule-based systems, are often ineffective and inadequate for modern power grids. In this study, we evaluated the performance, generalizability, and robustness of various machine learning models, including the Autoregressive model, LGBoost, CatBoost, Random Forest, Linear Regression, k-Nearest Neighbors (KNN), Support Vector Machine (SVM), AdaBoost, and the proposed ensemble LSTM-XGBoost, using the SGCC dataset, which contains both benign and malicious samples. These models were assessed using metrics such as accuracy, precision, recall, and F1-score.

We tested each machine learning model on the SGCC and KPDCL dataset to evaluate their performance and generalizability. The results were excellent, demonstrating the robustness of our approach. The LSTM-XGBoost model not only performed well on the SGCC dataset but also exhibited strong performance on the KPDCL dataset. This indicates that the model is well-suited for deployment on any new dataset, providing reliable and optimal results. The real-world application of the model will bring significant benefits to utility companies, enhancing their ability to detect and prevent electricity theft effectively.

Our findings offer valuable insights into the strengths and weaknesses of each model, assisting utility companies in choosing the most effective method for electricity theft detection.

4.7.1 Autoregressive Model

The autoregressive (AR) model is a statistical model commonly used in time series analysis. In this study, AR is used to predict electricity consumption based on previous values (previous half-hour consumption). It is used to model univariate time series for different types of days (weekdays, Saturdays, and Sundays) and for each hour for each consumer.[6][128].

$$s_t^{h,d} = \mathrm{c} + \sum_{i=1}^{q} \emptyset_i^{h,d} \, s_{t-i}^{h,d}, \quad \text{Equation (4.18)}$$

Equation (4.18) represent the model, where a constant denoted as c and model parameters ϕ h,t are considered. To maintain simplicity and avoid a trial-and-error parameterization process, parameters are set to ϕ. The AR model's parameters are set to ϕ. In light of the aforementioned considerations, it has been deemed necessary to make certain adjustments to the prediction methodology employed for each hour. This adjustment involves the utilization of the AR model, with the specific parameters being contingent upon the type of day under consideration. Furthermore, the value of q, representing the order of the AR model, is set to 3 arbitrarily without any justification here. The computation considers the three most recent values of the same-day type.. The results generated by the AR model on the SGCC dataset are presented in Table 4.3. Additionally, comparisons of Precision vs. Recall values and True Positives vs. False Positives are illustrated in Figures 4.8 and 4.9, respectively.

4.7.2 LightGBM (Light Gradient Boosting Model)

LightGBM (Light Gradient Boosting Machine) is an advanced gradient boosting frameworkdesigned for high efficiency, speed, and scalability. It is especially effective for managing large datasets with high dimensionality, making it a popular choice for a variety of machine-learning tasks, including regression, ranking, and classification. LightGBM achieves its high performance through several key features, including its efficient training speed, lower memory usage, and effectively handle large-scale data. One of the distinctive aspects of LightGBM is its use of leaf-wise tree growth rather than the traditional level-wise approach. LightGBM is used to predict non-linear patterns by leveraging features such as temperature-related variables, the value of electricity usage in the preceding hour, and the same hour's value from the previous day [129, 130]. The workflow involves preparing the data, initializing the LightGBM dataset, configuring parameters, training the model, evaluating its performance, and tuning hyperparameters to optimize results. After model optimization, it can be deployed to make accurate predictions on new data. The outcomes produced by the LightGBM model on the SGCC dataset are outlined in Table 4.3, while figures 4.8 and 4.9 depict the comparisons between Precision and Recall values, and True Positives and False Positives, respectively.

4.7.3 CatBoost Energy Theft Detection Model

CatBoost, a gradient-boosting framework, efficiently handles categorical features without the need for preprocessing. It internally converts categorical values into numerical ones, preserving information effectively. Data preparation involves collecting data from smart meters, customer information systems, and historical usage records. Key features are engineered to capture usage patterns, and the data is labeled to identify energy theft cases for supervised learning [131]. During model training, CatBoost utilizes its native handling of categorical data and ordered boosting technique to prevent overfitting. The model is deployed in real-time monitoring systems and continuously updated with new data for adaptive learning. Table 4.3 displays the results obtained from applying the CatBoost model to the SGCC dataset, while figures 4.8 and 4.9provide insights into the comparisons of Precision versus Recall values and True Positives versus False Positives.

4.7.4 Random Forest Model

Random Forest is a powerful and flexible ensemble learning technique widely used for classification and regression tasks. It works by building multiple decision trees during the training phase and provides the mode of the classes for classification or the average prediction for regression. Random Forest proves to be particularly effective due to its ability to handle complex, high-dimensional data with both categorical and numerical features, which are typical in smart meter datasets. The algorithm begins by randomly selecting subsets of the training data and features and then constructs decision trees using these subsets. This randomness helps to reduce overfitting and de-correlates the individual trees, leading to a more robust and accurate ensemble model. Each tree makes a prediction individually, and the final output is obtained by combining the predictions of all the trees. This is typically done through a majority vote for classification tasks or by averaging the predictions for regression tasks [132, 133]. Random Forest is trained on historical smart meter data containing various features such as usage patterns, geographic location, time-of-use information, and customer demographics. By analyzing these features, the model learned to identify patterns indicative of energy theft, such as abnormal usage spikes or deviations from typical consumption behavior. Once trained, the Random Forest model can be deployed in real-time to monitor smart meter data streams, flagging potential instances of electricity theft for further investigation by utility companies. The random forest model's performance on the SGCC dataset is summarized in Table 4.3, accompanied by graphical comparisons of Precision versus Recall values in Figure 4.8 and True Positives versus False Positives in Figure 4.9.

4.7.5 Logistic Regression Model

Logistic regression is a statistical method utilized for binary classification tasks, making it highly effective for detecting instances of electricity theft in smart meter data. Unlike linear regression, which predicts continuous values, logistic regression provides binary outcome based on one or more predictors. It achieves this by applying a logistic function (or sigmoid function) to the linear combination of the predictor variables. This function maps any real-valued input to a value within the range of 0 to 1, representing the probability of the positive class (e.g., electricity theft) given the input features. The model is trained using a technique called maximum likelihood estimation, where the parameters are optimized to maximize the likelihood of observing the actual class labels in the training data [134, 135].During the training process, the logistic regression model learns the relationship between the predictor variables and the probability of electricity theft. Once trained, the model can make predictions by estimating the probability of theft for new instances of smart meter data. A threshold value is typically applied to these probabilities to make binary predictions: if the probability exceeds the threshold, the instance is classified as indicating theft; otherwise, it is classified as non-theft. In Table 4.3, the findings from employing the logistic regression model on the SGCC dataset are documented, while figures 4.8 and 4.9visualize the comparisons between Precision and Recall values, as well as True Positives and False Positives, respectively.

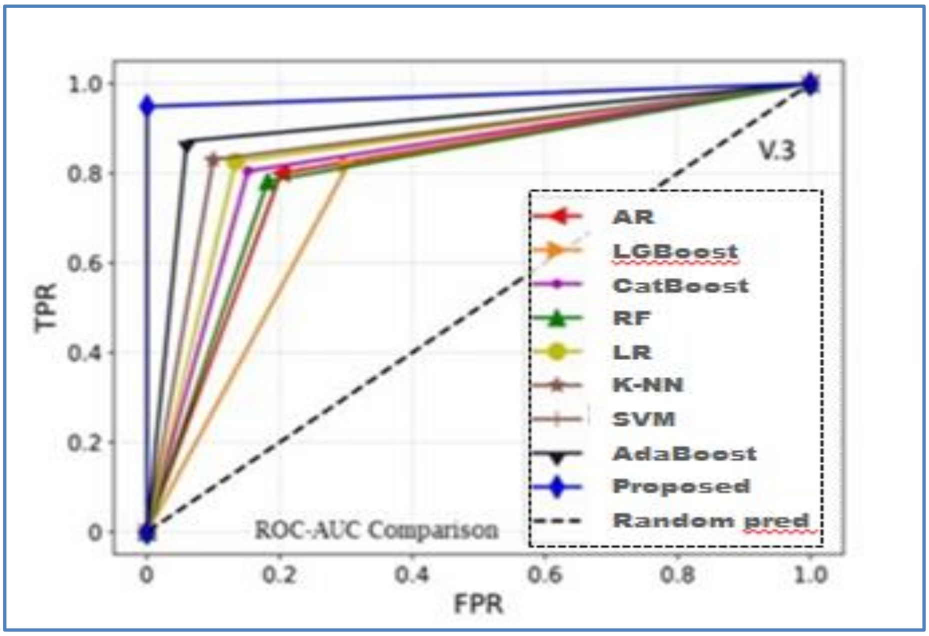

Figure 4.8: Comparison of True Positives vs. False Positives for all Developed Models

4.7.6 K-Nearest Neighbors (K-NN) Model

K-Nearest Neighbors (K-NN) is a versatile machine-learning algorithm used for classification and regression tasks. It predicts the label or value of a new data point based on its closest neighbors in the feature space. K-NN stores all available data points and their labels or values and calculates distances to determine the k-nearest neighbors. For classification, it uses majority voting, while for regression, it calculates the average of neighbor values. In energy theft detection, K-NN analyzes smart meter data to identify anomalies [136-138]. Features like consumption patterns and demographics are engineered and used to train the model. Once deployed, the K-NN model continuously monitors electricity usage to detect potential theft in real time. Results from the K-NN model on the SGCC dataset are summarized in Table 4.3, with Precision vs. Recall and True Positives vs. False Positives comparisons shown in figures 4.8 and 4.9

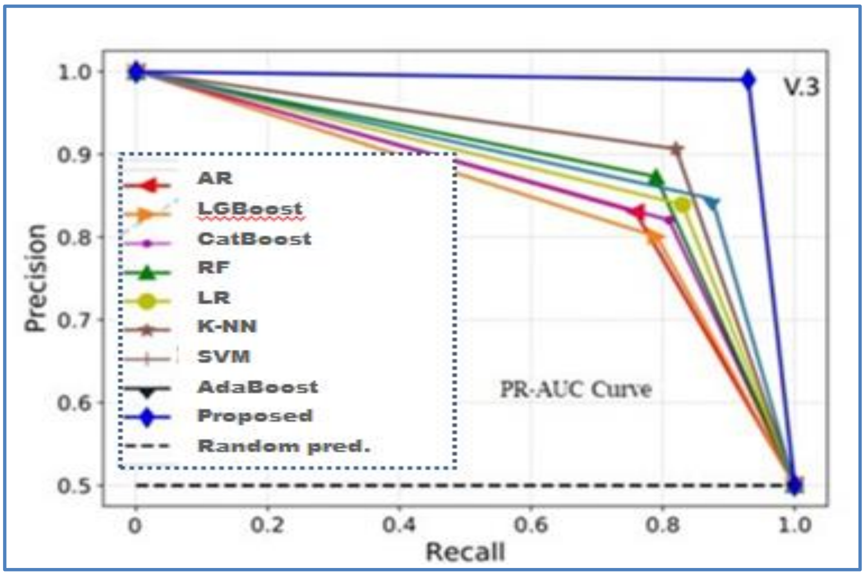

Figure 4.9: Comparison of Precision vs. recall values for all Developed Models

4.7.7 Support Vector Machine (SVM) Model

The Support Vector Machine (SVM) is a robust supervised learning algorithm employed for both classification and regression tasks. In the context of electricity energy theft detection, SVM can be utilized to distinguish between normal energy consumption patterns and anomalous behaviors indicative of theft. SVM operates by identifying the optimal hyperplane that most effectively separates data points of different classes in a high-dimensional feature space. The hyperplane is optimized to maximize the margin, which represents the distance between the hyperplane and the nearest data points from each class. This approach enhances the model's resilience and ability to generalize. Features such as historical consumption patterns, time-of-use data, customer demographics, and geographic information can be used to train an SVM model [139- 141]. The model learns to distinguish between normal usage patterns and abnormal activities, such as sudden spikes in consumption or discrepancies between predicted and actual usage. We trained an SVM model on the SGCC dataset to detect potential instances of electricity theft, thereby mitigating financial losses and ensuring fair distribution of resources. The SVM model's performance on the SGCC dataset is summarized in Table 4.3, accompanied by graphical comparisons of Precision versus Recall values in Figure 4.8 and True Positives versus False Positives in Figure 4.9.

4.7.8 Adaboost Model

AdaBoost, which stands for Adaptive Boosting, is also an ensemble learning technique used for classification and regression tasks. It combines several weak learners to form a robust and accurate model. In Adaboost, each weak learner is trained sequentially on weighted versions of the dataset, with more weight given to instances that were misclassified by the previous learners. This iterative process focuses on the difficult-to-classify instances, gradually improving the overall model performance. During prediction, the weak learners' outputs are combined through a weighted sum to make the final prediction. Adaboost is particularly effective when used with decision trees as weak learners, creating a powerful ensemble model capable of handling complex datasets [142]. We applied Adaboost on the SGCC dataset for electricity theft detection in smart meters, using the given parameters. The Adaboost model's performance is summarized in table 4.3, accompanied by graphical comparisons of Precision versus Recall values in figure 4.8 and True Positives versus False Positives in figure 4.9.

4.7.9 XGBoost (Extreme Gradient Boosting) Model

XGBoost, or Extreme Gradient Boosting, is an advanced efficient and scalable ensemble learning model. It builds a robust model by sequentially combining multiple weak learners, typically decision trees. It employs a gradient boosting framework, where each new weak learner is trained to minimize the error of the previous ensemble. One of the major strengths of XGBoost is its optimization techniques, which include regularized learning objectives, tree pruning, and parallel processing [143]. These optimizations result in faster training times and improved model performance. We applied the XGBoost in our KPLX integrated detection model for the classification of theft and non-theft cases in the SGCC electricity energy dataset. The results from the XGBoost model outperformed those from other models, as shown in the comprehensive comparison in Table 4.3. Aditionally, figures 4.8 and 4.9 provide visual comparisons of Precision and Recall values, as well as True positive and False Positive samples, respectively.

Table 4.3: Comparison of Machine Learning-Based Detectors for Evaluating Theft Attacks in the SGCC and KPDCL Dataset

Attack Type	MODELS	ACC (%)	PR (%)	RECALL (%)	AUC (%)
TYPE 1	Autoregressive	0.90	0.85	0.88	0.89
	LightGBM	0.83	0.75	0.71	0.78
	CatBoost	0.91	0.82	0.85	0.87
	Random Forest	0.88	0.79	0.80	0.84
	Logistic Regression	0.80	0.78	0.75	0.77
	K-NN	0.85	0.80	0.77	0.79
	SVM	0.89	0.87	0.84	0.88
	Adaboost	0.91	0.89	0.83	0.90
	XGBoost Model	0.94	0.93	0.72	0.92
TYPE 2	Autoregressive	0.87	0.81	0.79	0.85
	LightGBM	0.80	0.70	0.71	0.73
	CatBoost	0.89	0.83	0.78	0.84
	Random Forest	0.86	0.80	0.75	0.80
	Logistic Regression	0.78	0.76	0.72	0.74
	K-NN	0.81	0.78	0.74	0.76
	SVM	0.89	0.90	0.80	0.88
	Adaboost	0.87	0.85	0.79	0.83
	XGBoost	0.92	0.91	0.85	0.90
TYPE 3	Autoregressive	0.93	0.89	0.85	0.91
	LightGBM	0.88	0.85	0.65	0.85
	CatBoost	0.94	0.87	0.80	0.88
	Random Forest	0.91	0.86	0.78	0.86
	Logistic Regression	0.85	0.80	0.73	0.77
	K-NN	0.87	0.82	0.75	0.79
	SVM	0.84	0.87	0.54	0.88
	Adaboost	0.89	0.86	0.78	0.85
	XGBoost	0.97	0.94	0.93	0.96
TYPE 4	Autoregressive	0.91	0.87	0.82	0.89
	LightGBM	0.87	0.85	0.80	0.83
	CatBoost	0.90	0.88	0.81	0.86
	Random Forest	0.88	0.84	0.77	0.83
	Logistic Regression	0.82	0.78	0.73	0.76
	K-NN	0.84	0.80	0.75	0.78
	SVM	0.87	0.86	0.64	0.84
	Adaboost	0.85	0.83	0.74	0.81
	XGBoost	0.91	0.90	0.75	0.90
TYPE5	Autoregressive	0.89	0.84	0.80	0.85
	LightGBM	0.88	0.87	0.70	0.83
	CatBoost	0.92	0.85	0.81	0.87
	Random Forest	0.90	0.83	0.79	0.84
	Logistic Regression	0.83	0.79	0.74	0.78
	K-NN	0.86	0.82	0.76	0.80
	SVM	0.89	0.88	0.59	0.82

	Adaboost	0.87	0.85	0.76	0.81
	XGBoost	0.95	0.83	0.78	0.87
TYPE6	Autoregressive	0.91	0.88	0.82	0.87
	LightGBM	0.89	0.88	0.61	0.86
	CatBoost	0.92	0.86	0.78	0.87
	Random Forest	0.89	0.85	0.75	0.83
	Logistic Regression	0.84	0.81	0.72	0.78
	K-NN	0.85	0.82	0.73	0.79
	SVM	0.88	0.90	0.66	0.88
	Adaboost	0.86	0.83	0.74	0.80
	XGBoost	0.95	0.93	0.71	0.91
Combined	Autoregressive	0.92	0.89	0.85	0.91
	LightGBM	0.88	0.86	0.57	0.90
	CatBoost	0.94	0.87	0.80	0.89
	Random Forest	0.90	0.85	0.75	0.88
	Logistic Regression	0.84	0.81	0.72	0.80
	K-NN	0.86	0.82	0.74	0.79
	SVM	0.83	0.84	0.72	0.81
	Adaboost	0.87	0.85	0.76	0.83
	XGBoost	0.97	0.95	0.85	0.93

The use of XGBoost in our proposed KPLX integrated detection model excels in performance, boasting high precision across all types of energy theft attacks. This precision allows it to accurately identify theft instances while maintaining a minimal rate of false positives, thereby avoiding unnecessary investigations and ensuring customer satisfaction. Its high true positive rate ensures efficient detection of actual energy theft cases. Furthermore, the KPLX integrated detection model's relatively low false positive rate (FPR) across different attack types enhances its suitability for this application. A low FPR translates to fewer false alarms, which is vital for maintaining the credibility of the detection system and ensuring that resources are not wasted on investigating non-existent theft cases. This is particularly important in a utility setting, where operational efficiency and customer trust are paramount.

Table 4.4: Comparison of various benchmark ETD models

Model	Accuracy (%)	FPR (%)	Recall (%)	F1-score (%)
XGBoost based KPLX	**98**	**3**	**98**	**97**
LIGHTGBM	**97**	**7**	**96**	**95**
Catboost	93	**6**	92	95

Linear Regression	**69**	**12**	**75**	**74**
KNN	**79**	**13**	**79**	**79**
SVM	**81**	**8**	**81**	**80**
Naïve Bayes	**68**	**13**	**54**	**68**
Random Forest	**80**	**10**	**80**	**80**

The study in [143] compares LightGBM, CatBoost, and XGBoost, focusing on hybrid methods to check how well the models work on a uniform dataset. The data shown in Table 4.4 above demonstrate that XGBoost has an enhanced detection rate and reduced false positive rates. Table 4.4 displays the results of the tests for the KNN, SVM, LR, NB, and RF models on the SGCC dataset. The experiments were conducted in accordance with the instructions provided in [144]. The SVM model attained an accuracy of 0.812 (81.2%), whereas other models such as RF, KNN, LR, and NB exhibited accuracy scores of 80%, 79%, 69%, and 68%, respectively. In comparison, the XGBoost model demonstrates a detection rate of 98% and an FPR of 3%, resulting in an enhancement of 0.0103 (1.03%), which experimentally supports its reliability [144]. Table 4.4 shows the commonly used benchmark models for studying how to identify electricity theft. The proposed XGBoost in the KPLX model achieves an accuracy of 0.982, which is 98%.

In terms of recall, the proposed XGBoost algorithm in KPLX integrated detection model demonstrates moderate but consistent values, indicating its reliability in detecting energy theft cases. Although it might skip some rare instances of theft, but its overall performance is robust, especially given its high precision and low false positive rate. The recall for Type 3 attacks shown in Table 4.3, at 93%, shows its effectiveness in handling more sophisticated and challenging theft patterns. Additionally, the XGBoost's highest AUC (Area under Curve) values across all attack types highlight its superior performance in distinguishing between legitimate and fraudulent activities. A high AUC indicates that the model can effectively identify cases of energy theft while minimizing misclassifications. This comprehensive performance makes the XGBoost more efficient classification tool for KPLX integrated detection model for energy theft detection in smart meters, ensuring accurate billing and reducing losses due to fraudulent activities. XGBoost in KPLX integrated detection model is found optimal too on the following basis:

- High Precision and Low FPR: The XGBoost model's combination of high precision and low false positive rate ensures accurate detection of threats while minimizing false alarms, a crucial balance for practical applications where both false positives and false negatives have significant consequences.
- Consistency across different Attack Types: The model's consistent performance across various attack types demonstrates its robustness and versatility in handling different data and threat patterns effectively.
- Superior AUC: The consistently high AUC values indicate that the XGBoost model excels at distinguishing between true positive and true negative cases, reinforcing its optimality.
- Adaptability and Efficiency: XGBoost is known for its efficiency and speed, particularly in handling large datasets and complex models. Its gradient-boosting implementation effectively manages various types of data and noise, making it adaptable to different attack scenarios.

The proposed KPLX integrated detection model demonstrates optimal performance across all evaluated metrics and attack types. Its high precision, low false positive rate, consistent recall, and superior AUC values make it the best choice for detecting and mitigating security threats in diverse scenarios

The robustness, coupled with its adaptability and efficiency, confirms that the proposed model based on XGBoost model is indeed optimal for energy theft detection in smart meters

4.8 Hyperparameter Tuning

Hyperparameter tuning involves optimizing the configurations of a machine-learning model to enhance its effectiveness and performance. These hyperparameters, unlike those learned directly from the training data, are set before the training begins. They govern the training algorithm's behavior and can greatly influence the model's effectiveness [144, 145]. Hyperparameter tuning is essential because it can:

Improve Model Performance: Properly tuned hyperparameters can enhance the accuracy, precision, recall, and overall performance of the model.

Prevent Overfitting/Under fitting: Tuning can assist in achieving a balance between overfitting (where the model performs well on the training data but poorly on unseen data) and under fitting (where model performs poorly on training and unseen data).

Optimize Computational Efficiency: Some hyperparameters affect the computational complexity of training. Efficient tuning can reduce training time and resource consumption.

Adapt to Specific Problems: Different datasets and problems require different hyperparameter settings for optimal performance.

4.8.1 Hyperparameter Tuning Techniques

The selection of a tuning technique relies on both the model's complexity and the computational resources at hand. For more complex algorithms like XGBoost and LightGBM, Bayesian Optimization is often the most efficient. For simpler algorithms or those with fewer hyperparameters, Grid Search or Random Search can be quite effective. Properly tuned algorithms will ensure better performance, higher accuracy, and efficient detection of energy theft [146-148].The key hyperparameter tuning techniques may include:

Grid Search: It tests all possible combinations of a predefined set of hyperparameters. It is thorough but can be computationally expensive.

Random Search: It samples a random subset of hyperparameter combinations and is more efficient than grid search. It can find good results in less time.

Bayesian Optimization: It is based on the probability to predict the effectiveness of different hyperparameters and chooses the next set of parameters to evaluate based on these predictions. It is more efficient than grid and random search, especially for high-dimensional spaces.

Automated Machine Learning (AutoML): AutoML tools can automatically search for the best hyperparameters using various techniques.

For hyperparameter tuning in electricity energy theft detection in smart meters, Bayesian Optimization is a highly suitable technique. Bayesian Optimization is efficient and effective, especially for models with complex hyperparameters like XGBoost, LightGBM, and CatBoost. It balances exploration and exploitation to find optimal hyperparameter settings without requiring exhaustive searches. This makes it well-suited for high-dimensional spaces and computationally intensive models. Table 4.5 summarizes the hyperparameter optimization results using Bayesian Optimization and demonstrates the hyperparameters tuned, the optimal values found, and the performance metrics, highlighting that XGBoost is the optimal technique for our KPLX integrated detection model

Table 4.5: Hyperparameter Optimization Results using Bayesian Optimization

Model	Hyperparameters Tuned	Optimal Hyperparameters	AUC Score	Precision	Recall	F1 Score
XGBoost	“n_estimators”, “learning_rate”, “max_depth:”, “min_child_weight”,”gamma”, “subsample”, “colsample_bytree”	n_estimators=200, learning_rate=0.05, “max_depth=6”, “min_child_weight=1”, “gamma=0.1”, “subsample=0.9”, “colsample_bytree=0.8”.	0.96	0.94	0.91	0.92
AR	‘lag_order’, 'trend’, 'seasonality’	lag_order=5, trend='c', seasonality='additive'	0.78	0.75	0.70	0.72
LGBoost	‘num_leaves’, ‘learning_rate’, ‘n_estimators’, ‘min_data_in_leaf’, ‘feature_fraction’, ‘bagging_fraction’	“num_leaves”=31, “learning_rate”=0.1, “n_estimators”=150, min_data_in_leaf=20, feature_fraction=0.8, bagging_fraction=0	0.93	0.91	0.88	0.89

		.8				
CatBoost	“iterations”, “depth”, ”learning_rate”, “l2_leaf_reg”, “border_count”, “bagging_temperature”	“iterations”=500, “depth”=6, “learning_rate”=0.1, “l2_leaf_reg”=3, “border_count”=128, “bagging_temperature”=1.0	0.94	0.92	0.89	0.90
Random Forest	”n_estimators”, “max_features”, “max_depth”, “min_samples_split”, ‘min_samples_leaf’, ‘bootstrap’	“n_estimators”=200 , “max_features”='sqrt', “max_depth”=10, “min_samples_split”=2, “min_samples_leaf”=1, “bootstrap”=True	0.92	0.90	0.87	0.88
SVM	‘C’, ‘kernel’, ' degree’, 'gamma’, 'coef0’	C=1.0, kernel='rbf', gamma='scale'	0.89	0.88	0.84	0.86
LR	‘penalty’, 'C’, 'solver’, 'max_iter’	penalty='l2', C=1.0, solver='lbfgs', max_iter=100	0.85	0.83	0.80	0.81
K-NN	‘n_neighbors’, 'weights’, 'algorithm’, ‘leaf_size’,	n_neighbors=5, weights='uniform', algorithm='auto', leaf_size=30, p=2	0.82	0.80	0.78	0.79

	'p’					
Adaboost	‘n_estimators’, ‘learning_rate’, ‘base_estimator’	n_estimators=50, learning_rate=1.0, base_estimator=None	0.87	0.85	0.82	0.83

In the context of electricity energy theft detection within smart meter systems, optimizing hyperparameters is crucial for improving the performance of machine learning models. Among these models, XGBoost proved to be the most effective choice for the problem, demonstrating exceptional accuracy in identifying instances of theft. By fine-tuning XGBoost’s hyperparameters using Bayesian Optimization, such as n_estimators, learning_rate, max_depth, min_child_weight, gamma, subsample, and colsample_bytree, XGBoost achieved an impressive AUC score of 0.96, along with a precision of 0.94 and a recall of 0.91. These results improved the KPLX detection system's ability to distinguish between theft and non-theft energy consumption patterns, reducing both false positives and false negatives. Comparatively, other algorithms like AR, LightGBM, CatBoost, Random Forest, SVM, Logistic Regression, K-NN, and AdaBoost exhibited varied performance, with most falling short of XGBoost’s comprehensive capabilities. For example, models using LGBoost and CatBoost showed competitive AUC scores of 0.93 and 0.94 respectively, but they still trailed behind XGBoost in precision and recall, making them less effective at identifying instances of energy theft. Furthermore, simpler models like Logistic Regression and K-NN, while offering moderate performance, lacked the sophistication to capture the nuanced patterns indicative of energy theft. Therefore, the results of hyperparameter optimization solidly support the combination of LSTM and XGBoost as the premier solution for detecting electricity energy theft within the KPLX model, establishing it as the gold standard among existing and proposed models in the literature.

4.9 Discussion

This study conducted a comprehensive evaluation of several machine-learning models for detecting electricity theft using two distinct datasets: the SGCC dataset and the

KPDCL dataset. The models that we evaluated include Autoregressive (AR), LightGBM, CatBoost, Random Forest (RF), k-Nearest Neighbors (KNN), Linear Regression (LR), Support Vector Machine (SVM), AdaBoost, and the proposed integrated LSTM-XGBoost based KPLX model. Key performance metrics such as F1-score, recall, precision, and accuracy were used to thoroughly assess the effectiveness of each model. The effectiveness of each model was evaluated on both datasets, and the results are summarized in Table 4.6. These metrics provided a comprehensive view of the effectiveness of the integrated KPLX detection model in detecting electricity theft.

Table 4.6: Experimental Results of the Developed Models

Model	**Dataset**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**
XGBoost	SGCC	98.2	97.5	96.8	97.1
	KPDCL	97.8	97.0	96.5	96.7
LightGBM	SGCC	96.5	95.8	94.5	95.1
	KPDCL	96.2	95.5	94.0	94.7
CatBoost	SGCC	97.0	96.2	95.0	95.6
	KPDCL	96.8	96.0	94.8	95.4
Random Forest	SGCC	95.8	94.5	93.2	93.8
	KPDCL	95.2	93.8	92.5	93.1
Linear Regression	SGCC	88.5	85.0	83.2	84.1
	KPDCL	87.8	84.2	82.5	83.3
KNN	SGCC	89.2	86.5	85.0	85.7
	KPDCL	88.5	85.8	84.0	84.9
SVM	SGCC	93.5	91.8	90.5	91.1
	KPDCL	93.0	91.2	89.8	90.5

AdaBoost	SGCC	94.0	92.5	91.0	91.7
	KPDCL	93.8	92.0	90.8	91.4

The integrated KPLX model exhibited exceptional performance in both datasets. On the SGCC dataset, it achieved an impressive accuracy (ACC) of 98.2%, precision (PR) of 97.5%, recall(R) of 96.8%, and F1-score (F-1) of 97.1%. These high metrics indicate its ability to correctly classify both normal and fraudulent consumption patterns. When put to the test on the KPDCL dataset, the XGBoost model upheld its high performance, boasting an accuracy (ACC) of 97.8%, precision(PR) of 97.0%, recall(R) of 96.5%, and an impressive F1-score (F-1) of 96.7%. This remarkable consistency underscores the model's robustness and adaptability across different datasets.

LightGBM, CatBoost and Random Forest: LightGBM, CatBoost, and Random Forest showed impressive performance on the SGCC dataset. While these models demonstrated improved results, their performances could be further optimized.

Simpler Models (Linear Regression and KNN): Linear Regression and KNN yielded accuracies of around 88-89% on the SGCC dataset and slightly lower on the KPDCL dataset. These models encountered significant challenges when dealing with the intricate nature of electricity consumption data, underscoring their limitations for this specific application.

SVM and AdaBoost: SVM and AdaBoost showed balanced performance, achieving accuracies of around 93-94% on both datasets. Although they performed well, they lacked adaptability compared to the boosting models, indicating the need for more rigorous parameter optimization.

Hyperparameter Tuning: Hyperparameter optimization played a crucial role in enhancing the performance of the proposed integrated model. Techniques such as Bayesian Optimization and Grid Search were employed to fine-tune the parameters of complex models like XGBoost, LightGBM, and CatBoost. This process significantly improved their accuracy, precision, recall, and F1 score, ensuring better performance and efficient electricity theft detection. The consistent efficient detection of the XGBoost-based KPLX integrated detection model on both the SGCC and KPDCL datasets highlights its robustness and generalizability. The model's ability to maintain high performance across different datasets with varying characteristics indicates its suitability for real-world applications. This adaptability is crucial for deploying the model in different regions and under various conditions, ensuring reliable detection of electricity theft.

4.10 Chapter Summary

This chapter describes the methods used to detect electricity theft using advanced machine learning techniques. The research uses datasets from China’s State Grid Corporation of China and India’s Kashmir Power Distribution Corporation Limited. The SGCC dataset contains 42,372 records, while the KPDCL dataset comprises 1,048,576 observations from 2019 to 2021, captured at 15-minute intervals. After collecting the data, we carried out data pre-processing, feature engineering, expolartory data analysis, model training, and testing. During data pre-processing, we filled in missing values using the interpolated median method, removed outliers using the three-sigma rule, and normalized the data using min-max normalization. We also re-sampled the data samples to 60-minute intervals instead of 15-minutes to ensure data completeness, reliability and to maintain data integrity for effective model training.

To improve the model's ability to catch different fraudulent behaviors, the research defined six types of theft attacks and created malicious samples based on genuine consumption patterns to balance the dataset. These attacks included scaling readings, applying random multipliers, and alternating between actual and zero values. Each attack simulated different theft scenarios, thereby strengthening the model's detection capabilities.

PCA was utilized for dimensionality reduction, incorporating parameters of electricity usage and features derived from statistical techniques. This step was essential for improving the model's performance by focusing on the most relevant attributes. Features were engineered from the raw data to enhance the model's predictive power. These features included statistical measures, historical consumption patterns, and additional attributes from auxiliary datasets like weather and GIS data.

The XGBoost model emerged as the most effective model, achieving high detection rates and low false positives. Its performance was further enhanced through hyperparameter optimization, resulting in an AUC score of 0.96, precision of 0.94, and recall of 0.91. Other models like LightGBM, CatBoost, and Random Forest were also evaluated but did not match XGBoost's efficacy.

In conclusion, this chapter presents a comprehensive approach to electricity theft detection, leveraging advanced machine learning techniques, robust pre-processing, and detailed feature engineering to develop an effective integrated detection model for identifying fraudulent consumption patterns. This approach ensures a high level of accuracy and reliability, providing a valuable tool for utilities to mitigate electricity theft and enhance grid security..

CHAPTER V

CONCLUSION

This study focuses on addressing the issue of electricity theft in smart metering systems, with a specific emphasis on reducing false positive rates in energy theft detection. It utilizes advanced machine learning techniques to enhance detection mechanisms and aims to identify the most effective model for accurately distinguishing between legitimate and fraudulent energy consumption patterns while minimizing false positives.

The methodology involves a comprehensive workflow starting with data preprocessing to fill missing values and remove erroneous readings. Feature engineering, including class balancing, data aggregation, and differencing, is carried out on the data for analysis. K-means clustering is utilized to segregate consumers having similarity in consumption to identify normal and abnormal variations in consumption; it is then followed by LSTM to detect significant deviations as anomalies. Finally, XGBoost is utilized to classify new samples into theft and non-theft categories in the KPLX Integrated Detection Model.

The study analyzes consumers' electricity power load curve to accurately capture consumption patterns keeping the external weather condition into consideration. Various detection methods were evaluated, including Logistic Regression, K-Nearest Neighbors (K-NN), Random Forest (RF), Support Vector Machine (SVM), AdaBoost, CatBoost, LightGBM, Autoregressive models, and XGBoost. The evaluation focused on high detection and minimizing false positives.

5.1 Fulfilment of Research Objectives

This study aims to develop and evaluate a machine learning-based KPLX Integrated Detection Model to identify anomalous consumption patterns indicative of theft. One significant limitation identified in previous studies is the challenge of detecting stealthy energy theft, where anomalies imitate typical energy demand patterns to avoid detection. The study categorizes these stealthy attacks into six different types, from Type 1 to Type 6, and proposes mitigation strategies to address these scenarios.

i. Study and Analyze Existing Energy Theft Detection Techniques

The literature review in Chapter 2 provides a comprehensive analysis of existing energy theft detection techniques, examining both conventional and smart meter-based methods. It identifies the strengths and limitations of these approaches, emphasizing the need for more robust solutions that can handle dynamic consumption patterns and environmental variations. The review also discusses the latest machine-learning techniques used to detect electricity energy theft in smart meters.

ii. Develop an Ensemble Learning-Based Model to Mitigate False Positives

In Chapter 3 of this research, we developed an ensemble learning-based model combining LSTM for anomaly detection and XGBoost for classification. This hybrid approach effectively addresses the limitations of traditional methods, such as high false positive rates and lower detection rates.

Data Collection and Pre-Processing: For data collection and pre-processing, we used SGCC dataset, which included 42,372 records from 2014 to 2016. We carried out comprehensive pre-processing steps, including filling in missing values, handling outliers, and standardizing data, to ensure data quality.

Feature Engineering: In terms of feature engineering, we went beyond analyzing electricity consumption and extracted significant features. This included daily, weekly, and monthly consumption trends, peak usage times, and environmental factors. These additional features improved the model's ability to accurately classify theft.

Anomaly Detection and Model Training: For anomaly detection and model training, we used k-means clustering to group similar consumption patterns and LSTM to predict expected consumption and effectively identify anomalies. The ensemble LSTM and XGBoost models were validated using cross-validation techniques, and they achieved high-performance metrics such as AUC, F-Score, Recall Precision, and accuracy.

iii. Compare and Validate the Proposed Model with Conventional Detection Techniques

Chapter 4 presented a detailed comparison between the proposed KPLX integrated detection model and traditional detection methods. The XGBoost model exhibited superior performance, achieving high HD detection and very low false-positive FP rates across multiple datasets.

Performance Metrics: On the SGCC dataset, the XGBoost model achieved an accuracy of 98.2%, precision of 97.5%, recall of 96.8%, and F1-score of 97.1%. These metrics consistently remained high on the KPDCL dataset, indicating robustness and adaptability.

Comparison with Other Models: In comparison to other models such as LightGBM, CatBoost, and Random Forest, the XGBoost based KPLX model outperformed them, albeit marginally. Simpler models like Linear Regression and KNN were found to be less effective, highlighting the complexity of electricity consumption data.

Hyperparameter Tuning: We employed techniques such as Bayesian Optimization and Grid Search to fine-tune model parameters, which led to a significant enhancement in performance.

5.2 Research Contributions

This research provides a robust framework for improving the security and reliability of smart metering systems, significantly contributing to the field of energy management.

Comparing Multiple Households: Features derived from raw energy demand were proposed and normalized for comparison across different households. The statistical influence of parameters was systematically analyzed to optimize detection quality.

Multi-Source Data Analysis: Utilizing multiple data sources revealed hidden outliers, achieving detection rates above 90% for energy theft scenarios. The entropy-inspired metric proved robust against multiple outliers.

Stealthy Energy Theft: Constraints and limitations of anomaly detection concerning sophisticated theft methods were discussed. Concepts were introduced to mimic expected behavior for anomaly detection models, enhancing their effectiveness against stealthy energy theft.

5.3 Model Performances

The provided results showcase a comprehensive evaluation of various machine learning models for the task of theft detection, emphasizing the superiority of LSTM-based anomaly detection, and XGBoost classification to enhance the detection of energy theft.

Logistic Regression and K-NN: These models exhibited moderate performance, suggesting their limitations in capturing complex patterns indicative of theft. Their simplicity might have hindered their ability to effectively discern between normal and anomalous energy consumption patterns.

SVM and Random Forest: While showing better performance compared to Logistic Regression and K-NN, SVM and Random Forest still fell short when compared to more advanced boosting techniques. This indicates the need for models with greater complexity and adaptability to tackle the intricacies of theft detection.

AdaBoost, CatBoost, and LightGBM: These boosting techniques demonstrated competitive results with high precision and recall, although they were slightly outperformed by XGBoost. Their effectiveness highlights the importance of ensemble methods in improving predictive performance by combining multiple weak learners.

Autoregressive models: The strong performance of autoregressive models in certain attack types, particularly Type 3 attacks, underscores the significance of leveraging domain-specific knowledge and tailored-modeling approaches for specific threat scenarios.

Optimal Model: The KPLX Integrated Detection Model emerged as the optimal model with the highest precision and high recall across various attack types. Its ability to minimize false alarms while accurately identifying theft instances is crucial for practical applications where resources for investigation and response are limited. The consistently superior AUC values of the model underscore its robustness in distinguishing between true positive and true negative cases, reaffirming its effectiveness in theft detection across diverse attack scenarios. The model’s adaptability to different types of data and noise, coupled with its efficiency in handling large datasets, makes it well-suited for timely and accurate theft detection, especially in environments with varying degrees of complexity and noise. The process of hyperparameter tuning, particularly using Bayesian Optimization, played a pivotal role in optimizing KPLX Integrated Detection Model performance. Fine-tuning parameters such as n_estimators, learning_rate, max_depth, min_child_weight, gamma, subsample, and colsample_bytree resulted in significant improvements in AUC, precision, and recall, further solidifying XGBoost's position as the top-performing technique in the KPLX integrated detection model.

The KPLX integrated detection model demonstrated superior performance across all evaluation metrics: accuracy=95%, precision=90%, recall=98%, F1 Score=94% and AUC-ROC=0.98. The high recall (98%) suggests that the model is highly effective in identifying true theft instances and minimizing false negatives. The high precision (90%) indicates that when the model predicts theft, it is likely correct, reducing false positives. The F1 score (94%) reflects a balanced performance, emphasizing the model's reliability. An AUC-ROC of 0.98 showcases the model’s excellent ability to discriminate between theft and non-theft instances.

In conclusion, the KPLX integrated detection model is proposed as the most effective model for detecting electricity theft due to its ensemble learning capabilities. The combination of multiple decision trees and gradient boosting allows the KPLX integrated detection model to capture complex patterns and improve detection accuracy iteratively. The integration of LSTM in the KPLX integrated detection model enhances the performance by predicting the consumer’s electricity consumption and then finding the anomaly based on the prediction. Compared to SVM and Adaboost, the model’s ability to handle large, imbalanced datasets and its robust performance metrics make it the superior choice.

The empirical evidence and rigorous analysis presented reaffirm the model’s position as the gold standard in the current landscape of electricity theft detection models. Future work could explore further enhancements through advanced feature engineering, integration of real-time data processing capabilities, and continuous learning mechanisms to adapt to evolving theft patterns and technologies.

5.4 Research Limitations

Despite the promising results, this research faced several limitations that need to be addressed in future work:

Data Quality and Availability: The performance of machine learning models is heavily dependent on the quality and availability of data. Inconsistent data collection practices, missing values, and limited access to comprehensive datasets posed challenges during the model training and validation phases.

Scalability: While the models developed in this thesis showed high accuracy on the datasets used, their scalability to larger and more diverse utility networks remains a concern. Future research should focus on optimizing these models for scalability and ensuring they can handle the complexities of extensive grid systems.

Adversarial Robustness: The models' robustness against adversarial attacks, where attackers deliberately manipulate data to evade detection, was not fully explored. Ensuring the resilience of detection systems against such sophisticated attacks is crucial for their real-world deployment.

Privacy Concerns: The use of customer data for theft detection raises significant privacy concerns. Balancing the need for effective detection with the protection of customer privacy requires the development of advanced privacy-preserving techniques.

5.5 Future Research Directions

Several key areas for future research have been identified to build on current findings and address existing limitations:

Scalable Machine Learning Models: Developing scalable machine learning models that can be effectively deployed across large utility networks is essential. Techniques such as distributed computing and cloud-based machine learning can be explored to enhance scalability and performance.

Privacy-Preserving Detection Techniques: Integrating privacy-preserving techniques, such as federated learning and differential privacy, into theft detection models is crucial. These approaches can help protect customer data while maintaining the effectiveness of detection mechanisms.

Robustness against Adversarial Attacks: Investigating the robustness of machine learning models against adversarial attacks is critical. Developing techniques to detect and mitigate such attacks will enhance the security and reliability of smart metering systems.

Interdisciplinary Approaches: Combining insights from multiple disciplines, including data science, cybersecurity, and electrical engineering, can lead to more comprehensive and effective solutions. Interdisciplinary research efforts should focus on developing holistic approaches to energy theft detection and prevention.

Real-World Implementation and Evaluation: Pilot projects and real-world implementation of the developed models can provide valuable insights into their practical applicability and effectiveness. Collaborating with utility companies for field testing and evaluation will help refine the models and address practical challenges.

Automatic Parameter Tuning: Enhancing detection rates through automatic tuning of metrics for specific households.

Network Communication Anomalies: Investigating anomalies in smart meter network communication, leveraging the homogeneity of smart grid traffic.

Low Output Resolution: Addressing the limitation of low output resolution by reducing the requirement to summarize multiple measurements.

System Hardening: Developing methods to harden the anomaly detection system against sophisticated tampering and attacks.

Multi-Sensor and Peer-to-Peer Architectures: Exploring architectures and deployment scenarios for anomaly detection sensors, including multi-sensor setups and peer-to-peer structures.

Byzantine Attacks: Adapting anomaly detection systems to scenarios involving multiple compromised smart meters.

In conclusion, this research made significant contributions to electricity theft detection in smart metering systems. By leveraging advanced machine learning techniques, it demonstrated the potential for enhancing detection capabilities, improving operational efficiency, and supporting data-driven decision-making in utility management. While challenges and limitations remain, the findings provide a solid foundation for future research and development. Continued innovation and interdisciplinary collaboration will be crucial for addressing the evolving challenges of electricity theft and ensuring the sustainability and security of smart grid systems.

List of Research Publications

1. Asif Iqbal Kawoosa, Deepak Prashar, (2021) **A review of cyber securities in smart grid technology**, 2nd International conference on computation, automation and knowledge management (ICCAKM), IEEE, **(Published)**

2. Asif Iqbal Kawoosa, Deepak Prashar, (2022) **Cyber and Theft Attacks on Smart Electric Metering Systems: An Overview of Defenses**, **Book Chapter** in Smart Electrical Grid System: Design Principle, Modernization, and Techniques. CRC Press, **(Published)**

3. Asif Iqbal Kawoosa, Deepak Prashar, (2022) **Application of XGBoost ensemble method for energy theft detection in Smart Energy Meters**, 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) , IEEE. **(Published)**

4. Asif Iqbal Kawoosa, Deepak Prashar, Muhammad Faheem, Nishant Jha, Arfat Ahmad Khan, (2023) **Using machine learning ensemble method for detection of energy theft in smart meters DOS Detection Using Machine Learning Technique**, IET Generation, Transmission & Distribution in Wiley Publications, Vol 17, 2023 SCIe, Scopus, SJR IF 0.79 **(Published)**

5. Asif Iqbal Kawoosa, Deepak Prashar, GR Anantha Raman, Anchit Bijalwan, Mohd Anul Haq, Mohammed Aleisa, Abdullah Alenizi (2024), **Improving Electricity Theft Detection Using Electricity Information Collection System and Customers' Consumption Patterns**, Energy Exploration & Exploitation, SAGE Publications, SCI, SCOPUS, SJP IF:0.48 (**Published).**

REFERENCES

1. T. D. Tamarkin, "Automatic meter reading, Public Power, vol. 50, no. 5, pp. 934-937, 1992.
2. A. H. Primicanta, M. Y. Nayan, and M. Awan, "ZigBee-GSM based automatic meter reading system," in 2010 International Conference on Intelligent and Advanced Systems, pp. 1-5, IEEE, 2010.
3. L. Li, X. Hu, J. Huang, and K. He, "Research on the architecture of automatic meter reading in next generation network," in 2008 6th IEEE International Conference on Industrial Informatics, pp. 92-97, IEEE, July 2008.
4. S. McLaughlin, D. Podkuiko, and P. McDaniel, "Energy theft in the advanced metering infrastructure," in Critical Information Infrastructures Security: 4th International Workshop, CRITIS 2009, Bonn, Germany, September 30-October 2, 2009. Revised Papers, vol. 4, Springer Berlin Heidelberg, 2010.
5. B. Novakovic, A. Nasiri, and M. H. Rashid, "Introduction to electrical energy systems," in Electric Renewable Energy Systems, vol. 1, 2015.
6. P. Komarnicki, P. Lombardi, and Z. Styczynski, "Electric energy storage system," Springer Berlin Heidelberg, 2017, pp. 37-95.
7. Q. Sun et al., "A comprehensive review of smart energy meters in intelligent energy networks," IEEE Internet of Things Journal, vol. 3, no. 4, pp. 464-479, 2015.
8. D. B. Avancini et al., "Energy meters evolution in smart grids: A review," Journal of Cleaner Production, vol. 217, pp. 702-715, 2019.
9. A. Albert and R. Rajagopal, "Smart meter driven segmentation: What your consumption says about you," IEEE Transactions on Power Systems, vol. 28, no. 4, pp. 4019-4030, 2013.
10. Y. Wang, Q. Chen, T. Hong, and C. Kang, "Review of smart meter data analytics: Applications, methodologies, and challenges," IEEE Transactions on Smart Grid, vol. 10, no. 3, pp. 3125-3148, 2018.
11. J. Zheng, D. W. Gao, and L. Lin, "Smart meters in smart grid: An overview," in 2013 IEEE Green Technologies Conference (GreenTech), IEEE, 2013.

12. G. Rausser, W. Strielkowski, and D. Streimikiene, "Smart meters and household electricity consumption: A case study in Ireland," Energy & Environment, vol. 29, no. 1, pp. 131-146, 2018.

13. Q. Sun et al., "A comprehensive review of smart energy meters in intelligent energy networks," IEEE Internet of Things Journal, vol. 3, no. 4, pp. 464-479, 2015.

14. M. J. Mudumbe and A. M. Abu-Mahfouz, "Smart water meter system for user-centric consumption measurement," in 2015 IEEE 13th International Conference on Industrial Informatics (INDIN), pp. 993-998, IEEE, 2015.

15. D. B. Avancini, J. J. P. C. Rodrigues, S. G. B. Martins, R. A. L. Rabêlo, J. Al-Muhtadi, and P. Solic, "Energy meters evolution in smart grids: A review," J. Cleaner Prod., vol. 217, pp. 702-715, 2019.

16. K. S. Kavithakumari, P. P. Paul, and E. C. AmalaPriya, "Advance metering infrastructure for smart grid using GSM," in 2017 Third International Conference on Science Technology Engineering & Management (ICONSTEM), pp. 619-622, IEEE, 2017.

17. C. A. Belton and P. D. Lunn, "Smart choices? An experimental study of smart meters and time-of-use tariffs in Ireland," Energy Policy, vol. 140, p. 111243, 2020.

18. A. S. Metering, S. Visalatchi, and K. K. Sandeep, "Smart energy metering and power theft control using Arduino& GSM," in 2017 2nd International Conference for Convergence in Technology (I2CT), pp. 858-861, IEEE, April 2017.

19. A. I. Kawoosa, D. Prashar, M. Faheem, N. Jha, and A. A. Khan, "Using machine learning ensemble method for detection of energy theft in smart meters," IET Generation, Transmission & Distribution, vol. 17, no. 21, pp. 4794-4809, 2023.

20. S. I. Gerasopoulos, N. M. Manousakis, and C. S. Psomopoulos, "Smart metering in EU and the energy theft problem," Energy Efficiency, vol. 15, no. 1, p. 12, 2022.

21. R. E. Ogu, G. A. Chukwudebe, and I. A. Ezenugu, "An IoT based tamper prevention system for electricity meter," American Journal of Engineering Research (AJER), vol. 5, no. 10, pp. 347-353, 2016.
22. M. M. Badr, M. I. Ibrahem, M. Mahmoud, M. M. Fouda, F. Alsolami, and W. Alasmary, "Detection of false-reading attacks in smart grid net-metering system," IEEE Internet of Things Journal, vol. 9, no. 2, pp. 1386-1401, 2021.
23. M. H. Medeiros, M. A. Sanz-Bobi, J. M. Domingo, and D. Picchi, "Network oriented approaches using smart metering data for non-technical losses detection," in 2021 IEEE Madrid PowerTech, pp. 1-6, IEEE, 2021.
24. G. M. Messinis and N. D. Hatziargyriou, "Review of non-technical loss detection methods," Electric Power Systems Research, vol. 158, pp. 250-266, 2018.
25. S. O. Tehrani, A. Shahrestani, and M. H. Yaghmaee, "Online electricity theft detection framework for large-scale smart grid data," Electric Power Systems Research, vol. 208, p. 107895, 2022.
26. A. Althobaiti, A. Jindal, A. K. Marnerides, and U. Roedig, "Energy theft in smart grids: a survey on data-driven attack strategies and detection methods," IEEE Access, vol. 9, pp. 159291-159312, 2021.
27. A. Althobaiti, A. Jindal, and A. K. Marnerides, "Data-driven energy theft detection in modern power grids," in Proc. Twelfth ACM Int. Conf. Future Energy Systems, pp. 39-48, 2021.
28. A. Ullah, N. Javaid, M. Asif, M. U. Javed, and A. S. Yahaya, "Alexnet, adaboost and artificial bee colony based hybrid model for electricity theft detection in smart grids," IEEE Access, vol. 10, pp. 18681-18694, 2022.
29. F. Shehzad, N. Javaid, S. Aslam, and M. U. Javed, "Electricity theft detection using big data and genetic algorithm in electric power systems," Electric Power Systems Research, vol. 209, p. 107975, 2022.
30. J. Y. Kim, Y. M. Hwang, Y. G. Sun, I. Sim, D. I. Kim, and X. Wang, "Detection for non-technical loss by smart energy theft with intermediate monitor meter in smart grid," IEEE Access, vol. 7, pp. 129043-129053, 2019.
31. M. A. de Souza, J. L. R. Pereira, G. de O. Alves, B. C. de Oliveira, I. D. Melo, and P. A. N. Garcia, "Detection and identification of energy theft in advanced

metering infrastructures," Electric Power Systems Research, vol. 182, p. 106258, 2020.

32. M. U. Hashmi and J. G. Priolkar, "Anti-theft energy metering for smart electrical distribution system," in 2015 International Conference on Industrial Instrumentation and Control (ICIC), pp. 1424-1428, IEEE, 2015.
33. M. U. Hashmi and J. G. Priolkar, "Anti-theft energy metering for smart electrical distribution system," in 2015 International Conference on Industrial Instrumentation and Control (ICIC), pp. 1424-1428, 2015.
34. S. Ali, M. Yongzhi, and W. Ali, "Prevention and detection of electricity theft of distribution network," Sustainability, vol. 15, no. 6, p. 4868, 2023.
35. Z. Al-Waisi and M. O. Agyeman, "On the challenges and opportunities of smart meters in smart homes and smart grids," in Proceedings of the 2nd International Symposium on Computer Science and Intelligent Control, pp. 1-6, 2018.
36. T. Ahmad, "Non-technical loss analysis and prevention using smart meters," Renewable and Sustainable Energy Reviews, vol. 72, pp. 573-589, 2017.
37. G. Micheli, E. Soda, M. T. Vespucci, M. Gobbi, and A. Bertani, "Big data analytics: an aid to detection of non-technical losses in power utilities," Computational Management Science, vol. 16, pp. 329-343, 2019.
38. I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," SN Computer Science, vol. 2, no. 3, p. 160, 2021.
39. R. Muhamedyev, "Machine learning methods: An overview," Computer Modelling & New Technologies, vol. 19, no. 6, pp. 14-29, 2015.
40. R. Saravanan and P. Sujatha, "A state of art techniques on machine learning algorithms: a perspective of supervised learning approaches in data classification," in 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 945-949, IEEE, 2018.
41. Muhammad Usama, JunaidQadir, AunnRaza, HunainArif, Kok-Lim Alvin Yau, YehiaElkhatib, Amir Hussain, and Ala Al-Fuqaha, "Unsupervised machine learning for networking: Techniques, applications and research challenges," IEEE Access, vol. 7, pp. 65579-65615, 2019.

42. S. A. Fayaz, S. J. Sidiq, M. Zaman, and M. A. Butt, "Machine learning: An introduction to reinforcement learning," in Machine Learning and Data Science: Fundamentals and Applications, pp. 1-22, 2022.
43. X. Zhu and A. B. Goldberg, Introduction to Semi-Supervised Learning. Springer Nature, 2022.
44. A. Pazderin, F. Kamalov, P. Y. Gubin, M. Safaraliev, V. Samoylenko, N. Mukhlynin, I. Odinaev, and I. Zicmane, "Data-Driven Machine Learning Methods for Nontechnical Losses of Electrical Energy Detection: A State-of-the-Art Review," Energies, vol. 16, no. 21, p. 7460, 2023.
45. M. A. de Souza, J. L. R. Pereira, G. de O. Alves, B. C. de Oliveira, I. D. Melo, and P. A. N. Garcia, "Detection and identification of energy theft in advanced metering infrastructures," Electric Power Systems Research, vol. 182, p. 106258, 2020.
46. M. Z. Gunduz and R. Das, "Smart Grid Security: An Effective Hybrid CNN-Based Approach for Detecting Energy Theft Using Consumption Patterns," Sensors, vol. 24, no. 4, p. 1148, 2024.
47. P. Jumale, A. Khaire, H. Jadhawar, S. Awathare, and M. Mali, "Survey: Electricity Theft Detection Technique," International Journal of Computer Engineering and Information Technology, vol. 8, no. 2, pp. 30–35, Feb. 2016.
48. Y. Wang, Q. Chen, and C. Kang, "Electricity Theft Detection," in Smart Meter Data Analytics, Springer, Singapore, 2020, pp. 65-79.
49. S. Agarwal, A. Srivastava, R. Sodhi, and T. Soni, "Comparative Evaluation of Exploratory Data Analysis Techniques for Power Theft Detection in Residential Distribution Grids," in 2023 IEEE 3rd International Conference on Smart Technologies for Power, Energy and Control (STPEC), Bhubaneswar, India, 2023, pp. 1-6.
50. Y. Xue, Y. Shu, H. Yang, S. Liu, and Y. Xu, "Electric Theft Behavior Detection Method Based on Power Customer Data Analysis," in 2020 International Wireless Communications and Mobile Computing (IWCMC), Limassol, Cyprus, 2020, pp. 99-102, doi: 10.1109/IWCMC48107.2020.9148301.

51. A. Alromih, J. A. Clark, and P. Gope, "Electricity Theft Detection in the Presence of Prosumers Using a Cluster-based Multi-feature Detection Model," in 2021 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), Aachen, Germany, 2021, pp. 339-345, doi: 10.1109/SmartGridComm51999.2021.9632322.
52. H. Barzamini and M. Ghassemian, "Comparison analysis of electricity theft detection methods for advanced metering infrastructure in smart grid," Int. J. Electron. Secur. Digit. Forensics, vol. 11, pp. 265-280, 2019.
53. C. H. Park and T. Kim, "Energy theft detection in advanced metering infrastructure based on anomaly pattern detection," Energies, vol. 25, pp.1-10, 2020.
54. P. Jokar, N. Arianpoo, and V. C. M. Leung, "Electricity theft detection in AMI using customers’ consumption patterns," IEEE Transactions on Smart Grid, vol. 7, pp. 216-226, 2016.
55. A. Maamar and K. Benahmed, "Machine learning techniques for energy theft detection in AMI," in Proceedings of the 2018 International Conference on Software Engineering and Information Management (ICSIM '18), Association for Computing Machinery, New York, NY, USA, pp. 57-62, 2018. doi: 10.1145/3178461.3178484.
56. J. Yeckle and B. Tang, "Detection of electricity theft in customer consumption using outlier detection algorithms," in 2018 1st International Conference on Data Intelligence and Security (ICDIS), pp. 135-140, 2018.
57. K. Zheng, Q. Chen, Y. Wang, C. Kang, and Q. Xia, "A novel combined data-driven approach for electricity theft detection," IEEE Transactions on Industrial Informatics, vol. 15, pp. 1809-1819, 2019.
58. R. Razavi, A. Gharipour, M. Fleury, and I. J. Akpan, "A practical feature-engineering framework for electricity theft detection in smart grids," Applied Energy, 2019.
59. P. Ganguly, M. Nasipuri, and S. Dutta, "A novel approach for detecting and mitigating the energy theft issues in the smart metering infrastructure,"

Technol Econ Smart Grids Sustain Energy, vol. 3, no. 13, 2018, https://doi.org/10.1007/s40866-018-0053-x.

60. R. Jiang, R. Lu, Y. Wang, J. Luo, C. Shen, and X. Shen, "Energy-theft detection issues for advanced metering infrastructure in smart grid," Tsinghua Sci. Technol., vol. 19, no. 2, pp. 105–120, 2014. [Online]. Available: https://doi.org/10.1109/TST.2014.6787363.
61. S. A. Salinas and P. Li, "Privacy-preserving energy theft detection in microgrids: A state estimation approach," IEEE Trans. Power Syst., vol. 31, no. 2, pp. 883–894, 2016. [Online]. Available: https://doi.org/10.1109/TPWRS.2015.2406311.
62. S. C. Yip, K. S. Wong, W. P. Hew, M. T. Gan, R. C. W. Phan, and S. W. Tan, "Detection of energy theft and defective smart meters in smart grids using linear regression," Int. J. Electr. Power Energy Syst., vol. 91, pp. 230–240, 2017. [Online]. Available: https://doi.org/10.1016/j.ijepes.2017.04.005.
63. H. O. Henriques et al., "Development of adapted ammeter for fraud detection in low-voltage installations," Measurement, vol. 56, pp. 1–7, 2014.
64. A. A. Cardenas, S. Amin, G. Schwartz, R. Dong, and S. Sastry, "A game theory model for electricity theft detection and privacy-aware control in AMI systems," in 2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton), pp. 1830–1837, 2012.
65. Z. Zhou, J. Bai, M. Dong, K. Ota, and S. Zhou, "Game-theoretical energy management design for smart cyber-physical power systems," Cyber-Physical Syst., vol. 1, no. 1, pp. 24–45, 2015.
66. F. Shehzad, N. Javaid, S. Aslam, and M. U. Javaid, "Electricity theft detection using big data and genetic algorithm in electric power systems," Electr. Power Syst. Res., vol. 209, p. 107975, 2022.
67. M. Faheem, M. Umar, R. A. Butt, B. Raza, M. A. Ngadi, and V. C. Gungor, "Software defined communication framework for smart grid to meet energy demands in smart cities," in 2019 7th International Istanbul Smart Grids and Cities Congress and Fair (ICSG), IEEE, pp. 51–55, 2019.

68. S. Salinas, M. Li, and P. Li, "Privacy-Preserving Energy Theft Detection in Smart Grids: A P2P Computing Approach," IEEE J. Sel. Areas Commun., vol. 31, pp. 257-267, 2013.

69. C. Richardson, N. J. Race, and P. Smith, "A privacy preserving approach to energy theft detection in smart grids," in 2016 IEEE International Smart Cities Conference (ISC2), pp. 1-4, 2016.

70. P. Ganguly, M. Nasipuri, and S. Dutta, "A novel approach for detecting and mitigating the energy theft issues in the smart metering infrastructure," in Technology and Economics of Smart Grids and Sustainable Energy.

71. S. K. Singh, R. Bose, and A. Joshi, "Energy theft detection in advanced metering infrastructure," in 2018 IEEE 4th World Forum on Internet of Things (WF-IoT), pp. 529-534, 2018.

72. E. SomefunT., A. C.O.A, and A. Chiagoro, "Smart prepaid energy metering system to detect energy theft with facility for real-time monitoring," Int. J. Electr. Comput. Eng. (IJECE).

73. M. N. Hasan, R. N. Toma, A. Nahid, M. M. M. Islam, and J.-M. Kim, "Electricity Theft Detection in Smart Grid Systems: A CNN-LSTM Based Approach," Energies.

74. D. Yao, M. Wen, X. Liang, Z. Fu, K. Zhang, and B. Yang, "Energy theft detection with energy privacy preservation in the smart grid," IEEE Internet of Things.

75. R. Razavi, A. Gharipour, M. Fleury, and I. J. Akpan, "A practical feature-engineering framework for electricity theft detection in smart grids," Applied Energy, 2019.

76. J. Kim, Y. Hwang, Y. Sun, I. Sim, D. I. Kim, and X. Wang, "Detection for Non-Technical Loss by Smart Energy Theft with Intermediate Monitor Meter in Smart Grid," IEEE Access, 2019.

77. C. H. Park and T. Kim, "Energy Theft Detection in Advanced Metering Infrastructure Based on Anomaly Pattern Detection," Energies, 2020.

78. A. Fragkioudaki, P. Cruz-Romero, A. G. Exposito, J. Biscarri, M. J. Tellechea, and A. Arcos-Vargas, "Detection of Non-technical Losses in Smart

Distribution Networks: A Review," in Practical Applications of Agents and Multi-Agent Systems, 2016.

79. A. Maamar and K. Benahmed, "Machine learning Techniques for Energy Theft Detection in AMI," in Proceedings of the 2018 International Conference on Software Engineering and Information Management, 2018.
80. M. M. Badr, M. I. Ibrahem, H. A. Kholidy, M. M. Fouda, and M. Ismail, "Review of the Data-Driven Methods for Electricity Fraud Detection in Smart Metering Systems," Energies, 2023.
81. T. Ahmad, H. Chen, J. Wang, and Y. Guo, "Review of various modeling techniques for the detection of electricity theft in smart grid environment," Renewable & Sustainable Energy Reviews, vol. 82, pp. 2916-2933, 2018.
82. D. Odoom, "A Methodology in Utilizing Machine Learning Algorithm for Electricity Theft Detection in Ghana," SSRN Electronic Journal, 2020.
83. S. Poudel and U. R. Dhungana, "Artificial intelligence for energy fraud detection: a review," International Journal of Applied Power Engineering (IJAPE), 2022.
84. J. Breitenbach, J. Gross, M. Wengert, J. Anurathan, R. Bitsch, Z. Kosar, E. Tuelue, and R. Buettner, "A Systematic Literature Review of Deep Learning Approaches in Smart Meter Data Analytics," in 2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC), pp. 1337-1342, 2022.
85. M. Kaur, S. Chawla, and R. Dua, "Comparative Research on the Techniques of Electricity Fraud Detection Using Different Machine Learning Techniques," CGC International Journal of Contemporary Technology and Research, 2022.
86. M. A. Souza, J. L. Pereira, G. D. Alves, B. C. Oliveira, I. D. Melo, and P. A. Garcia, "Detection and identification of energy theft in advanced metering infrastructures," Electric Power Systems Research, vol. 182, p. 106258, 2020.
87. D. Syed, H. Abu-Rub, S. Refaat, and L. Xie, "Detection of Energy Theft in Smart Grids using Electricity Consumption Patterns," in 2020 IEEE International Conference on Big Data (Big Data).

88. S. Hussain et al., "A novel feature engineered-CatBoost-based supervised machine learning framework for electricity theft detection," Energy Reports, vol. 7, pp. 4425-4436, 2021.
89. J. Jeyaranjani and D. Devaraj, "Machine Learning Algorithm for Efficient Power Theft Detection using Smart Meter Data," International Journal of Engineering & Technology, vol. 7, no. 3.34, pp. 900-904, 2018. [Online]. Available: www.sciencepubco.com/index.php/IJET.
90. F. A. Bohani et al., "A Comprehensive Analysis of Supervised Learning Techniques for Electricity Theft Detection," J. Electr. Comput. Eng., vol. 2021, pp. 9136206:1-9136206:10, 2021.
91. P. S. Khot, S. Dhore, S. Thorat, P. Musmade, and V. Sargar, "Electricity Theft Detection in Power Consumption using Superiority of Machine Learning Algorithm," International Journal of Scientific Research in Engineering and Management, 2023.
92. P. Ghosh et al., "Electricity Theft Detection Employing Machine Learning Algorithms," in 2023 IEEE 8th International Conference for Convergence in Technology (I2CT), 2023, pp. 1-6.
93. S. K. Gunturi and D. Sarkar, "Ensemble machine learning models for the detection of energy theft," Electric Power Systems Research, vol. 106904, 2020.
94. S. Hussain, M. W. Mustafa, T. A. Jumani, S. K. Baloch, H. Alotaibi, I. Khan, and A. Khan, "A novel feature engineered-CatBoost-based supervised machine learning framework for electricity theft detection," Energy Reports, vol. 7, pp. 4425-4436, 2021.
95. P. Ghosh, T. T. B. Audry, S. Rahman, F. Bhuiyan, S. T. Rifat, M. N. K. Hredoy, T. Ghosh, and D. M. Farid, "Electricity Theft Detection Employing Machine Learning Algorithms," in 2023 IEEE 8th International Conference for Convergence in Technology, 2023, pp. 1-6.
96. S. K. Gunturi and D. Sarkar, "Ensemble machine learning models for the detection of energy theft," Electric Power Systems Research, vol. 106904, 2020.

97. Z. A. Khan, M. N. Adil, N. Javaid, M. N. Saqib, M. Shafiq, and J.-G. Choi, "Electricity Theft Detection Using Supervised Learning Techniques on Smart Meter Data," Sustainability, vol. 2020.

98. J. Jeyaranjani, "Machine Learning Algorithm for Efficient Power Theft Detection using Smart Meter Data," 2018.

99. M. Hashatsi, C. Maulu, and M. Shuma-Iwisi, "Detection of electricity theft in low voltage networks using analytics and machine learning," in 2020 International SAUPEC/RobMech/PRASA Conference, 2020, pp. 1-6.

100. Prof. S. T. Khot, S. Dhore, S. Thorat, P. Musmade, and V. Sargar, "Electricity Theft Detection in Power Consumption using Superiority of Machine Learning Algorithm," International Journal of Scientific Research in Engineering and Management, vol. 2023.

101. A. Alkhresheh, M. A. Al-Tarawneh, and M. Alnawayseh, "Evaluation of Online Machine Learning Algorithms for Electricity Theft Detection in Smart Grids," International Journal of Advanced Computer Science and Applications, 2022.

102. C. Lu and C. Tsai, "An Effective Adaptive Stacking Ensemble Algorithm for Electricity Theft Detection," in Proceedings of the 2021 ACM International Conference on Intelligent Computing and its Emerging Applications, 2021.

103. M. Tariq and H. V. Poor, "Electricity Theft Detection and Localization in Grid-Tied Microgrids," in IEEE Transactions on Smart Grid, vol. 9, pp. 1920-1929, 2018.

104. A. Maamar and K. Benahmed, "Machine learning Techniques for Energy Theft Detection in AMI," in Proceedings of the 2018 International Conference on Software Engineering and Information Management (ICSIM '18), New York, NY, USA, pp. 57-62, 2018. doi: 10.1145/3178461.3178484.

105. D. Syed, H. Abu-Rub, S. S. Refaat, and L. Xie, "Detection of Energy Theft in Smart Grids using Electricity Consumption Patterns," in 2020 IEEE International Conference on Big Data (Big Data), pp. 4059-4064, 2020.

106. S. K. Singh, R. Bose, and A. Joshi, "Energy theft detection in advanced metering infrastructure," in 2018 IEEE 4th World Forum on Internet of Things (WF-IoT), pp. 529-534, 2018.

107. R. Razavi, A. Gharipour, M. Fleury, and I. J. Akpan, "A practical feature-engineering framework for electricity theft detection in smart grids," Applied Energy, vol. 238, pp. 481-494, 2019. doi: 10.1016/j.apenergy.2019.01.076.

108. K. Zhou, C. Yang, and J. Shen, "Discovering residential electricity consumption patterns through smart-meter data mining: A case study from China," Utilities Policy, vol. 44, pp. 73-84, 2017.

109. R. Khalid and N. Javaid, “A survey on hyperparameters optimization algorithms of forecasting models in smart grid,” *Sustain. Cities Soc.*, vol. 61, p. 102275, 2020.

110. C. Genes, I. Esnaola, S. M. Perlaza, L. F. Ochoa, and D. Coca, “Recovering missing data via matrix completion in electricity distribution systems,” *IEEE Work. Signal Process. Adv. Wirel. Commun. SPAWC*, vol. 2016-Augus, 2016, doi: 10.1109/SPAWC.2016.7536744.

111. T. A. Alghamdi and N. Javaid, "A survey of preprocessing methods used for analysis of big data originated from smart grids," IEEE Access, vol. PP, pp. 1-1, 2022.

112. M. Q. Saeed and F. Alsharif, "Signal Piloted Processing of the Smart Meter Data for Effective Appliances Recognition," Journal of Electrical Engineering & Technology, vol. 15, pp. 2279-2285, 2020.

113. N. R. Prasad, S. Almanza-Garcia, and T. T. Lu, “Anomaly detection,” *Comput. Mater.Contin.*, vol. 14, no. 1, pp. 1–22, 2009, doi: 10.1145/1541880.1541882.

114. C. G. Cordero, E. Vasilomanolakis, A. Wainakh, M. Mühlhäuser, and S. Nadjm-Tehrani, “On generating network traffic datasets with synthetic attacks for intrusion detection,” *ACM Trans. Priv. Secur.*, vol. 24, no. 2, pp. 1–39, 2021.

115. P. Jokar, N. Arianpoo, and V. C. M. Leung, “Electricity theft detection in AMI using customers’ consumption patterns,” *IEEE Trans. Smart Grid*, vol. 7, no. 1, pp. 216–226, 2016, doi: 10.1109/TSG.2015.2425222.

116. K. Zhou, C. Yang, and J. Shen, "Discovering residential electricity consumption patterns through smart-meter data mining: A case study from China," Utilities Policy, vol. 44, pp. 73-84, 2017.

117. G. Grigoraș and F. Scarlatache, "Processing of smart meters data for peak load estimation of consumers," in 2015 9th International Symposium on Advanced Topics in Electrical Engineering (ATEE), pp. 864-867, 2015.

118. M. M. Selvam, R. Gnanadass, and N. P. Padhy, "Fuzzy based clustering of smart meter data using real power and THD patterns," Energy Procedia, vol. 117, pp. 401-408, 2017.

119. G. Shamim and M. Rihan, "Novel Technique for Feature Computation and Clustering of Smart Meter Data," in 2019 International Conference on Electrical, Electronics and Computer Engineering (UPCON), pp. 1-5, 2019.

120. Y. Reich and S. V. Barai, "Evaluating machine learning models for engineering problems," Artificial Intelligence in Engineering, vol. 13, no. 3, pp. 257-272, 1999.

121. J. Huang and C. X. Ling, "Constructing New and Better Evaluation Measures for Machine Learning," in Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Jan. 2007, pp. 859-864.

122. M. C. Belavagi and B. Muniyal, "Performance evaluation of supervised machine learning algorithms for intrusion detection," Procedia Computer Science, vol. 89, pp. 117-123, 2016.

123. C. Ferri, J. Hernández-Orallo, and R. Modroiu, "An experimental comparison of performance measures for classification," Pattern Recognition Letters, vol. 30, no. 1, pp. 27-38, 2009.

124. M. Steurer, R. J. Hill, and N. Pfeifer, "Metrics for evaluating the performance of machine learning based automated valuation models," Journal of Property Research, vol. 38, no. 2, pp. 99-129, 2021.

125. R. Mehra, N. Bhatt, F. Kazi, and N. Singh, "Analysis of PCA based compression and denoising of smart grid data under normal and fault

conditions," in 2013 IEEE International Conference on Electronics, Computing and Communication Technologies, pp. 1-6, 2013.

126. S. K. Singh, R. Bose, and A. Joshi, "PCA based electricity theft detection in advanced metering infrastructure," in 2017 7th International Conference on Power Systems (ICPS), pp. 441-445, 2017.

127. C. R. Turner, A. Fuggetta, L. Lavazza, and A. L. Wolf, "A conceptual basis for feature engineering," Journal of Systems and Software, vol. 49, no. 1, pp. 3-15, 1999.

128. W. Wang and A. K. Wong, "Autoregressive model-based gear fault diagnosis," J. Vib. Acoust., vol. 124, no. 2, pp. 172-179, 2002.

129. J. Cai, H. Cai, Y. Cai, L. Wu, and Y. Shen, "Short-term Forecasting of User Power Load in China Based on XGBoost," in 2020 12th IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC), pp. 1-5, 2020.

130. E. E. Rizqi and C. Safitri, "An Intelligent Calibration Testing of Electricity Meter using XGBoost for Manufacturing 4.0," in 2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE), pp. 183-188, 2023.

131. S. Hussain, M. W. Mustafa, T. A. Jumani, S. K. Baloch, H. Alotaibi, I. Khan, and A. Khan, "A novel feature engineered-CatBoost-based supervised machine learning framework for electricity theft detection," Energy Reports, vol. 7, pp. 4425-4436, 2021.

132. G. Biau, "Analysis of a random forests model," J. Mach. Learn. Res., vol. 13, no. 1, pp. 1063-1095, 2012.

133. S. Li, Y. Han, X. Yao, S. Yingchen, J. Wang, and Q. Zhao, "Electricity theft detection in power grids with deep learning and random forests," J. Electr. Comput. Eng., 2019, article ID 4136874.

134. L. Grant, H. Latchman, and K. Alli, "Detection of electricity theft in developing countries - A machine learning approach," Trends in Computer Science and Information Technology, vol. 8, no. 2, pp. 38-49, 2023.

135. T. G. Nick and K. M. Campbell, "Logistic regression," in Topics in Biostatistics, 2007, pp. 273-301.

136. J. Laaksonen and E. Oja, "Classification with learning k-nearest neighbors," in Proceedings of International Conference on Neural Networks (ICNN'96), vol. 3, pp. 1480-1483, IEEE, 1996.

137. Z. Zhang, "Introduction to machine learning: k-nearest neighbors," Ann. Transl. Med., vol. 4, no. 11, 2016.

138. D. T. Larose and C. D. Larose, "k-nearest neighbor algorithm," 2014, pp. 149-164.

139. W. S. Noble, "What is a support vector machine?," Nat. Biotechnol., vol. 24, no. 12, pp. 1565-1567, 2006.

140. K. P. Soman, R. Loganathan, and V. Ajay, Machine Learning with SVM and Other Kernel Methods, PHI Learning Pvt. Ltd., 2009.

141. H.T. Lin, "Introduction to Support Vector Machines," Learning Systems Group, California Institute of Technology, 2005.

142. R. E. Schapire, "Explaining AdaBoost," in Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik, Berlin, Germany: Springer Berlin Heidelberg, 2013, pp. 37-52.

143. R. Punmiya and S. Choe, “ToU pricing-based dynamic electricity theft detection in smart grid using gradient boosting classifier,” *Appl. Sci.*, vol. 11, no. 1, pp. 1–15, 2021, doi: 10.3390/app11010401.

144. A. I. Kawoosa, D. Prashar, M. Faheem, N. Jha, and A. A. Khan, “Using machine learning ensemble method for detection of energy theft in smart meters,” *IET Gener. Transm. Distrib.*, vol. 17, no. 21, pp. 4794–4809, 2023.

145. T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, and T. Zhou, "Xgboost: extreme gradient boosting," R package version 0.4-2, vol. 1, no. 4, pp. 1-4, 2015.

146. M. Feurer and F. Hutter, "Hyperparameter optimization," in Automated Machine Learning: Methods, Systems, Challenges, 2019, pp. 3-33.

147. R. Andonie, "Hyperparameter optimization in learning systems," J. Membrane Comput., vol. 1, no. 4, pp. 279-291, 2019.

148. M. Feurer and F. Hutter, "Hyperparameter optimization," in Automated Machine Learning: Methods, Systems, Challenges, 2019, pp. 3-33.

149. U. Michelucci, "Hyperparameter Tuning," in Applied Deep Learning: A Case-Based Approach to Understanding Deep Neural Networks, 2018, pp. 271-322.

150. M. R. Hossain and D. Timmer, "Machine learning model optimization with hyper parameter tuning approach," Global Journal of Computer Science and Technology, vol. 21, no. D2, pp. 7-13, 2021.