Comprehensive Analysis of Cost-Effective Data-Driven Approaches for Electricity Theft Detection

A PROJECT REPORT

For the partial fulfilment for the award of the major degree in

Electrical and Electronics Engineering

Submitted By:

Ayush Kumar (211230013)

under the guidance of

Dr. Manoj Kumawat

Department of Electrical Engineering

NIT Delhi

Delhi-110036

June - July 2024

DECLARATION

I hereby declare that the project report entitled " **Comprehensive Analysis of Cost-Effective Data-Driven Approaches for Electricity Theft Detection** " submitted to the National Institute of Technology Delhi during the academic year 2023-24 in partial fulfillment of the requirements for the award of Degree of Bachelor of Technology in Electrical and Electronics Engineering is a record of bonafide project work carried out under the guidance and supervision of Dr. Manoj Kumawat. I further declare that the work reported in this project has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other University.

Ayush Kumar (211230013)

Place: NIT Delhi
Date: October 1, 2024

Department of Electrical Engineering
National Institute of Technology Delhi
Delhi – 110036

CERTIFICATE

This is to certify that the report entitled **"Comprehensive Analysis of Cost-Effective Data-Driven Approaches for Electricity Theft Detection"** submitted by **Ayush Kumar (Roll no. 211230013)**, to the National Institute of Technology Delhi in Electrical and Electronics Engineering is a bonafide record of the project work carried out by him under my guidance and supervision. This report in any form has not been submitted to any other University or Institute for any purpose.

Dr. Manoj Kumawat
Supervisor
Assistant Professor
Dept. of Electrical Engineering

CONTENTS

ACKNOWLEDGEMENT

Many noble hearts contributed immense inspiration and support for the successful completion of the project. We are unable to express our gratitude in words to such individuals.

I would like to express my deep regard to Ḍr. (Prof.) Ajay K Sharma, Director, NIT Delhi, for providing facilities throughout the works of my project.

I would like to take this opportunity to express my profound gratitude to HOD, Head of the Department, Dr. Pankaj Mukhija, Department of Electrical Engineering, NIT Delhi, for providing permission and availing all required facilities for undertaking the project in a systematic way. I am extremely grateful to Ṣupervisor, Dr. Manoj Kumawat Assistant Professor, Department of EE, NIT Delhi, who guided us with his kind, ordinal and valuable suggestions. We would also like to thank all the teaching and non-teaching staff of the Department of EE, NIT Delhi, for the sincere directions imparted and the cooperation in connection with the project.

I will be failing in our duty if we do not acknowledge with grateful thanks to the authors of the references and other literatures referred in this project.

Last, but not the least, I take pleasant privilege in expressing my heartful thanks to our friends who were of precious help in completing this project.

Ayush Kumar

ABSTRACT

Electricity theft is a major challenge for utility companies and power grids, leading to significant financial losses, grid overloading, and safety hazards for workers. Overloading can result in voltage instability, potentially causing grid failures and large-scale blackouts, which disrupt everyday life and business operations. Unsafe wiring from theft also poses risks of fires, shocks, and equipment damage.

To address these issues, various anti-theft strategies have been employed, including manual inspections, hardware-based methods like tamper-proof meters, and data-driven approaches. While hardware methods offer direct detection, they are often costly and may not catch sophisticated theft. Data-driven techniques, on the other hand, rely on patterns in energy consumption to identify anomalies. Traditional models like SVM, Random Forests (RF), LSTM, and Decision Trees have been used but exhibit limitations in metrics such as precision, sensitivity, recall, and specificity.

This paper focuses on overcoming these limitations using advanced machine learning and deep learning methods, including Convolutional Neural Networks (CNN), Wide and Deep CNN architectures, and Time Series Forecasting models. CNNs excel at identifying complex patterns in energy consumption data, while the Wide and Deep architecture combines the strengths of both linear models and deep neural networks, enhancing accuracy. Time Series Forecasting allows for the prediction of future consumption patterns, identifying deviations that may indicate theft.

By leveraging these advanced techniques, this paper demonstrates improved performance over traditional methods in detecting electricity theft. These approaches offer greater scalability, accuracy, and efficiency, reducing false positives and negatives. The findings suggest that deep learning models hold the key to developing more robust and proactive theft detection systems, safeguarding financial resources, improving grid reliability, and enhancing worker safety.

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 OVERVIEW

Electricity theft poses a significant challenge for utility companies and power grid management, leading to considerable financial losses and the potential for grid overloading. Such overloading not only threatens the stability of power supply but also raises safety concerns for maintenance personnel and infrastructure. In response to this pressing issue, utilities have implemented various detection strategies, which can broadly be categorized into manual involvement, hardware detection systems, and data-driven methodologies[1].

Among these strategies, data-driven methods have emerged as a promising approach, offering enhanced accuracy and efficiency in identifying instances of electricity theft. These methods can be further divided into fundamental and advanced categories. While fundamental techniques, such as Support Vector Machines (SVM), Random Forests (RF), Long Short-Term Memory (LSTM) networks, and Decision Trees, have laid the groundwork for electricity theft detection, they often fall short in terms of key performance metrics like precision, sensitivity, recall, and specificity.

To address the limitations inherent in these traditional approaches, this paper explores the application of advanced data-driven techniques, including Deep Learning, Convolutional Neural Networks (CNN), Wide and Deep CNN, and Time Series Forecasting. By leveraging these sophisticated models, we aim to improve the detection capabilities of electricity theft, thereby enhancing the overall performance of utility operations and contributing to a more reliable power grid.

1.1.1 Challenges in the Power Sector

The power sector faces a myriad of challenges that significantly impact its efficiency, reliability, and sustainability[2]. Understanding these challenges is crucial for developing effective strategies to mitigate risks and enhance service delivery. The key challenges include:

(i) Electricity Theft

Electricity theft remains one of the most pressing issues in the power sector. Unauthorized consumption of electricity leads to significant revenue losses for utilities, esti-

mated to run into billions annually. The difficulty in accurately detecting and addressing theft undermines the financial viability of power providers and affects their ability to invest in infrastructure improvements.

(ii) Aging Infrastructure

Many power grids are built on aging infrastructure that is not equipped to handle the increasing demand for electricity. Outdated equipment and technologies can lead to frequent outages, inefficiencies, and higher maintenance costs. Upgrading this infrastructure requires substantial investment and strategic planning, which can be challenging for utilities facing budget constraints.

(iii) Grid Overloading

With the growing population and industrial demands, power grids are often subject to overloading, particularly during peak usage periods. Overloading can lead to equipment failures, prolonged outages, and safety hazards for both the grid operators and consumers. Managing demand through effective load forecasting and distribution is essential but remains a complex challenge.

(iv) Integration of Renewable Energy

The shift towards renewable energy sources, such as solar and wind, presents both opportunities and challenges. While renewable energy can enhance sustainability and reduce carbon emissions, it also introduces variability and unpredictability in power generation. Utilities must develop advanced grid management strategies to integrate these resources effectively while maintaining reliability.

(v) Cybersecurity Threats

The increasing digitization of power systems exposes them to cybersecurity threats. Malicious attacks on critical infrastructure can disrupt service delivery, compromise sensitive data, and pose safety risks. Ensuring robust cybersecurity measures is vital to protect the integrity and reliability of power systems.

(vi) Regulatory and Compliance Issues

The power sector operates under a complex framework of regulations and compliance requirements, which can vary significantly across regions. Navigating these regulations can be challenging for utilities, particularly in terms of meeting environmental standards, reporting requirements, and maintaining consumer protection laws.

1.1.2 The Global Context of Electricity Theft

Electricity theft poses significant challenges not only in India but across the globe. It leads to substantial financial losses for utility companies and contributes to customer dissatisfaction, manifesting in frequent power outages, voltage fluctuations, and increased tariffs[3]. These issues are often exacerbated by the activities of unauthorized users. Various methods of power theft are employed, including bypassing meters and utilizing illegal techniques to access electricity. Common tactics include power line tapping, electricity siphoning, and meter tampering. Addressing these theft methods is crucial for maintaining the integrity and reliability of power supply systems worldwide[4].

Table 1.1: Profits & Power vanished in Electricity illegal usage

Nation	Power Robbed (%)	Returns Shortfalls in US Dollars
America	3.5	$10 B
India	30	$16 B
SA	33	$1.5 B
Netherlands	23	$1.2 B
Brazil	28	$3.7 B
Bangladesh	14	$51 M
Malaysia	20	$229 M
Turkey	15	$1 B
Jamaica	18	$46 M

1.1.3 Electricity Theft in India

Electricity theft is a pervasive issue in India, significantly impacting the financial health of utility companies and the overall power sector. Despite the government's considerable investment in improving infrastructure and expanding electricity access, the problem persists, primarily due to socio-economic factors and inadequate regulatory enforcement[5].

In India, electricity theft accounts for approximately 30% of total power losses, translating to an annual revenue loss of around 16 billion USD. This alarming figure highlights the urgent need for effective measures to combat this issue, especially in a developing economy where reliable electricity is crucial for growth.

Several factors contribute to the prevalence of electricity theft in India. First, many consumers lack awareness of the consequences of theft, often viewing it as a victimless crime. Second, the high levels of poverty and unemployment in certain regions drive individuals to resort to illegal methods to access electricity. Third, the existing power infrastructure is often outdated and vulnerable, making it easier for unauthorized users to exploit loopholes.

Common methods of electricity theft in India include meter tampering, direct connec-

tions to power lines, and bypassing meters[6]. Unauthorized users often employ tactics such as power line tapping and siphoning, which not only result in financial losses but also pose safety risks to both consumers and utility workers.

To address the challenge of electricity theft, the Indian government has initiated various strategies, including the implementation of smart metering technologies and increased surveillance of power distribution networks. Additionally, public awareness campaigns aim to educate consumers about the importance of reporting theft and the potential consequences of engaging in such practices.

Despite these efforts, tackling electricity theft remains a complex challenge that requires a multi-faceted approach involving technological innovation, regulatory reform, and community engagement.

Table 1.2: Percentage Electricity Theft in Different States of India

State/UT	**Electricity Theft (%)**
Andhra Pradesh	26%
Bihar	40%
Goa	18%
Gujarat	12%
Jammu & Kashmir	37%
Karnataka	14%
Kerala	16%
Maharashtra	15%
Nagaland	34%
Odisha	27%
Punjab	21%
Rajasthan	29%
Telangana	25%
Uttar Pradesh	38%
West Bengal	32%

1.2 GOVERNMENT INITIATIVES AND STRATEGIES

The Indian government has recognized electricity theft as a critical issue undermining the financial stability of the power sector. In response, several initiatives and strategies have been implemented at both the central and state levels to combat this persistent challenge.

One of the primary strategies has been the widespread installation of smart meters across urban and rural areas. Smart meters provide real-time monitoring of electricity consumption, making it easier for utilities to detect discrepancies and unauthorized usage. This technology enables remote disconnection of services in cases of theft, thus

reducing manual intervention and losses.

Additionally, the government has launched initiatives like the Pradhan Mantri Sahaj Bijli Har Ghar Yojana (Saubhagya Scheme) to ensure universal access to electricity. By improving infrastructure in rural areas, the government aims to minimize technical losses and create a more reliable supply, which can reduce the incentive for theft.

To foster community participation, various awareness campaigns have been launched to educate consumers about the implications of electricity theft. These campaigns aim to encourage reporting of theft and promote responsible usage of electricity.

Table 1.3: Most common methods of illegal usage of electricity

Apparatus	**Theft's Mode**
Meters	Manipulating meters
Wires/Cables	Tapping into the overhead cables
Transformers	Unauthorized connection to the transformer
Billing Irregularities	Incorrect meter readings by the meter readers
Unpaid Bills	Non-payment of bills by institutions and individuals

The government is also collaborating with technology firms and research institutions to develop innovative solutions for theft detection. Advanced data analytics, machine learning, and AI-based systems are being explored to enhance monitoring and predictive capabilities in the power distribution network.

Moreover, investments have been made to strengthen the operational capacity of utility companies, including training programs for staff and upgrades to existing infrastructure. By enhancing the capabilities of utility providers, the government aims to improve their efficiency in monitoring and preventing electricity theft.

The legal framework has also been strengthened to address electricity theft. Laws have been enacted to ensure swift legal action against offenders, thus serving as a deterrent to potential theft.

Finally, the government is encouraging public-private partnerships to leverage private sector expertise in combating electricity theft. These partnerships can facilitate the sharing of resources and technology, leading to more effective theft detection and prevention strategies.

Through these initiatives and strategies, the Indian government aims to create a more sustainable and reliable power sector, reduce electricity theft, and enhance the financial health of utility companies.

CHAPTER 2
LITERATURE REVIEW

2.1 INTRODUCTION TO ELECTRICITY THEFT DETECTION

Electricity theft is a widespread and challenging issue that poses significant concerns for both utility companies and governments, particularly in developing countries. The illegal consumption of electricity not only results in substantial revenue losses but also affects the country's Gross Domestic Product (GDP), leading to a range of consequences such as reduced power quality and increased strain on the electrical grid. This issue impacts regular consumers as well, who often suffer from voltage drops, service interruptions, and higher utility costs due to theft.

As electricity theft continues to grow, traditional methods such as manual inspections and hardware-based solutions are proving to be inefficient and costly. Manual inspections are labor-intensive and difficult to scale, while tamper-proof meters and other hardware approaches are susceptible to advanced bypass techniques. In response, the focus has shifted towards data-driven approaches that harness the power of big data, machine learning, and artificial intelligence to detect and mitigate electricity theft in real-time.

In this section, we will explore how advanced data-driven techniques can be implemented to address electricity theft. By leveraging the abundant data collected from smart meters and intelligent sensors, machine learning models can classify and cluster users based on their electricity usage patterns[7]. These techniques range from fundamental machine learning methods, such as decision trees and support vector machines (SVM), to more advanced approaches using deep learning and artificial intelligence, which have proven to be more accurate and cost-effective for detecting theft. We will also delve into the gaps in the current research and highlight the most efficient and economically viable methods for electricity theft detection.

2.2 EXISTING METHODS AND TECHNOLOGIES

2.2.1 Manual and Hardware-Based Approaches

Historically, the detection of electricity theft relied heavily on manual inspections, regular meter checks, and hardware solutions such as tamper-resistant meters. These approaches, while straightforward, come with high costs and scalability issues, especially in regions with a large number of consumers[8]. As electricity theft methods become

more sophisticated, even advanced metering infrastructure (AMI) and smart meters face challenges in providing secure, tamper-proof solutions.

2.2.2 Data-Driven Approaches

With the growing availability of data from smart meters, utilities are increasingly turning to data-driven methods for theft detection[9]. These approaches utilize machine learning (ML) algorithms to analyze consumption data and identify anomalies that may indicate theft. The deployment of smart meters and intelligent sensors enables the collection of large volumes of data, which can be processed to extract valuable features for classification and clustering of electricity users[10].

Machine learning methods include:

- **Support Vector Machine (SVM)**: SVMs are commonly applied to classify consumption patterns, separating legitimate users from potential electricity thieves. Despite their effectiveness in linear data separations, SVMs often struggle with complex, non-linear relationships in the data[11].

- **Random Forest (RF)**: As an ensemble learning method, RF constructs multiple decision trees and combines their outputs to improve classification accuracy. However, RF models may overfit the training data, and their interpretability is often limited[12].

- **Artificial Neural Networks (ANN)**: ANNs are widely used for non-linear classification tasks. By mimicking the neural networks of the human brain, ANNs can learn complex patterns in electricity consumption data, though they require a large amount of training data to function effectively.

- **Principal Component Analysis (PCA)**: PCA is used for dimensionality reduction and feature extraction in large datasets, helping to simplify the data without losing valuable information.

- **Decision Trees (DT)**: Decision Trees are simple to implement and interpret but can become biased toward overfitting, especially with imbalanced datasets.

While these machine learning methods have shown promise, they are not without limitations. For example, traditional models often suffer from low precision and recall in the context of detecting electricity theft, particularly in cases where the theft methods are sophisticated or when the data is noisy[13].

2.3 ADVANCED ML & DL TECHNIQUES

Recent advancements in deep learning (DL) have revolutionized the field of electricity theft detection, offering superior performance compared to conventional machine learning methods[14]. Deep learning models, such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, have shown higher accuracy in detecting theft due to their ability to automatically learn and extract complex features from large datasets.

2.3.1 Deep Learning Methods in Electricity Theft Detection

Several DL techniques have been proposed to address the limitations of traditional ML approaches:

- **CNN-Based Learning Schemes**: A comprehensive learning scheme combining both wide and deep CNN modules has been developed to detect illegal electricity usage more accurately by analyzing long-standing smart meter data[15].
- **Privacy-Preserving Methods**: Privacy-preserving theft detection methods that incorporate CNNs have been implemented to ensure user data privacy while improving detection accuracy[16].
- **Hybrid Methods Using LSTM and CNN**: A hybrid method utilizing LSTM and CNN models has been designed to process smart meter data and auxiliary operation data separately, improving theft detection accuracy and efficiency[17].
- **Class Imbalance Learning with CNN**: CNN models combined with class imbalance learning techniques have been developed to address the issue of the lower proportion of electricity thieves in comparison to normal users, thereby reducing false positives[18].

These deep learning approaches significantly outperform traditional methods, particularly in their ability to extract relevant features from large and complex datasets. However, deep learning models require significant amounts of labeled data for training, which can be a challenge in the context of electricity theft detection.

2.4 CHALLENGES IN ELECTRICITY THEFT DETECTION

One of the major challenges in electricity theft detection is the need for large, high-quality labeled datasets. Supervised learning techniques rely on accurate labels to classify data correctly, but the process of labeling data is often expensive and time-consuming. To address this issue, research has explored the use of semi-supervised

and unsupervised learning techniques, which require fewer labeled instances to build accurate models.

2.4.1 Semi-Supervised and Unsupervised Learning

While unsupervised learning methods, such as clustering, can perform electricity theft detection using unlabelled data, they often face issues related to data mislabeling, which can degrade model performance. Semi-supervised learning (SSL), on the other hand, only requires a small number of labeled examples, but it lacks the ability to automatically correct mislabeling during iterative learning processes. As a result, supervised learning methods, although reliant on labeled datasets, continue to offer the highest accuracy and reliability in detecting electricity theft.

2.4.2 Cost-Effective Data Labeling

There has been limited research into reducing the costs associated with data labeling for electricity theft detection. Recent studies have proposed strategies such as Active Learning (AL) schemes to minimize the amount of labeled data required, thus lowering the overall cost of model implementation without sacrificing performance.

2.5 RESEARCH GAPS AND FUTURE DIRECTIONS

Despite the advancements in machine learning and deep learning techniques, several research gaps remain in the field of electricity theft detection:

- **Cost-Effective Data Labeling**: There is a need for more research into methods that can reduce the cost of data labeling without compromising detection accuracy.
- **Privacy Concerns**: While some privacy-preserving methods have been developed, further research is needed to ensure data privacy in electricity theft detection.
- **Hybrid Approaches**: Combining electricity data with other utility data, such as water usage, has shown promise in improving detection accuracy. However, this area remains under-explored and requires further investigation.

In conclusion, while data-driven approaches and deep learning techniques offer significant potential in combating electricity theft, future research should focus on addressing the challenges of data labeling costs, privacy, and the development of more efficient and accurate hybrid models.

CHAPTER 3
METHODOLOGY

This chapter outlines the systematic approach followed in this study to detect electricity theft using advanced machine learning and deep active learning (DAL) techniques. The process is divided into three main sections: data collection, model selection, and evaluation metrics. Each section describes the step-by-step methods employed to ensure accurate, efficient, and cost-effective electricity theft detection.

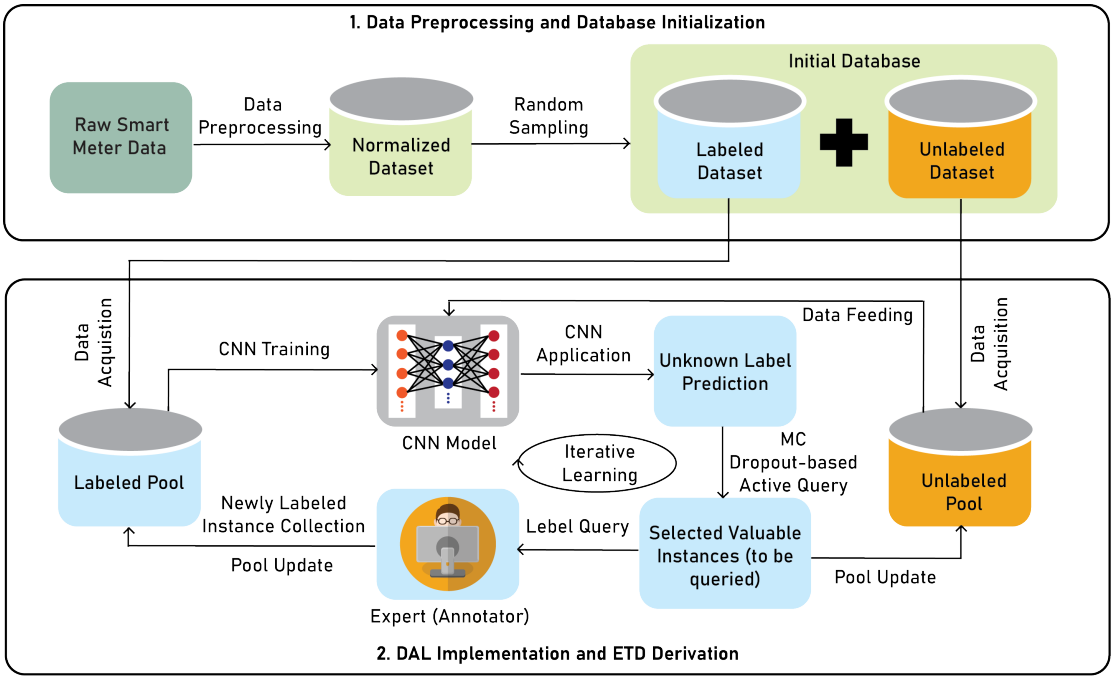


Figure 3.1: The suggested DAL-enabled ETD methodology's general framework

3.1 DATA COLLECTION

The success of any machine learning model largely depends on the quality and quantity of the data used for training. In this study, a comprehensive dataset of electricity consumption was collected from smart meters installed across various regions. These datasets were supplemented with auxiliary data, including user-specific characteristics and environmental factors, to enhance the detection capability of our models.

3.1.1 Data Preprocessing

Data preprocessing is an essential step to ensure clean, usable data for model training. The raw datasets often contain anomalies such as missing values, outliers, and inconsistencies. The preprocessing pipeline consisted of the following stages:

- **Missing Value Updates:** Missing values, often marked as NaN, were treated using an interpolation technique to ensure the dataset's completeness and continuity. The formula used for interpolation is:

$$f(x) = \begin{cases} \frac{x_{i-1}+x_{i+1}}{2}, & \text{if } x_i \in NaN, x_{i-1}, x_{i+1} \neq NaN \\ 0, & \text{if } x_i \in NaN, x_{i-1} \text{ or } x_{i+1} \in NaN \\ x_i, & \text{if } x_i \neq NaN \end{cases}$$

  This ensures that the values used for training are complete and within acceptable ranges.

- **Outlier Filtering:** Outliers can negatively affect model accuracy. The three-sigma rule was employed to identify and replace extreme values with more representative estimates:

$$f(x) = \begin{cases} avg(x) + 2 \cdot std(x), & \text{if } x_i > avg(x) + 2 \cdot std(x) \\ x_i, & \text{otherwise} \end{cases}$$

- **Data Normalization:** After handling missing values and outliers, the data was normalized using Min-Max scaling. This process ensures that the features are on a comparable scale, reducing bias in the model's predictions:

$$f(x) = \frac{x_i - min(x)}{max(x) - min(x)}$$

3.2 MODEL SELECTION

The model selection process is a critical aspect of this study, as the aim was to balance performance and cost-efficiency. The deep active learning (DAL) methodology was chosen due to its ability to reduce manual labeling efforts while maintaining model performance. The proposed DAL-enabled electricity theft detection (ETD) framework was divided into two main stages:

3.2.1 Data Initialization and Pooling

The collected data was split into labeled and unlabeled pools, with the former used to train an initial convolutional neural network (CNN) model. The initial model provided rough labels for the majority of the dataset, significantly reducing the need for manual labeling.

3.2.2 Active Learning Loop

After the initial training, valuable data points from the unlabeled pool were selected using the Monte Carlo failure method. These instances were sent to human annotators, whose feedback was incorporated into the model for further refinement. This iterative process continued until the ETD model reached optimal performance, reducing costs while improving detection accuracy.

3.3 EVALUATION METRICS

To assess the model's performance, various evaluation metrics were utilized, each offering insights into different aspects of the model's accuracy and reliability.

- **Accuracy:** This measures the percentage of correctly classified instances out of the total instances.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision:** Precision assesses the proportion of true positives out of all instances classified as positive.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall:** This metric, also known as sensitivity, calculates the ratio of true positives to the actual number of positives.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-Score:** The F1-score provides a balance between precision and recall, serving as a more comprehensive evaluation metric.

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

These metrics offered a thorough and detailed evaluation of the model's performance, enabling a balanced approach to managing the trade-offs between data labeling costs and detection efficiency. By providing a comprehensive insight into the model's strengths and limitations, they ensured that both cost-effectiveness and the accuracy of electricity theft detection were optimized.

CHAPTER 4
RESULTS AND DISCUSSION

4.1 OVERVIEW OF MODEL PERFORMANCE

The Classification Learner application was utilized to systematically assess a range of machine learning models, including Decision Trees, Binary GLM Logistic Regression, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Neural Networks. Each model's performance was rigorously analyzed using confusion matrices and scatter plot visualizations, which provided a comprehensive evaluation of classification accuracy and overall effectiveness. Among these models, the Decision Tree emerged as the standout performer, achieving exceptional accuracy and robustness in detecting electricity theft. In contrast, other models, such as Binary GLM Logistic Regression and KNN, exhibited notable limitations, struggling to deliver comparable results. This highlights the Decision Tree's superior capability in accurately identifying instances of electricity theft, positioning it as a highly reliable choice for real-world applications.

4.2 DECISION TREE PERFORMANCE

The Decision Tree model demonstrated outstanding performance, with an accuracy of 99.96% on the validation dataset and 99.92% on the test dataset. Sensitivity was 99.85%, and specificity reached a perfect 100.00%, indicating the model's exceptional ability to detect instances of electricity theft while flawlessly identifying non-theft cases. This balance of high sensitivity and specificity makes the Decision Tree model highly reliable for real-world deployment, minimizing both false positives and false negatives.

Table 4.1: Performance Metrics for Electricity Theft Detection Models

Model Type	**Sensitivity**	**Specificity**	**Precision**	**Recall**
Decision Tree	99.85%	100.00%	100.00%	99.85%
Linear GLM LR	53.03%	53.61%	27.31%	53.03%
Fine KNN	28.32%	98.34%	84.85%	28.32%
Linear SVM	16.76%	99.00%	84.67%	16.76%
Quadratic SVM	39.74%	97.01%	81.36%	39.74%
Optimizable SVM	99.85%	99.95%	99.85%	99.85%
Narrow NN	54.48%	90.98%	66.49%	54.48%
Medium NN	51.01%	92.74%	69.76%	51.01%
Bilayered NN	54.19%	90.98%	66.37%	54.19%
Trilayered NN	53.18%	91.55%	67.40%	53.18%

4.3 PERFORMANCE OF OTHER MODELS

The performance of models such as Binary GLM Logistic Regression and Fine KNN significantly lagged behind that of the Decision Tree. The Linear GLM Logistic Regression model recorded a sensitivity of only 53.03% and a specificity of 53.61%, rendering it ineffective for electricity theft detection. Similarly, while Fine KNN and Linear SVM models exhibited high precision, their low sensitivity resulted in a considerable number of missed theft cases, making them suboptimal for practical deployment. This highlights the critical importance of sensitivity in detection models, as it directly impacts their ability to identify instances of electricity theft effectively. As such, models must strike a balance between precision and sensitivity to ensure reliable real-world applications.

4.4 COMPARATIVE PERFORMANCE OF NEURAL NETWORKS

Neural Networks, configured with various architectures, demonstrated a balanced performance across the board. The Narrow NN, Medium NN, Bilayered NN, and Trilayered NN models achieved moderate accuracy levels, as illustrated in Table 4.1. While these models exhibited reasonable effectiveness, they fell short of the exceptional sensitivity and specificity attained by the Decision Tree model. For instance, the Narrow NN achieved an accuracy of 83.24%, but its sensitivity was comparatively lower at 54.48%. This deficiency in sensitivity indicates that the Narrow NN struggles to accurately identify instances of electricity theft, highlighting the importance of selecting models that not only provide balanced performance but also excel in detecting critical cases with high accuracy.

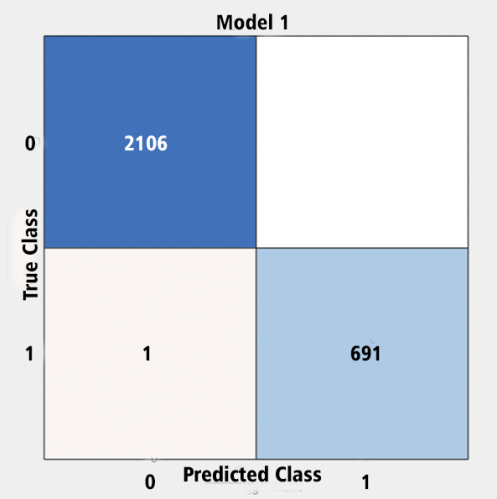
Table 4.2: Accuracy of Models

Model Type	**Accuracy (Validation)**	**Accuracy (Test)**
Decision Tree	99.96%	99.92%
Binary GLM LR	53.47%	53.34%
Linear SVM	78.66%	78.71%
Quadratic SVM	82.84%	82.64%
Optimizable SVM	99.93%	99.92%
Fine KNN	81.02%	80.30%
Narrow NN	83.24%	83.56%
Medium NN	82.42%	83.14%
Trilayered NN	81.95%	83.14%

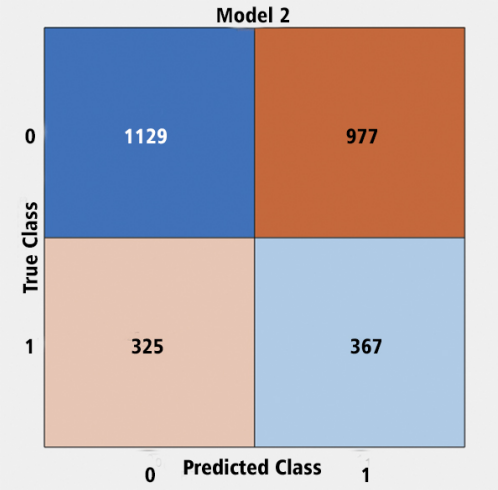
4.5 CONFUSION MATRIX AND SCATTER PLOT ANALYSIS

(i) Confusion Matrix

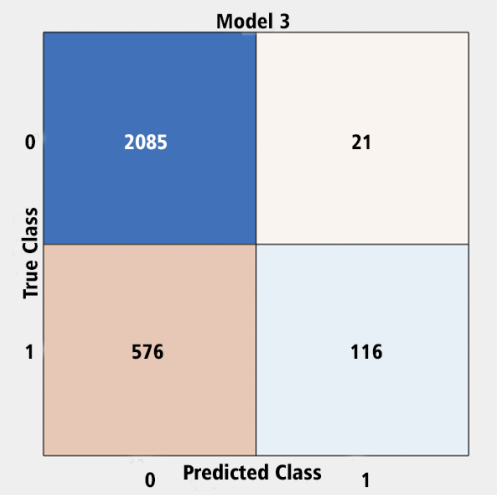
The Confusion Matrix analysis demonstrates that the Fine Decision Tree model excels, with exceptionally high True Positives and True Negatives and minimal classification errors, establishing it as a top performer. In stark contrast, the Linear GLM Logistic Regression and Linear SVM models exhibit considerably higher rates of False Positives and False Negatives, undermining their effectiveness and reliability. Neural Networks, while offering a balanced performance, vary depending on their specific configuration, providing a moderate level of accuracy.



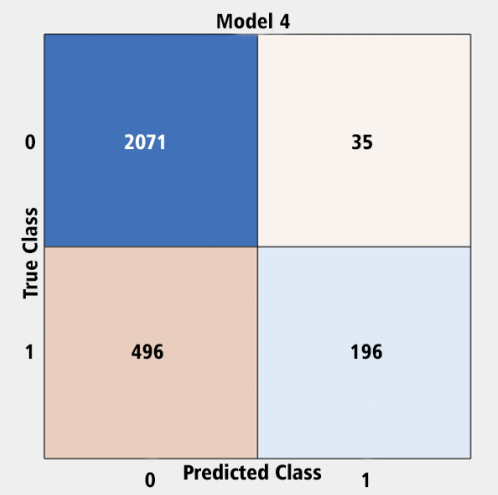
(a) Fine Tree Algorithm



(b) Binary GLM LR Algorithm



(c) Linear SVM Algorithm

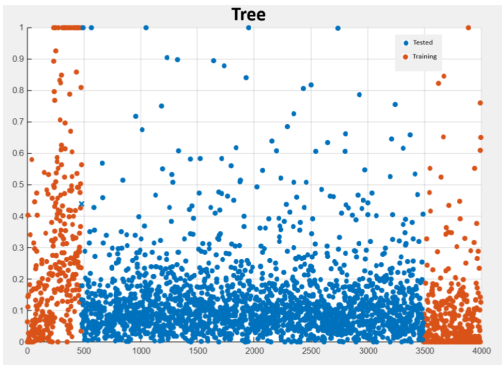


(d) Fine KNN Algorithm

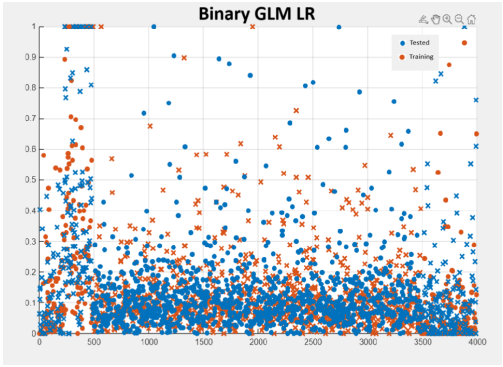
Figure 4.1: Confusion Matrices for Various Algorithms

(ii) Scatter Plot Analysis

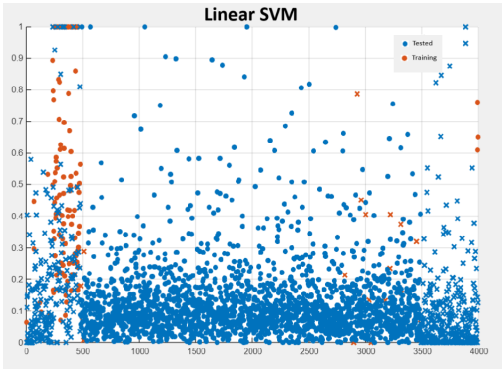
The scatter plot analysis further supports these findings, highlighting that the Decision Tree model achieves robust class separation with minimal overlap, underscoring its superior accuracy and efficiency. Conversely, the Linear GLM LR, Fine KNN, and Linear SVM models show significant class overlap, indicating less effective performance. Neural Networks present moderate class separation with varying degrees of overlap, which aligns with their moderate classification efficiency. This nuanced view of the scatter plots reinforces the varying performance levels of each model and their effectiveness in different scenarios.



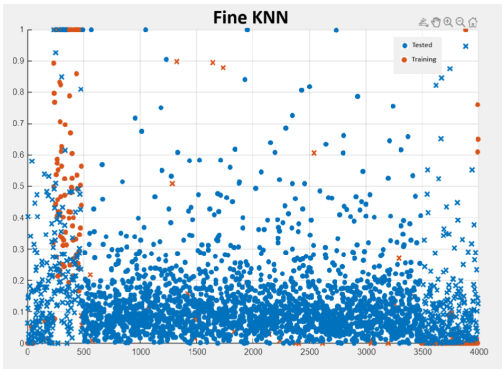
(a) Fine Tree Algorithm



(b) Binary GLM LR Algorithm



(c) Linear SVM Algorithm



(d) Fine KNN Algorithm

Figure 4.2: Scatter Plots for Various Algorithms

In conclusion, the evaluation of performance metrics reveals that the Decision Tree model excels in electricity theft detection, achieving a sensitivity of 99.85% and perfect specificity of 100.00%. In contrast, the Linear GLM Logistic Regression model showed low sensitivity (53.03%) and specificity (53.61%), making it ineffective for practical use. While Fine KNN and Linear SVM exhibited high precision, their insufficient sensitivity resulted in missed theft cases. These results highlight the necessity of choosing models that balance precision and sensitivity to enhance detection effectiveness.

CHAPTER 5
CONCLUSION

In the realm of electricity theft detection, existing research often underestimates the profound costs associated with data labeling—a critical process that significantly influences the quality and effectiveness of predictive models. While numerous studies delve into the sophistication of various detection methodologies, they frequently overlook the foundational necessity of high-quality, accurately labeled data. This oversight can result in models that, despite their advanced algorithms, fail to deliver reliable and actionable insights.

The labor-intensive nature of manual data labeling is a substantial challenge in this domain. This process not only consumes valuable time and human resources but also incurs significant financial costs, creating a bottleneck that hampers the scalability of detection initiatives. The reliance on manual efforts leads to inconsistencies and potential biases in the data, which ultimately undermine the performance of predictive models. As electricity theft continues to evolve in sophistication, it is crucial that the datasets used to train detection algorithms are not only abundant but also meticulously labeled to reflect the complexities of real-world scenarios.

Furthermore, addressing the challenges associated with data labeling will not only strengthen the reliability of electricity theft detection systems but also pave the way for their practical application on a larger scale. The implications of improved data labeling extend beyond mere model accuracy; they enhance the overall ecosystem of electricity theft prevention by enabling utility companies to deploy effective strategies tailored to specific contexts and challenges.

By tackling the inefficiencies in the data labeling process, we can enhance the scalability, accuracy, and overall performance of electricity theft detection systems. This advancement is crucial for making electricity theft detection solutions more effective and accessible, thereby facilitating widespread implementation in the fight against this pervasive issue.

Ultimately, a concerted effort to optimize the data labeling process will empower stakeholders—including utility companies, regulators, and policymakers—to adopt more robust detection strategies. This, in turn, will lead to a significant reduction in electricity theft and its associated economic impacts, fostering a more sustainable and equitable energy landscape.

REFERENCES

[1] Tanveer Ahmad, Dr Qadeer Ul Hasan, and Saleem Zada. Non-technical loss detection, prevention and suppression issues for ami in smart grid. *International Journal of Scientific & Engineering Research*, 6(3):217–228, 2015.

[2] Sameer Al-Dahidi, Osama Ayadi, Jehad Adeeb, and Mohamed Louzazni. Assessment of artificial neural networks learning algorithms and training datasets for solar photovoltaic power production prediction. *Frontiers in Energy Research*, 7:130, 2019.

[3] Husam H. Alkinani, Abo Taleb T. Al-Hameedi, and Shari Dunn-Norman. Data-driven recurrent neural network model to predict the rate of penetration: Upstream oil and gas technology. *Upstream Oil and Gas Technology*, 7:100047, 2021.

[4] M. Cao, J. Zou, L. Wei, X. Zhao, L. Zhang, and P. Li. Detection of power theft behavior of distribution network based on rbf neural network. *J. Yunnan Univ. Nat. Sci. Ed.*, 40(5):872–878, 2018.

[5] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[6] Divam Lehri and Arjun Choudhary. A survey of energy theft detection approaches in smart meters. In *Intelligent Energy Management Technologies: ICAEM 2019*, pages 9–24. Springer Singapore, 2020.

[7] Anish Jindal, Amit Dua, Kuljeet Kaur, Mukesh Singh, Neeraj Kumar, and Sukumar Mishra. Decision tree and svm-based data analytics for theft detection in smart grid. *IEEE Transactions on Industrial Informatics*, 12(3):1005–1016, 2016.

[8] Zhongzong Yan and He Wen. Electricity theft detection based on extreme gradient boosting in ami. *IEEE Transactions on Instrumentation and Measurement*, 70:1–9, 2021.

[9] Sravan Kumar Gunturi and Dipu Sarkar. Ensemble machine learning models for the detection of energy theft. *Electric Power Systems Research*, 192:106904, 2021.

[10] Jawad Nagi, Keem Siah Yap, Sieh Kiong Tiong, Syed Khaleel Ahmed, and Malik Mohamad. Nontechnical loss detection for metered customers in power utility us-

ing support vector machines. *IEEE Transactions on Power Delivery*, 25(2):1162–1171, 2009.

[11] Hao Huang, Shan Liu, and Katherine Davis. Energy theft detection via artificial neural networks. In *2018 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)*, pages 1–6. IEEE, 2018.

[12] Sook-Chin Yip, Chia-Kwang Tan, Wooi-Nee Tan, Ming-Tao Gan, and Ab-Halim Abu Bakar. Energy theft and defective meters detection in ami using linear regression. In *2017 IEEE International Conference on Environment and Electrical Engineering and 2017 IEEE Industrial and Commercial Power Systems Europe (EEEIC/I&CPS Europe)*, pages 1–6. IEEE, 2017.

[13] Paria Jokar, Nasim Arianpoo, and Victor CM Leung. Electricity theft detection in ami using customers' consumption patterns. *IEEE Transactions on Smart Grid*, 7(1):216–226, 2015.

[14] Paria Jokar, Nasim Arianpoo, and Victor CM Leung. Electricity theft detection in ami using customers' consumption patterns. *IEEE Transactions on Smart Grid*, 7(1):216–226, 2015.

[15] Min Xiang, Huayang Rao, Tong Tan, Zaiqian Wang, and Yue Ma. Abnormal behaviour analysis algorithm for electricity consumption based on density clustering. *The Journal of Engineering*, (10):7250–7255, 2019.

[16] Fei Xiao and Qian Ai. Electricity theft detection in smart grid using random matrix theory. *IET Generation, Transmission & Distribution*, 12(2):371–378, 2018.

[17] Sandeep Kumar Singh, Ranjan Bose, and Anupam Joshi. Energy theft detection for ami using principal component analysis based reconstructed data. *IET Cyber-Physical Systems: Theory & Applications*, 4(2):179–185, 2019.

[18] Xinlin Wang, Insoon Yang, and Sung-Hoon Ahn. Sample efficient home power anomaly detection in real time using semi-supervised learning. *IEEE Access*, 7:139712–139725, 2019.