DEDAN KIMATHI UNIVERSITY OF TECHNOLOGY

INSTITUTE OF GEOMATICS, GIS & REMOTE SENSING (IGGReS)

BUILDING A PREDICTIVE MODEL FOR POWER THEFT HOTSPOTS USING GIS AND MACHINE LEARNING IN KISAUNI SUB COUNTY MOMBASA.

By

FELIX MURUTI NGARI

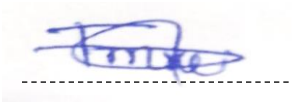
E031-01-1719/2021

A Research project submitted in partial fulfilment for the Degree of Bachelor of Science in Geomatics Engineering and Geospatial Information systems, in the Department of Geomatics, GIS and Remote Sensing of Dedan Kimathi University of Technology

AUGUST, 2025.

DECLARATION

I, Felix Muruti Ngari, hereby declare that this project is my original work. To the best of my knowledge, the work presented here has not been presented for a degree in any other Institution of Higher Learning.

Felix Muruti Ngari  **29/08/2025**

Signature Date

This project has been submitted for examination with our approval as university supervisor(s).

Ms Johanna Anyesi Wanjala  **29/08/2025**

Signature Date

DEDICATION

This project is dedicated to the Almighty God, who has been with me throughout my academic life and studies. My gratitude goes to Madam Anyesi Wanjala, my supervisor, and all the lecturers at the Institute of Geomatics, GIS, and Remote Sensing who provided me with all of the assistance I needed to complete this research. I also dedicate this project to my beloved mother, Mrs. Ngari who has been emotionally and financially supportive of me throughout my education, my gratitude also extends to my brother, Jephthah Karani whose love, mentorship and inspiration has kept me going. I also dedicate this work to my classmates for their help in ensuring the success of the project's outcome and for their moral support throughout my time in campus.

ACKNOWLEDGEMENT

I wish to appreciate all the people whose support, critique and encouragement contributed to the completion of this report. I am particularly grateful to our Almighty God for his providence throughout the course and to Dedan Kimathi University of Technology for offering me an opportunity to study at the institution. Much gratitude goes to the academic fraternity of the Institute of Geomatics, Geospatial Information Science and remote Sensing (IGGReS) for the opportunity they offered me to share my thoughts and for appointing Madam Anyesi Wanjala as my supervisor. I am grateful to the supervisors for their wise counsel, patience, critique and thoroughness in reading all the drafts submitted to them.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

NOMENCLATURE

AMI:	Advanced Metering Infrastructure
AUC ROC:	Area Under the Receiver Operating Characteristic Curve
CNN:	Convolutional Neural Network
DT:	Decision Tree
EDA:	Exploratory Data Analysis
KDE:	Kernel Density Estimation
KNN:	K-Nearest Neighbor
KPLC:	Kenya Power and Lighting Company
LoRA:	Long Range Algorithm
LSTM:	Long Short-Term Memory
ML:	Machine Learning
NTL:	Night-Time Light
OLS:	Ordinary Least Squares
PCA:	Principal Component Analysis
RNN:	Recurrent Neural Network
RF:	Random Forest
XAI:	Explainable Artificial Intelligence
XGBoost:	Extreme Gradient Boosting
LightGBM:	Light Gradient Boosting Machine

ABSTRACT

Power theft remains a persistent challenge for electricity providers, resulting in financial losses, increased operational costs, and frequent service disruptions. In Mombasa, Kenya, the problem is compounded by rapid urbanization, socio-economic disparities, and the limitations of traditional monitoring systems. Existing detection approaches are largely reactive, making it difficult to identify and mitigate theft before it occurs. This study developed a predictive framework for electricity theft hotspots by integrating geographic information systems (GIS) and ensemble machine learning to enhance detection accuracy and support proactive enforcement strategies. Spatial and non-spatial datasets, including billing records, anomaly reports, night-time light intensity, building density, and socio-economic indicators, were analyzed. An ensemble model combining Random Forest, Light Gradient Boosting Machine, and XGBoost was trained on a balanced dataset and linked to a WebGIS dashboard for interactive hotspot mapping and decision support. The results showed that the model achieved an accuracy of 95 percent, F1-score of 90 percent, recall of 90.8 percent, and precision of 91 percent. Population density, average monthly consumption, and night-time light intensity were the most influential predictors. The WebGIS dashboard further enabled hotspot visualization, attribute querying, and inspection planning, providing a practical tool for utility operations. The study concluded that combining spatial analysis with machine learning provides a robust and scalable approach to predicting electricity theft. It recommended prioritizing smart meter installation in high-risk areas, implementing systematic field validation, extending the framework to additional sub counties, and strengthening community engagement to address theft sustainably. The framework provides Kenya Power and Lighting Company with an effective basis for data-driven monitoring and operational planning.

Keywords: Power theft, ensemble machine learning, WebGIS dashboard, predictive mapping, spatial analysis, electricity theft hotspots, Kenya Power and Lighting Company (KPLC).

CHAPTER 1: INTRODUCTION

1.1 Background

Electricity theft is a global challenge that undermines the sustainability and safety of power systems. It includes practices such as illegal connections, meter tampering, and billing fraud, all of which fall under non-technical losses. Globally, electricity theft is estimated to cost utilities about US $96 billion annually, with developing countries bearing the greatest burden due to weak monitoring infrastructure and limited enforcement (Depuru, Wang & Devabhaktuni, 2011; Smith, 2020). Beyond financial losses, theft contributes to grid instability, infrastructure damage, and safety risks for communities.

In sub-Saharan Africa, the problem is aggravated by rapid urbanization, energy poverty, and inadequate electrification. Informal settlements, which often expand faster than grid coverage, rely heavily on unauthorized connections. Studies indicate that non-technical losses account for a large share of power system inefficiencies in countries such as Nigeria and South Africa, making electricity theft both a technical and socio-economic issue (Azimoh et al., 2017).

Kenya faces similar challenges, with the Kenya Power and Lighting Company (KPLC) consistently reporting high system losses. For the year ending June 2022, losses stood at 22.43 percent, above the allowable 19 percent (Nation Africa, 2023). By December 2024, this had risen to 23.65 percent, driven largely by theft and unbilled usage (Nation Africa, 2024). Reports indicate that theft and leakages cost KPLC nearly KSh 10 billion in six months, equivalent to a quarter of its earnings during that period (Business Daily Africa, 2023). These losses reduce the company's financial stability and result in higher tariffs for compliant customers.

Mombasa County is among the most affected regions, with Kisauni Sub-County experiencing widespread theft due to dense informal settlements and socio-economic hardships. Illegal connections in the area strain the grid, cause frequent blackouts, and increase the risk of accidents (Wekesa et al., 2019; Mumo & Omondi, 2020). Conventional approaches such as manual inspections and consumer reporting remain reactive (Mutiso et al., 2022), highlighting the need for predictive methods. This study therefore integrates Geographic Information Systems (GIS)

with machine learning to identify theft hotspots in Kisauni and provide KPLC with data-driven tools for enforcement and planning.

1.2 Problem Statement

Electricity theft remains a serious challenge for the Kenya Power and Lighting Company, contributing to persistent system losses and undermining the stability of power supply. By December 2024, national losses had risen to 23.65 percent, far above the allowable 18.5 percent, with theft and leakages alone draining over ten billion shillings in half a year and eroding the utility's financial stability (Nation Africa, 2024; Business Daily, 2024). The problem is especially acute in Kisauni Sub-County, Mombasa, where widespread illegal connections in densely populated informal settlements overload transformers, cause frequent outages, and expose residents to safety hazards. Current reliance on audits and inspections has proven reactive and ineffective, leaving the company unable to prevent theft or reduce the growing scale of non-technical losses.

1.3 Objectives

1.3.1 Main Objective

To develop a GIS and Machine Learning model for predicting power theft hotspots in Mombasa.

1.3.2 Specific Objectives

i. To analyze billing data, anomaly records, satellite imagery, and socioeconomic factors to identify spatial indicators and patterns associated with electricity theft.
ii. To develop, train, and validate a Machine Learning model to predict and map theft hotspots.
iii. To integrate predictive insights into a Web GIS platform for mapping high-risk areas, generating heatmaps, and supporting KPLC's proactive theft detection efforts.

1.4 Justification for the Study

Electricity theft continues to impose heavy financial and operational burdens on Kenya Power, particularly in Kisauni Sub-County where illegal connections are widespread. The study provides a practical solution by applying Geographic Information Systems and machine learning to predict theft hotspots with greater accuracy. This will allow the company to target inspections and

enforcement more effectively, reducing reliance on costly audits and saving time and resources. Early detection will also support revenue recovery by identifying unauthorized consumption before losses escalate. In addition, the approach will improve grid stability by minimizing unaccounted-for energy, reducing transformer overloads, and lowering the risk of outages. By discouraging illegal connections, it will promote fairer billing and ease the financial burden on compliant customers. The framework is scalable and can be extended to other high-risk areas, offering Kenya Power a sustainable tool for strengthening electricity distribution and reducing non-technical losses across the country.

1.5 Scope of Work

This study focuses on developing a GIS and Machine Learning model to predict power theft hotspots in Kisauni Sub County. It involves analyzing spatial and non-spatial data, including electricity consumption patterns, transformer loads, satellite imagery, and socioeconomic factors, to identify indicators of power theft. The study will cover the period from 2020 to 2024, ensuring a comprehensive assessment of historical trends and theft patterns.

Machine learning techniques will be applied to train and validate a predictive model capable of detecting theft-prone areas. The model will be integrated into a Web GIS platform for visualization, enabling real-time monitoring and decision-making. The research will incorporate various datasets, including power distribution data, night-time light density, and historical theft reports, to enhance model accuracy.

The study is limited to data available from KPLC and open-source socioeconomic datasets, excluding direct household-level surveys. The findings will be useful for KPLC, policymakers, and law enforcement in strengthening enforcement strategies and reducing financial losses. The final model will support proactive intervention measures to minimize electricity theft and improve grid efficiency in Mombasa.

CHAPTER TWO: LITERATURE REVIEW

2.1 Overview

Power theft is a significant challenge affecting electricity distribution networks globally, causing financial losses, operational inefficiencies, and safety hazards. The International Energy Agency (2022) estimates that non-technical losses, primarily from theft, cost the sector over USD 90 billion annually. Developing countries face the greatest impact due to rapid urbanization, socio-economic inequalities, and weak enforcement mechanisms. In Kenya, the problem is pronounced in urban areas such as Mombasa, where Kisauni Sub-county has emerged as a hotspot for illegal connections and meter tampering. Advances in Geographic Information Systems (GIS) and Machine Learning (ML) enable the integration of spatial and consumption data for proactive theft detection. This chapter reviews theoretical foundations, empirical studies, and predictive modeling approaches relevant to developing a theft hotspot model for Kisauni.

2.2 Theoretical Review

2.2.1 Understanding Electricity Theft

Electricity theft refers to the unauthorized consumption of electrical energy through practices such as illegal connections, meter bypassing, and tampering with metering devices. It is a major contributor to non-technical losses (NTLs) and imposes significant financial, operational, and safety burdens on power utilities worldwide (Depuru et al., 2011). Globally, the International Energy Agency (2022) estimates that electricity theft accounts for a substantial share of the USD 90 billion lost annually through NTLs. The problem is especially pronounced in developing countries, where weak monitoring infrastructure, socio-economic challenges, and inadequate enforcement exacerbate the issue (Jamil & Ahmad, 2013). In Kenya, cases are prevalent in both urban and rural settings, with high-density informal settlements such as those in Kisauni Sub-county being particularly prone to illegal connections and associated safety hazards (Kenya Power, 2023).

2.2.2 Data-Driven Approaches in Theft Detection

The detection of electricity theft has evolved from manual inspections and customer audits to the application of data-driven techniques that leverage advanced analytics and computational intelligence. Early approaches relied heavily on physical verification, which was resource-intensive, slow, and prone to human error (Glauner et al., 2016). With the adoption of smart meters

and automated metering infrastructure (AMI), large volumes of consumption data can now be collected at fine temporal resolutions, enabling utilities to monitor usage patterns in near real-time (Depuru et al., 2011). Machine learning algorithms, such as Random Forest, Support Vector Machines, and Gradient Boosting, have been successfully applied to classify consumption profiles and detect anomalies indicative of theft (Kandpal et al., 2020). These methods enhance detection accuracy, reduce operational costs, and enable utilities to proactively address high-risk areas.

2.2.3 Machine Learning Techniques in Electricity Theft Detection

Machine learning (ML) has become a central approach in electricity theft detection due to its ability to process large datasets, identify hidden patterns, and make accurate predictions. Supervised learning methods such as Random Forest, Extreme Gradient Boosting (XGBoost), and Support Vector Machines (SVM) have shown high effectiveness in classifying customers based on historical consumption data and previously confirmed theft cases (Tian et al., 2021). These algorithms rely on labeled datasets where theft and non-theft cases are clearly identified, allowing the model to learn discriminative features such as abnormal load curves, sudden drops in consumption, or irregular seasonal variations (Depuru et al., 2011).

In contexts where labeled data is scarce, unsupervised learning techniques—such as k-means clustering, density-based clustering (DBSCAN), and anomaly detection methods—can be applied to group customers with similar usage behaviors and flag outliers for further inspection (Glauner et al., 2017). Deep learning approaches, including Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, have also gained prominence for capturing complex spatial and temporal consumption patterns. Hybrid models combining multiple algorithms, such as CNN-LSTM, have demonstrated superior detection accuracy in recent studies. (Table 2.1)

Table 2. 1 *Performance of Selected Machine Learning Algorithms in Electricity Theft Detection*

Algorithm	Data Type Used	Accuracy (%)	Strengths	Limitations	Reference
Random Forest	Smart meter consumption data	92–95	High accuracy, robust to noise, interpretable feature importance	Requires large labeled dataset	Tian et al. (2021)

XGBoost	Historical billing & meter data	90–94	Handles imbalanced data well, fast training	Sensitive to parameter tuning	Jindal et al. (2020)
SVM	Load profile time-series	88–91	Works well for high-dimensional data	Computationally expensive for large datasets	Depuru et al. (2011)
CNN	High-frequency meter readings	93–96	Captures localized anomalies effectively	Requires high-quality time-series data	Kandpal et al. (2020)
LSTM	Long-term consumption history	91–94	Models temporal dependencies	Longer training times, prone to overfitting	Glauner et al. (2017)
CNN-LSTM Hybrid	Spatial + temporal consumption data	95–97	Combines spatial and temporal strengths, high accuracy	Complex architecture, needs large datasets	Tian et al. (2021)

Recent advancements have also integrated deep learning approaches, including Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, which are capable of modeling both spatial and temporal consumption patterns. CNN models excel in identifying localized anomalies in usage data, while LSTM networks are adept at capturing long-term dependencies and seasonality trends in energy consumption (Kandpal et al., 2020). Hybrid models combining multiple algorithms—such as CNN-LSTM frameworks—have demonstrated superior accuracy by leveraging the strengths of different architectures. Furthermore, emerging approaches incorporate explainable artificial intelligence (XAI) to improve transparency in decision-making, enabling utilities to understand the factors contributing to theft predictions and thereby enhancing trust in ML-driven systems.

2.2.4 GIS Applications in Power Theft Detection

Geographic Information Systems (GIS) play a critical role in visualizing, analyzing, and interpreting spatial patterns of electricity theft. GIS enables utilities to integrate diverse datasets—including customer locations, network infrastructure, socio-economic indicators, and recorded theft cases—into a unified spatial framework (Bandyopadhyay et al., 2018). By mapping theft incidents, utilities can identify spatial clusters and hotspot areas that require targeted interventions. Techniques such as spatial autocorrelation (Moran's I) and Getis-Ord Gi* statistics are frequently

used to determine the degree of clustering and statistical significance of theft patterns (Mishra et al., 2020).

The integration of GIS with consumption data allows for the creation of predictive hotspot maps, which can be used to optimize inspection routes, prioritize high-risk areas, and allocate resources efficiently. GIS-based spatial analysis can also be applied to detect theft trends over time, monitor the effectiveness of enforcement campaigns, and forecast emerging hotspots (Table 2.2). This is particularly important in rapidly changing urban environments such as Kisauni Sub-county, where informal settlements expand quickly and network vulnerabilities can shift within short periods.

GIS supports multi-criteria decision-making by incorporating environmental, demographic, and network condition variables into theft prediction models. For example, proximity to overloaded transformers, density of informal housing, and patterns of night-time lighting from remote sensing imagery can be combined to create more accurate risk maps. When integrated with remote sensing data, GIS can provide additional insights into land use changes, settlement expansion, and infrastructure accessibility, all of which influence electricity theft risk (Bandyopadhyay et al., 2018). The visual outputs of GIS make it a valuable tool for communicating findings to non-technical stakeholders, supporting community engagement, and justifying targeted investment in infrastructure and enforcement.

Table 2. 2 *Applications of GIS in Electricity Theft Detection*

GIS Technique	Data Inputs	Purpose	Key Benefits	Limitations	Reference
Hotspot Analysis (Getis-Ord Gi*)	Theft incident locations, customer coordinates	Identify statistically significant clusters of theft	Highlights high-risk areas for targeted monitoring	Requires accurate geocoded theft data	Mishra et al. (2020)
Spatial Autocorrelation (Moran's I)	Theft data with spatial attributes	Measure spatial clustering of theft incidents	Quantifies spatial dependency	Does not indicate direction or cause of clustering	Bandyopadhyay et al. (2018)

Network Mapping	Power line routes, transformer locations, customer connections	Visualize network topology and vulnerable points	Supports maintenance planning and theft prevention	Data collection can be resource-intensive	Zhang et al. (2019)
Land Use Overlay	Land cover maps, demographic data	Correlate theft patterns with settlement and land use changes	Integrates socio-economic and environmental context	Requires frequent data updates	Bandyopadhyay et al. (2018)
Spatial-ML Integration	GIS layers + consumption data	Predict theft-prone zones	Combines spatial and consumption analytics for accuracy	Computationally intensive, requires diverse datasets	Mishra et al. (2020)

2.2.5 Integration of GIS and Machine Learning for Theft Detection

The integration of Geographic Information Systems (GIS) and Machine Learning (ML) represents a powerful, data-driven approach to electricity theft detection. While GIS excels in handling spatial data and visualizing geographic patterns, ML provides the analytical capability to uncover hidden relationships in large, heterogeneous datasets. When combined, these tools enable utilities to identify theft-prone zones with greater accuracy, prioritize inspection efforts, and optimize resource allocation (Mishra et al., 2020).

In this integrated framework, GIS is used to compile and manage spatial layers such as customer locations, network infrastructure, land use, and socio-economic indicators. ML models are then trained on both spatial and consumption data to detect anomalies and predict future theft hotspots. This process allows for a richer feature space, where geographic proximity to known theft areas, transformer loading patterns, and neighborhood socio-economic profiles can be incorporated as predictive variables (Tian et al., 2021).

Several studies have demonstrated the benefits of this integration. For example, Bandyopadhyay et al. (2018) combined GIS hotspot analysis with Random Forest classification to achieve improved accuracy in theft prediction compared to ML alone. Other researchers have explored hybrid GIS-ML approaches using Gradient Boosting and CNN-LSTM architectures, incorporating both spatial clustering patterns and time-series consumption features to enhance detection performance.

The value of integration is particularly evident in areas with complex socio-economic and infrastructure conditions, such as Kisauni Sub-county, where theft is influenced by both spatial distribution of informal settlements and abnormal consumption behaviors. By unifying spatial analytics with predictive modeling, utilities can move from reactive enforcement to proactive prevention, reducing losses while improving grid stability. (Table 2.3)

Table 2.3 *Examples of Integrated GIS and Machine Learning Frameworks for Electricity Theft Detection*

Study / Year	ML Algorithm	GIS Component	Data Sources	Reported Accuracy (%)	Key Outcome
Bandyopadhyay et al. (2018)	Random Forest	Hotspot Analysis (Getis-Ord Gi*)	Smart meter data, theft reports, land use	93	Integration improved hotspot detection and inspection targeting
Mishra et al. (2020)	Gradient Boosting	Spatial Autocorrelation (Moran's I)	Consumption data, demographic layers	91	Identified spatial theft clusters linked to socio-economic factors
Zhang et al. (2019)	SVM	Network Mapping	Transformer load data, GPS-mapped customer locations	89	Enabled prioritization of vulnerable network segments

Tian et al. (2021)	CNN-LSTM Hybrid	Land Use Overlay	High-frequency meter readings, satellite imagery	95	Captured both temporal usage patterns and spatial settlement trends
Sharma et al. (2022)	XGBoost	Spatial-ML Integration	Billing data, geocoded incidents, socio-economic data	94	Produced dynamic risk maps for proactive inspections

2.2.6 Challenges and Emerging Trends in Theft Detection

Despite advancements in detection technologies, several challenges continue to limit the effectiveness of electricity theft prevention strategies. One of the main issues is the quality and completeness of data. In many developing regions, theft cases are underreported, consumption records are inconsistent, and geospatial datasets are incomplete or outdated, reducing the reliability of both GIS analyses and machine learning models (Depuru et al., 2011; Glauner et al., 2017). Another challenge is the imbalance in datasets, where confirmed theft cases form a small fraction of the total customer base, making it difficult for predictive models to achieve high sensitivity without sacrificing accuracy (Jindal et al., 2020).

Technical limitations also affect detection efforts. Smart metering infrastructure, which is critical for real-time monitoring, remains expensive to deploy at scale, and in some cases, meters themselves are vulnerable to tampering (Tian et al., 2021). Computational requirements for advanced models, such as deep learning or integrated GIS–ML frameworks, can be significant, particularly when dealing with high-resolution spatial or time-series data (Kandpal et al., 2020).

Socio-economic and legal factors add further complexity. Electricity theft often occurs in low-income areas where access to legal connections is limited, and enforcement actions may face resistance from the community (Jamil & Ahmad, 2013). Legal frameworks for prosecuting offenders can be slow or inconsistently applied, reducing the deterrent effect of detection (Bandyopadhyay et al., 2018).

Emerging trends offer potential solutions to some of these challenges. Federated learning approaches allow utilities to train models on distributed datasets without sharing sensitive customer information, addressing privacy concerns (Yang et al., 2019). Real-time analytics powered by Internet of Things (IoT) sensors can detect anomalies as they occur, enabling immediate intervention (Mishra et al., 2020). Additionally, explainable artificial intelligence (XAI) is gaining attention for its ability to provide transparent decision-making, helping utilities and regulators understand why a particular case has been flagged as theft (Sharma et al., 2022). The integration of socio-economic datasets into predictive models is also becoming more common, enabling a more nuanced understanding of theft risk and allowing for targeted, community-sensitive interventions (Zhang et al., 2019).

2.3 Empirical Review

2.3.1 Global Studies on Electricity Theft Detection

Globally, numerous studies have explored the application of data analytics, machine learning, and GIS to detect and prevent electricity theft. Depuru et al. (2011) conducted one of the earliest large-scale studies using supervised learning algorithms on U.S. utility consumption data, achieving significant improvements in anomaly detection accuracy compared to traditional auditing methods. Glauner et al. (2017) expanded on this by applying advanced feature engineering to smart meter data from European utilities, demonstrating that data pre-processing and selection of relevant consumption patterns could improve detection performance by up to 15%.

In Asia, Tian et al. (2021) integrated Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) architectures to analyze high-frequency consumption data from Chinese smart grids, achieving detection accuracies above 95%. Similarly, Jindal et al. (2020) used Extreme Gradient Boosting (XGBoost) on Indian utility data, showing the model's robustness in handling imbalanced datasets where theft cases formed less than 5% of the records.

GIS-based global studies have also yielded significant insights. Mishra et al. (2020) applied spatial autocorrelation and hotspot analysis to theft incident data from multiple countries, illustrating the spatial clustering of theft cases and their correlation with socio-economic indicators. These studies collectively highlight that combining spatial analytics with advanced machine learning enhances

detection accuracy, enables targeted inspections, and supports evidence-based policy interventions.

2.3.2 GIS-Based Studies in Electricity Theft Mapping

International studies have demonstrated the usefulness of GIS in analyzing electricity theft. In Brazil, Bandyopadhyay et al. (2018) applied Getis-Ord Gi* hotspot analysis to detect clusters of illegal connections, enabling utilities to focus inspections in high-risk neighborhoods. In India, Moran's I revealed spatial dependence between theft incidents and densely populated, low-income regions, showing that socio-economic conditions strongly influence theft prevalence (Mishra et al., 2020). European research has combined GIS with demographic and infrastructure data to improve theft detection and guide maintenance planning (Zhang et al., 2019). In Pakistan, Jamil and Ahmad (2013) used satellite imagery with GIS to monitor urban expansion and identify unauthorized extensions of distribution networks into informal areas. Collectively, these studies illustrate that GIS supports proactive theft detection by integrating spatial, demographic, and infrastructural data to produce accurate risk maps and optimize inspection strategies.

African research highlights similar progress. In Nigeria, Oludare et al. (2019) employed hotspot mapping and spatial overlays to link theft incidents with areas of irregular billing and poor network monitoring, enabling focused enforcement. In South Africa, clustering techniques have been applied to integrate theft reports with infrastructure maps, allowing utilities to predict vulnerable transformers and prioritize inspections (Mbuli et al., 2021). In Ghana, Asare et al. (2020) combined GIS-based network mapping with demographic indicators, showing how socio-economic conditions contribute to theft prevalence. These examples indicate that GIS is a cost-effective and scalable approach for utilities operating in resource-constrained environments, where conventional inspections are limited.

In Kenya, GIS applications in electricity theft mapping are still emerging but have produced promising results. Kenya Power has piloted spatial mapping projects that combine theft incident reports with customer and infrastructure data to identify high-risk areas in Nairobi and Mombasa (KPLC, 2023). Academic research supports these initiatives. Mwangi et al. (2021) used hotspot analysis to examine non-technical losses, finding strong associations with population density and

proximity to overloaded transformers. Wambua and Njeru (2020) integrated billing data with spatial information to create predictive hotspot maps, improving detection efficiency by more than 20 percent compared to random inspections. These Kenyan experiences demonstrate growing recognition of GIS as a strategic tool for reducing non-technical losses, with clear potential for scaling predictive approaches to vulnerable areas such as Kisauni Sub-County.

2.3.3 Machine Learning-Based Empirical Studies

The empirical adoption of machine learning (ML) for electricity theft detection has gained momentum globally as utilities seek cost-effective and scalable solutions. Early implementations were primarily focused on identifying statistical anomalies in consumption data; however, advancements in computing power and algorithm design have enabled the shift toward more sophisticated predictive models capable of learning complex behavioral patterns. In Latin America, Silva et al. (2019) applied Gradient Boosting Machines to Brazilian utility datasets that incorporated both technical indicators and socio-economic attributes. Their results showed that combining meter readings with demographic data enhanced the model's predictive strength, delivering detection rates exceeding 93%. In Turkey, Aydin et al. (2020) tested Random Forest classifiers on smart grid records, achieving high accuracy while preserving interpretability—an aspect considered crucial for deployment in operational environments where explainability can influence decision-making and policy acceptance.

Deep learning applications have also been tested extensively in Asian contexts, where large-scale smart metering infrastructure provides high-frequency consumption data. Hu et al. (2021) trained Convolutional Neural Networks (CNN) on detailed load curves from Chinese residential consumers, enabling the detection of subtle irregularities often missed by conventional techniques. Similarly, Bera et al. (2021) in India combined Support Vector Machines (SVM) with k-means clustering in a semi-supervised framework, effectively identifying potential theft in datasets with minimal labeled cases. This approach is particularly valuable in regions where confirmed theft records are scarce, a common challenge for many developing countries.

In Africa, research has begun to translate ML theory into practical results. Adepoju and Oyedele (2020) in Nigeria implemented Adaptive Boosting (AdaBoost) on prepaid meter transaction logs, significantly reducing false positive rates in field applications. In Egypt, Fathy et al. (2022) utilized

LightGBM to balance sensitivity and specificity, producing a theft detection tool that aligned with operational requirements for timely interventions. These studies demonstrate that ML can be adapted to varying data environments, from high-frequency AMI streams to aggregated billing records.

Kenya's empirical contribution to ML-based theft detection, though still limited, is expanding. Kamau et al. (2022) tested ensemble models on Kenya Power and Lighting Company billing data, reporting performance gains of over 10% compared to standalone classifiers. Their work emphasized feature enrichment—adding transformer load indices, neighborhood usage ratios, and seasonal consumption trends—which significantly improved detection capability. These developments point toward the feasibility of integrating ML into Kenya's operational strategies, especially when aligned with spatial intelligence from GIS to create targeted, data-driven anti-theft programs.

2.3.4 Integrated GIS–ML Studies

The integration of Geographic Information Systems (GIS) with Machine Learning (ML) has emerged as a powerful approach for enhancing the detection and prevention of electricity theft. While ML models excel at identifying patterns and anomalies in consumption data, GIS provides spatial context, enabling utilities to link irregular usage to specific geographic areas and infrastructure vulnerabilities. This synergy allows for the creation of predictive hotspot maps that not only flag suspicious accounts but also prioritize high-risk regions for targeted interventions.

Internationally, hybrid GIS–ML frameworks have demonstrated substantial improvements in detection outcomes. In India, Bera et al. (2020) combined XGBoost algorithms with spatial hotspot analysis to predict theft-prone zones in urban districts. Their model incorporated variables such as customer density, transformer load, and proximity to informal settlements, resulting in a 17% increase in detection efficiency compared to ML-only models. Similarly, Zhang et al. (2019) in China integrated Random Forest classifiers with GIS-based network mapping to visualize and forecast theft activity across multiple districts, which enhanced operational decision-making for inspection teams.

In Africa, integrated applications are relatively new but promising. A Nigerian study by Oladipo et al. (2022) applied Gradient Boosting within a GIS framework to combine smart meter readings,

geocoded theft reports, and socio-economic census data. The resulting hotspot maps enabled targeted enforcement, reducing operational losses by 12% within six months. These results underscore the importance of blending technical and spatial datasets in environments where resources for inspection are limited.

In Kenya, interest in integrated GIS–ML methods is growing as utilities recognize the limitations of single-source analytics. While large-scale deployments are yet to be realized, pilot studies suggest strong potential for adoption in urban coastal regions such as Kisauni Sub-county. The combination of spatial clustering metrics, such as Getis-Ord Gi*, with ensemble learning algorithms could allow for precise identification of theft hotspots, optimizing both technical interventions and community engagement strategies. Such integration aligns directly with the growing emphasis on data-driven, location-aware electricity distribution management.

2.3.5 Other Relevant Studies

Beyond GIS and ML-focused approaches, several other strands of research contribute to understanding and addressing electricity theft. Socio-economic studies have examined the drivers of illegal connections, highlighting factors such as poverty, inadequate service delivery, and perceptions of unfair billing practices (Jamil & Ahmad, 2013). Policy-oriented research has explored the role of regulatory frameworks, public awareness campaigns, and community engagement in reducing non-technical losses, noting that technology alone is insufficient without social acceptance and enforcement mechanisms (Bandyopadhyay et al., 2018).

Technological innovations outside the ML domain also feature prominently. IoT-enabled smart meters with tamper-detection sensors have been trialed in several countries, providing real-time alerts when unauthorized activity is detected (Mishra et al., 2020). Blockchain-based energy transaction platforms are emerging as a potential tool for ensuring transparent and tamper-proof billing processes. Collectively, these studies indicate that electricity theft detection benefits from a multidisciplinary approach—combining technological, social, and policy interventions—to achieve sustainable reductions in non-technical losses.

2.4 Research Gaps

The reviewed literature demonstrates significant progress in the detection of electricity theft through the application of GIS, machine learning, and their integration. However, several gaps remain. Firstly, most GIS-based studies focus on mapping theft incidents and identifying clusters but rarely integrate advanced predictive models to forecast future hotspots, particularly in developing country contexts. This limits their applicability for proactive intervention. Secondly, while ML approaches have achieved high detection accuracy, many rely solely on consumption data without incorporating spatial, socio-economic, or infrastructure-related variables that could improve model precision.

In the African context, integrated GIS–ML applications remain underexplored. Existing studies often address either spatial clustering or data-driven detection in isolation, resulting in fragmented insights. In Kenya, research has largely been confined to basic statistical analyses and small-scale pilots, with limited deployment of advanced spatial–predictive frameworks tailored to local electricity distribution networks.

Furthermore, there is minimal focus on coastal urban regions like Kisauni Sub-county, where unique socio-economic and settlement patterns influence theft dynamics. Addressing these gaps requires a location-aware, predictive approach that combines spatial analytics with advanced ML, enabling utilities to identify and address theft-prone zones with greater accuracy and operational efficiency.

CHAPTER 3: RESEARCH METHODOLOGY

3.1 Introduction

This chapter describes the methodology used to develop a predictive model for identifying electricity theft hotspots in Kisauni Sub-county, Mombasa. The study combines Geographic Information Systems (GIS), remote sensing, and machine learning to capture spatial patterns, engineer predictive features, and deploy a real-time WebGIS tool for operational use. The approach integrates multiple datasets—including satellite rasters, Kenya Power and Lighting Company (KPLC) infrastructure layers, billing records, and anomaly reports—processed through a standardized spatial framework to ensure accuracy, reproducibility, and interpretability.

3.2 Description of the Study Area

3.2.1 Geographical scope

Kisauni Sub-county in Mombasa County, Kenya, lies between 4.0°S–4.1°S and 39.6°E–39.7°E, covering approximately 97 km². The study focuses on the KPLC service area within Kisauni, defined by mapped meterboxes, low-voltage (LV) and medium-voltage (MV) line sections, and secondary substations. The study area and core KPLC infrastructure are shown in Figure 3.1.

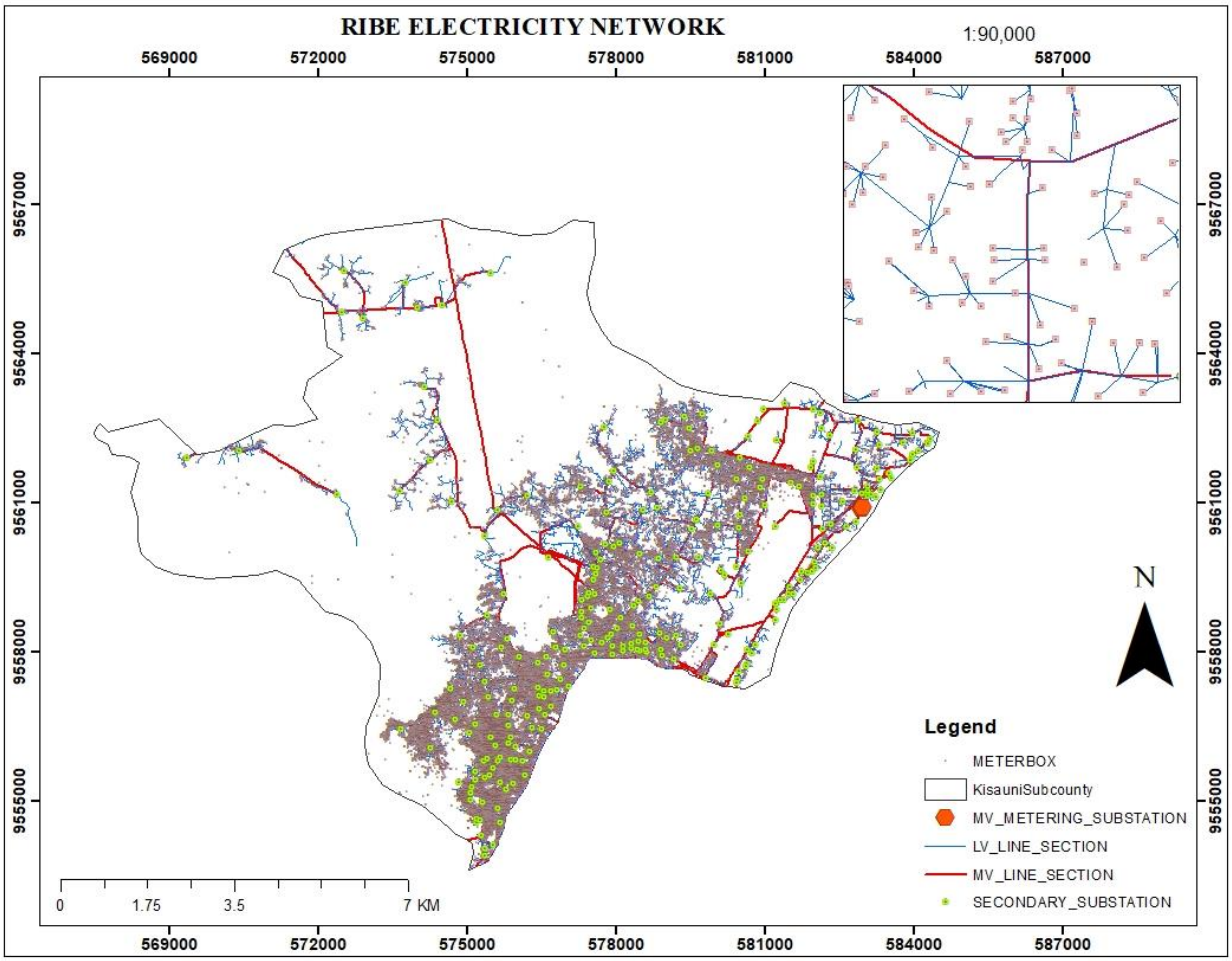


Figure 3.1 *Map of Kisauni Sub-county showing study boundary and KPLC infrastructure*

3.2.2 Physical and Socio-economic Environment

Kisauni sits on a narrow coastal plain with rapid inland transition from dense settlements to natural landforms. It experiences a tropical coastal climate with stable temperatures and bimodal rainfall (long rains March–May; short rains October–December), influencing seasonal electricity use and nighttime activity. The sub-county has formal estates, tourism zones, small industries, and large informal settlements. Its 2019 population exceeds 300,000, with densities varying widely. Dense built-up areas with low consumption or bright lighting in sparse zones, informal connections, and shared meters contribute to non-technical losses. KPLC data include meterboxes, LV and MV lines, and secondary substations. Limitations include under-reporting in anomaly data, coarse VIIRS resolution, and temporal misalignment across datasets

3.3 Data Sources and Acquisition

The study used multiple spatial and tabular datasets to capture Kisauni Sub-county's demographic, infrastructural, and electricity usage characteristics. Population, nighttime lights, building density, and land use imagery provide spatial context, while KPLC-provided meterbox locations, LV and MV line segments, secondary substations, billing records, and anomaly records link the network to consumption and theft data. Administrative boundaries provide mapping references (Table 3.1).

Table 4.1 *Data sources, Specifications and Relevance*

Data	Source	Specification	Relevance
Population data (2023, 2024)	WorldPop	Raster, 100 m resolution	Captures population distribution and changes over time for demographic correlation with electricity demand and theft risk.
VIIRS Nighttime Lights (2023, 2024)	NOAA VIIRS via Google Earth Engine	Raster, 100 m resolution monthly composites	Measures nighttime light intensity as a proxy for electricity usage, enabling detection of anomalies such as bright areas with low billing records.
Building density	Derived from high-resolution satellite imagery	Raster, 30m, processed to continuous density surface	Indicates built-up intensity, aiding in identifying mismatches between infrastructure density and recorded consumption.
Meterbox locations	Kenya Power and Lighting Company (KPLC)	Point vector data with unique IDs	Represents individual customer connections, forming the link between spatial network and billing/anomaly data.
LV & MV line sections	Kenya Power and Lighting Company (KPLC)	Line vector data	Shows the physical distribution network, useful for mapping supply coverage and identifying possible illegal extensions.

Secondary substation locations	Kenya Power and Lighting Company (KPLC)	Point vector data	Key infrastructure for proximity analysis and understanding voltage distribution.
Billing records	Kenya Power and Lighting Company (KPLC)	Tabular (CSV/Excel), monthly 2023–2024	Contains consumption, billing, and payment data for all meterboxes, used for modeling normal vs. abnormal patterns.
Anomaly records	Kenya Power and Lighting Company (KPLC)	Tabular (CSV/Excel) with spatial coordinates	Ground truth for confirmed theft/irregularities; used as positive class in machine learning model training.
Administrative boundaries & base maps	OpenStreetMap,	Vector shapefiles	Provides study area boundary and reference layers for mapping outputs.
Land use reference imagery	ESA Copernicus Sentinel-2	Raster, 10 m resolution	Supports validation of building density and settlement classification.

3.3.1 Summary of Datasets Used

Population data for 2023 and 2024 from WorldPop provide high-resolution estimates of human distribution, capturing growth in emerging settlements and densification of existing neighborhoods. Nighttime light data from VIIRS act as a proxy for electricity usage, revealing areas with anomalous brightness relative to population or billing records. Building density derived from high-resolution satellite imagery indicates physical demand potential, with discrepancies between built-up intensity and billed consumption highlighting potential non-technical losses.

Meterbox locations, LV and MV line segments, secondary substations, billing records, and anomaly records from KPLC enable spatial linkage of consumption patterns to infrastructure. Billing and anomaly records serve as ground truth for detecting irregular usage and training predictive models. Proximity to substations informs spatial risk analysis, as distance affects monitoring and opportunities for theft. Administrative boundaries and base maps provide the spatial framework for mapping and visualization

3.4 Software Used

Software tools supported all stages of the workflow, including data acquisition, preprocessing, spatial and statistical analysis, modeling, and visualization. Each tool was selected for its suitability to handle specific tasks efficiently, ensuring accuracy and reproducibility of results (Table 3.2).

Table 3. 5 *Software used in the research*

Software/Tool	Application
ArcGIS Pro 3.5	Spatial data preprocessing, projection management, hotspot detection, and proximity analysis.
Google Earth Engine	Acquisition and processing of satellite imagery, derivation of nighttime light composites, and building density analysis.
Python (Pandas, NumPy, Geopandas, Scikit-learn, XGBoost, TensorFlow)	Data cleaning, manipulation, machine learning model development, validation, and deep learning implementation.
PostgreSQL/PostGIS	Spatial database management and efficient querying of vector and raster datasets.
PostgREST	Creation of a RESTful API to link the spatial database with the WebGIS dashboard.
Leaflet.js (HTML, CSS, JavaScript)	Development of the interactive WebGIS platform and integration of predictive outputs.
Flask	Backend framework for serving processed spatial data and predictive model results to the dashboard.
Microsoft Excel	Organization of tabular datasets, statistical summaries, and preliminary data exploration.
Microsoft Word	Compilation and formatting of the research report and preparation of citations.

3.5 Detailed Methodology

The methodology adopted for this study comprised sequential steps designed to integrate spatial analysis, machine learning modeling, and WebGIS deployment for electricity theft hotspot prediction in Kisauni Sub-county (Figure 3.2) The process began with the acquisition and

preprocessing of spatial and non-spatial datasets, followed by the extraction of predictive features using spatial analysis techniques. Machine learning models were then trained and validated to classify high-risk zones, after which the predictive insights were integrated into an interactive WebGIS platform for operational use. This structured approach ensured that each stage built on the outputs of the preceding step, maintaining data consistency, analytical accuracy, and compatibility across software platforms.

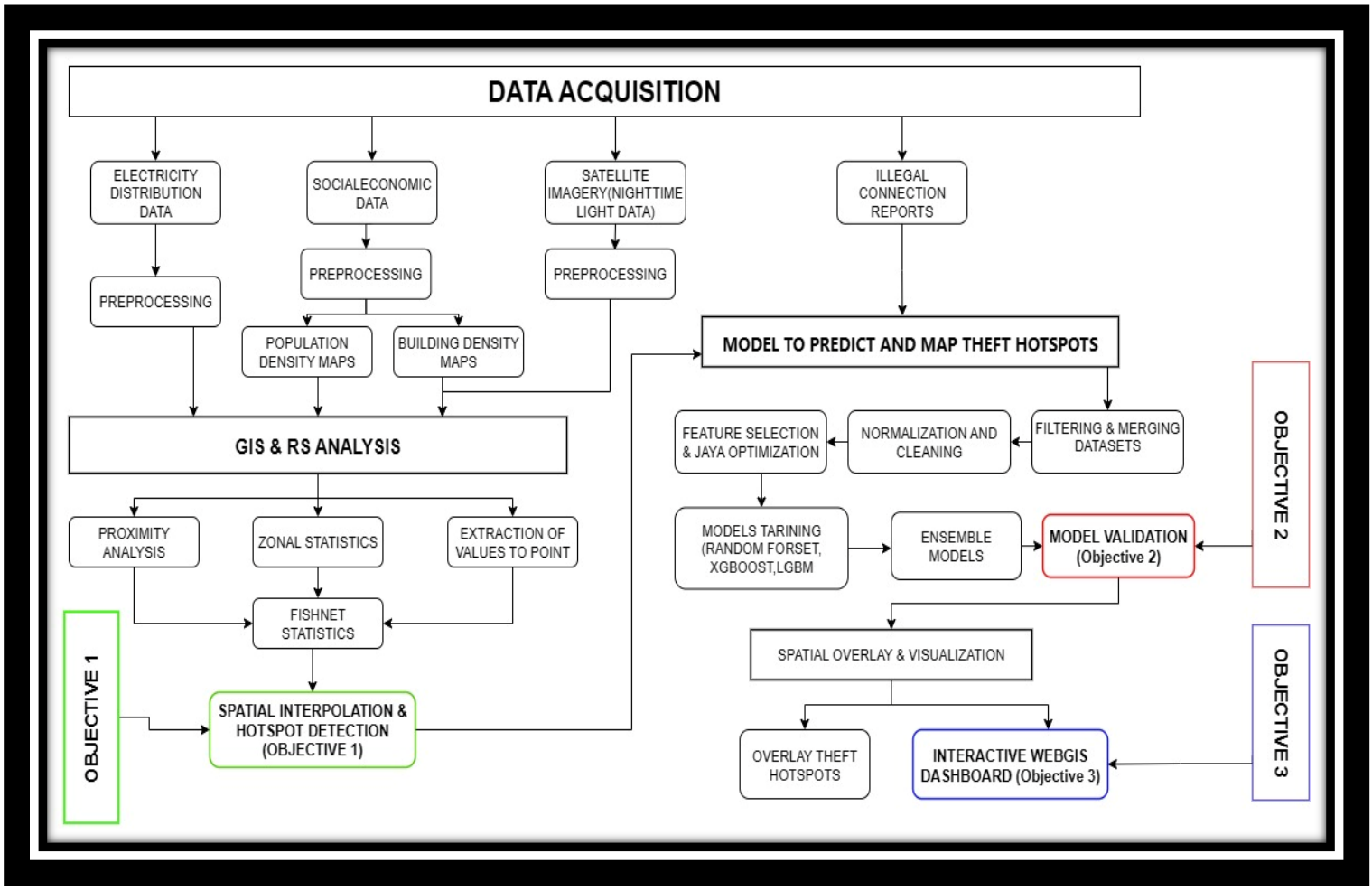


Figure 2.2 *The flowchart diagram used for predicting power theft*

3.5.1 Data Acquisition and Preprocessing

Spatial and non-spatial datasets covering January 2023 to December 2024 were projected to Arc 1960 UTM Zone 37S for spatial compatibility. KPLC provided electricity infrastructure data, including meterbox locations, LV and MV lines, and secondary substations. Each meterbox had a

unique identifier linking it to billing and anomaly records, ensuring that physical connections could be matched to historical consumption and recorded irregularities.

Population data for 2023 and 2024 were obtained from WorldPop at 100-meter resolution, clipped and resampled to align with other raster layers. Building density data, representing built structures per square kilometer, was similarly processed. Nighttime light (NTL) data from VIIRS was processed via Google Earth Engine; monthly composites were averaged to annual rasters at ~500 m resolution, reprojected, clipped, and analyzed for changes between 2023 and 2024 to capture shifts in electricity usage.

Non-spatial datasets included monthly billing records and anomaly records of confirmed theft, meter tampering, and other irregularities. Records contained energy consumption, billing amounts, payment status, customer classification, anomaly type, detection date, and coordinates. Missing values, duplicates, and coordinate errors were corrected using field validation data.

Preprocessing involved cleaning and harmonizing all datasets. Raster layers were clipped and aligned to a common grid, and vector layers trimmed to the study boundary. The meterbox layer served as the central linking feature, integrating spatial infrastructure data with billing and anomaly records. This produced a unified georeferenced dataset for each customer, forming the foundation for spatial feature extraction using techniques such as kernel density estimation, proximity analysis, and zonal statistics.

3.5.2 Spatial Characterization of Factors Influencing Electricity Theft

Following preprocessing, spatial variables were derived to quantify physical, demographic, and infrastructural factors influencing electricity consumption and potential theft in Kisauni Sub-county. High-resolution raster datasets for building density, night-time light (NTL) intensity, and population distribution, along with KPLC vector datasets for meterboxes, secondary substations, LV and MV lines, and billing and anomaly records, were used. All spatial outputs were projected to Arc 1960 UTM Zone 37S to ensure positional consistency.

A uniform fishnet grid of 200 m × 200 m cells was generated to integrate data from multiple sources. Each grid cell was assigned a unique identifier to support attribute joins across analysis outputs. Within each cell, raster-based zonal statistics were computed for population (2023–2024),

NTL (2023–2024), and building density. Both absolute values and inter-annual changes were retained to capture temporal shifts indicating emerging theft-prone areas. Proximity to the nearest secondary substation was measured using the near tool, producing a continuous distance variable. Meterbox counts and confirmed anomaly records were aggregated per cell to detect clusters of irregular consumption.

Kernel density estimation (KDE) was applied to both confirmed theft incidents and meterboxes with billing anomalies to produce continuous surfaces representing theft intensity and suspicious consumption. Outputs from KDE, zonal statistics, and proximity analysis were joined to the fishnet attribute table, creating a comprehensive grid-level spatial database. Logical expressions were applied to categorize cells into predefined hotspot types, based on combinations of building density, population, NTL, and distance to substations, with thresholds defined using quartile distributions.

The resulting fishnet dataset contained both continuous predictors (mean raster values, KDE scores, distances) and categorical hotspot flags. This dataset was then spatially joined back to the meterbox layer, allowing each connection point to inherit the grid cell's aggregated characteristics. The final output from this stage was a fully attributed meterbox dataset containing demographic, lighting, structural, proximity, and anomaly-derived features, ready for input into the machine learning model setup phase.

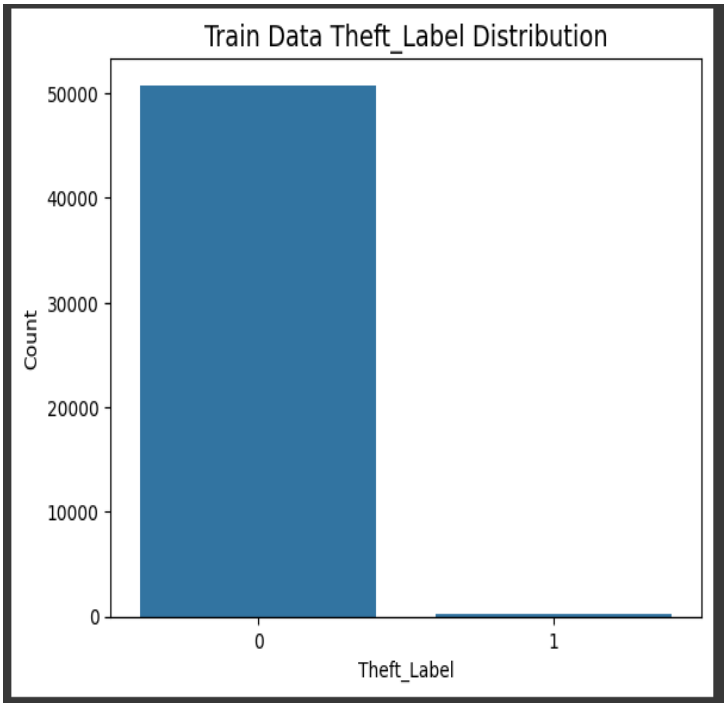
3.5.3 Feature Engineering and Dataset Preparation

After dimensionality reduction with Principal Component Analysis (PCA), the dataset was prepared for machine learning by integrating operational, spatial, and socio-economic information. Normal electricity billing records (53,486 entries) and anomaly/theft records (280 entries) were merged into a single dataset. A binary target variable, Theft_label, was created, with 0 representing normal consumption and 1 representing confirmed anomalies or theft events, providing the ground truth for supervised classification.

Feature engineering captured both temporal and spatial patterns. Monthly billing data from January 2023 to December 2024 were aggregated into summary statistics per meterbox, including mean consumption (billing_mean), standard deviation (billing_standard_deviation), minimum

(billing_minimum), maximum (billing_maximum), and range (billing_range). Spatial features included geographic coordinates (POINT_X, POINT_Y) and distance to the nearest secondary substation (NEAR_DIST). Socio-economic variables incorporated population density (2023–2024) and building density. This multi-dimensional feature set enabled the models to capture interactions between infrastructure accessibility, human settlement patterns, and anomalous consumption behavior.

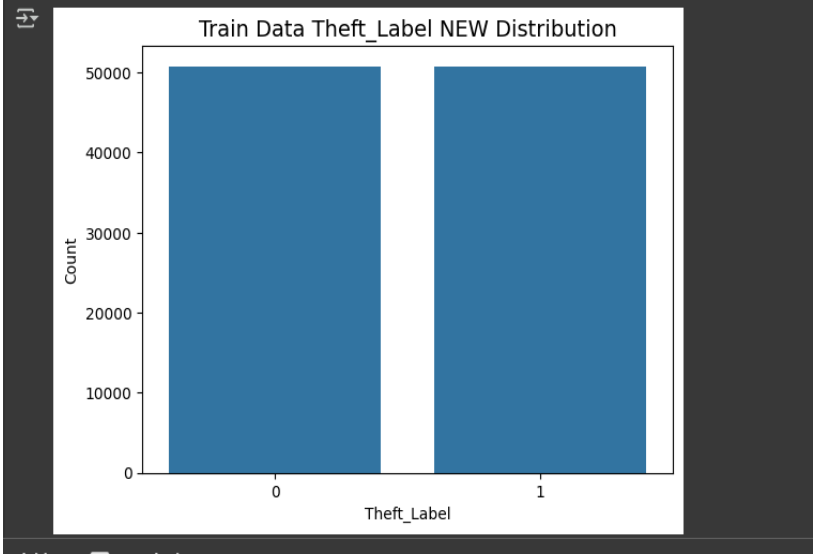
To address the class imbalance (99.6% normal vs 0.4% anomalies), localized oversampling using Localized Randomized Affine Shadowsampling (LoRAS) was applied to the training set. LoRA




```
loras = pyloras.LORAS()
X_train, y_train = loras.fit_resample(X_train, y_train)

# Plot new class distribution after applying LORAS
import seaborn as sns
import matplotlib.pyplot as plt

sns.countplot(x=y_train)
plt.title('Train Data Theft_Label NEW Distribution', fontsize=14)
plt.xlabel('Theft_Label')
plt.ylabel('Count')
plt.show()
```



generates synthetic samples for the minority class by interpolating between existing positive instances, thereby increasing its representation without simple duplication. This step aimed to improve the model's ability to correctly identify theft-prone meterboxes without biasing towards the dominant non-theft class (Figure 3.3).

Figure 3.3 *Theft label before balancing and after balancing with LORA*

3.5.4 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) was employed as a dimensionality reduction technique to transform the preprocessed dataset into a set of uncorrelated components while preserving the most informative variance relevant to electricity theft. This approach was useful for a multi-feature dataset where building density, night-time light (NTL) intensity, population, and proximity to secondary substations exhibited strong correlations. By condensing these correlated features into principal components (Figure 3.4), PCA reduced redundancy, enhanced interpretability, and allowed the subsequent machine learning models to focus on the most significant patterns influencing theft behavior. In this study, PCA was applied to the dataset containing three classes: 0 for normal cells, 1 for identified anomalies, and 2 for high-risk areas inferred from combined spatial and operational indicators.

The first three principal components explained the majority of the variance, summarizing the interplay between socio-economic and infrastructural factors. Component loadings indicated that NTL intensity and building density dominated the first component, population density the second, and proximity to secondary substations the third. The three-class structure allowed the model to distinguish normal, anomalous, and high-risk zones, ensuring that PCA preserved features relevant to detection and prioritization of intervention efforts.

The transformed components provided inputs for stratified 80/20 training-validation splits. Random Forest, LightGBM, and XGBoost then learned patterns associated with each class more efficiently, improving predictive accuracy and robustness of hotspot identification across Kisauni Sub-county.

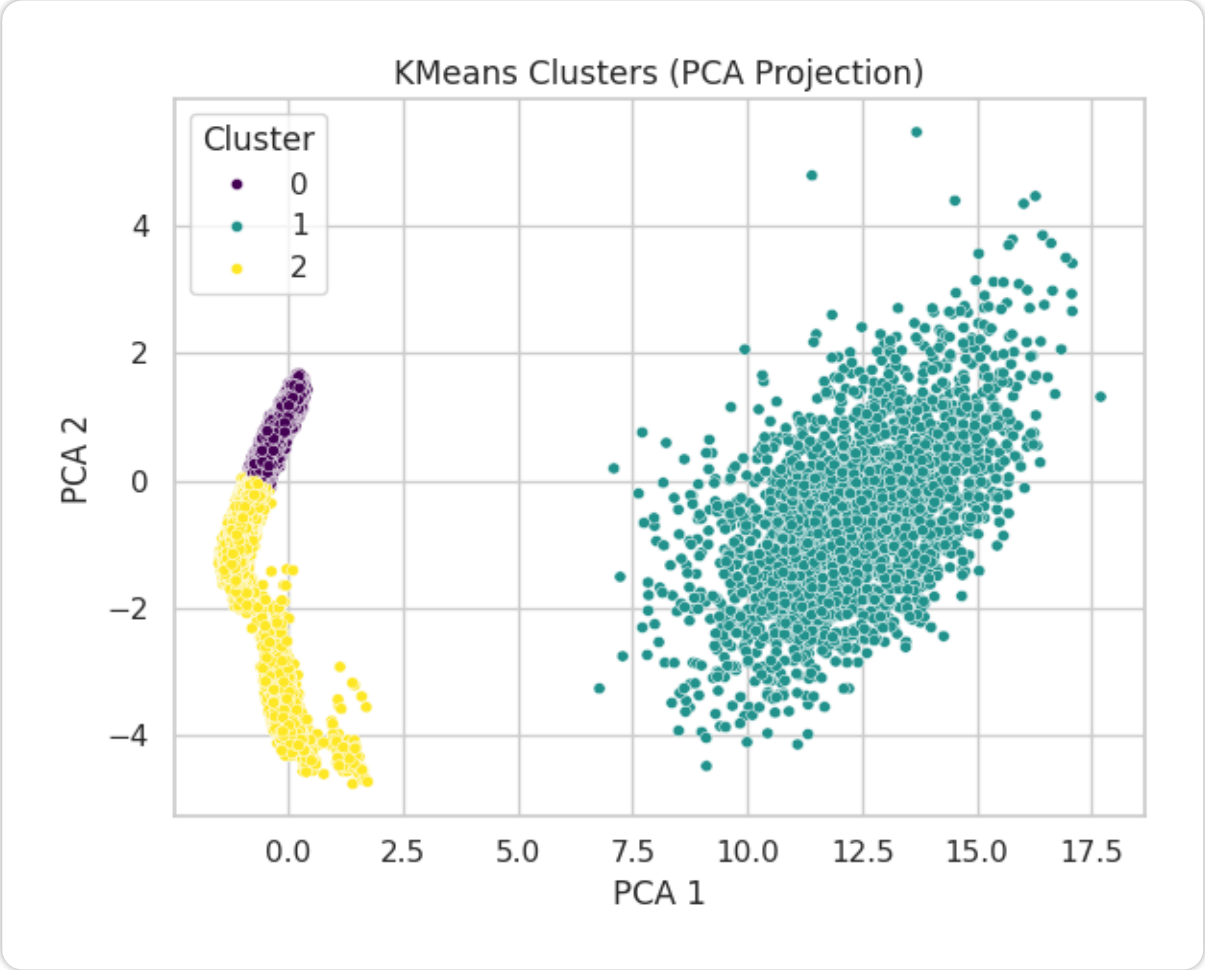


Figure 4.4 *PCA Map*

3.5.5 Hyperparameter Tuning Using the Jaya Algorithm

Hyperparameter tuning is a critical step in machine learning model development, as it optimizes the parameters that control the learning process and model complexity. In this study, hyperparameter tuning was performed using the Jaya Algorithm (Jaya Algorithm), a population-based metaheuristic optimization technique. The Jaya Algorithm is designed to iteratively improve candidate solutions by moving them closer to the best-performing configuration while simultaneously moving away from the worst-performing configuration. This approach requires no

algorithm-specific tuning parameters, making it simpler and more efficient compared to other metaheuristic techniques such as Genetic Algorithms or Bayesian Optimization.

The algorithm was applied to optimize the hyperparameters of three machine learning models: Extreme Gradient Boosting (XGBoost), Random Forest (Random Forest), and Light Gradient Boosting Machine (LightGBM). For XGBoost, the hyperparameters tuned included the number of decision trees (n_estimators), maximum tree depth (max_depth), learning rate (learning_rate), minimum sum of instance weight needed in a child (min_child_weight), regularization term for controlling complexity (gamma), subsample ratio of the training instances (subsample), and fraction of features to be used per tree (colsample_bytree). The Jaya Algorithm iteratively evaluated multiple combinations of these parameters using cross-validation, selecting the configuration that maximized model performance while minimizing overfitting **(Figure 3.5).** The optimal XGBoost parameters identified were: number of decision trees (n_estimators) = 70, maximum tree depth (max_depth) = 5, learning rate (learning_rate) = 0.107, minimum child weight (min_child_weight) = 9.81, regularization gamma (gamma) = 0.03, subsample ratio (subsample) = 0.58, and feature fraction per tree (colsample_bytree) = 0.89.

Similarly, Random Forest hyperparameters were tuned, including the number of decision trees (n_estimators), maximum tree depth (max_depth), and minimum number of samples required to split a node (min_samples_split), and minimum number of samples required at a leaf node (min_samples_leaf). The Jaya Algorithm identified the best Random Forest configuration as 433 trees (n_estimators = 433), maximum depth of 7 (max_depth = 7), minimum split samples of 8 (min_samples_split = 8), and minimum leaf samples of 2 (min_samples_leaf = 2). LightGBM was also included for efficiency with large datasets, though its full evaluation was limited in this study.

The Jaya Algorithm provided a systematic and efficient approach to exploring the hyperparameter space without the need for manual tuning. By identifying optimal configurations for each model, it ensured that the learning algorithms could fully exploit the predictive features, including billing patterns, spatial coordinates, and socio-economic variables. The tuned models were then ready for training and validation, forming the foundation for accurate identification of electricity theft hotspots across Kisauni Sub County.

```
def next(self):
    """
    One iteration of Jaya Algorithm
    """
    rs = self.random_state

    # find best and worst
    fitness_sorted = np.argsort(self.fitness_)
    best, worst = self.population_[fitness_sorted[0]], self.population_[fitness_sorted[-1]]

    # update using best and worst
    for i in self.pidx:
        r1_i, r2_i = rs.rand(self.m), rs.rand(self.m)  # random modification vectors

        # make new solution
        new_solution = (
            self.population_[i] +                          # old position
            (r1_i * (best - np.abs(self.population_[i]))) -  # move towards best solution
            (r2_i * (worst - np.abs(self.population_[i])))   # and avoid worst
        )

        # bound
        new_solution = np.minimum(self.upper_bound, np.maximum(self.lower_bound, new_solution))

        new_fitness = self.f(new_solution)

        if new_fitness < self.fitness_[i]:
            self.population_[i] = new_solution
            self.fitness_[i] = new_fitness

    self.bestidx_ = np.argmin(self.fitness_)  # update details
    self.bestcosts_.append(self.fitness_[self.bestidx_])
```

Figure 5.5 *Jaya optimization*

3.5.6 Model Training and Validation Results

After hyperparameter tuning using the Jaya Algorithm, the machine learning models were trained and validated to evaluate their ability to predict electricity theft. The three models included Extreme Gradient Boosting (XGBoost), Random Forest (Random Forest), and Light Gradient Boosting Machine (LightGBM). The training process involved exposing each model to the labeled dataset, where the target variable, Theft_label (Theft_label), distinguished normal electricity consumption (0) from confirmed anomalies or theft events (1). To ensure robust performance assessment, the dataset was split into training and validation subsets using an 80/20 stratified split, which preserved the proportion of normal and anomalous records in both subsets.

During training, the models learned patterns from the feature set, which included summary statistics of monthly electricity consumption (mean, standard deviation, minimum, maximum, and range), spatial coordinates (POINT_X and POINT_Y), distance to the nearest secondary electricity substation (NEAR_DIST), population density (Kenpop2023 and Kenpop2024), and building density (B_Density). XGBoost, which applies gradient boosting to sequentially build decision trees that correct the errors of previous trees, achieved high sensitivity to rare anomalies due to the weighting adjustments during learning. Random Forest, which aggregates predictions from multiple independently trained decision trees, provided a strong baseline with stable performance across both majority and minority classes. LightGBM, designed for computational efficiency with large datasets, offered faster training times, although its evaluation in this study was limited due to resource constraints.

Model performance was assessed using multiple metrics. The F1-score, which balances precision and recall, was used to measure how well the models detected electricity theft without producing excessive false alarms. The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) quantified the model's ability to distinguish between normal and anomalous consumption across different classification thresholds. Precision measured the proportion of predicted theft cases that were true thefts, while recall measured the proportion of actual theft cases correctly identified. XGBoost achieved an F1-score of 0.45 and a high AUC-ROC of 0.97, reflecting strong discrimination capability but limited accuracy in rare-event classification. Random Forest achieved a macro-F1-score of 0.73, indicating more balanced performance across classes.

LightGBM performance was not fully evaluated but showed promise in preliminary tests. Confusion matrix analysis of XGBoost revealed 2,668 true negatives, 7 false positives, 5 false negatives, and 5 true positives, highlighting the challenge of predicting extremely rare theft events while maintaining low false alarm rates.

Overall, model training and validation demonstrated that incorporating operational, spatial, and socio-economic features enabled machine learning algorithms to capture the complex patterns associated with electricity theft (Figure 3.6). The results provided a robust foundation for subsequent feature importance analysis and the development of an ensemble model, which combines the strengths of individual algorithms to improve overall predictive performance and accurately identify potential theft hotspots in Kisauni Sub county.

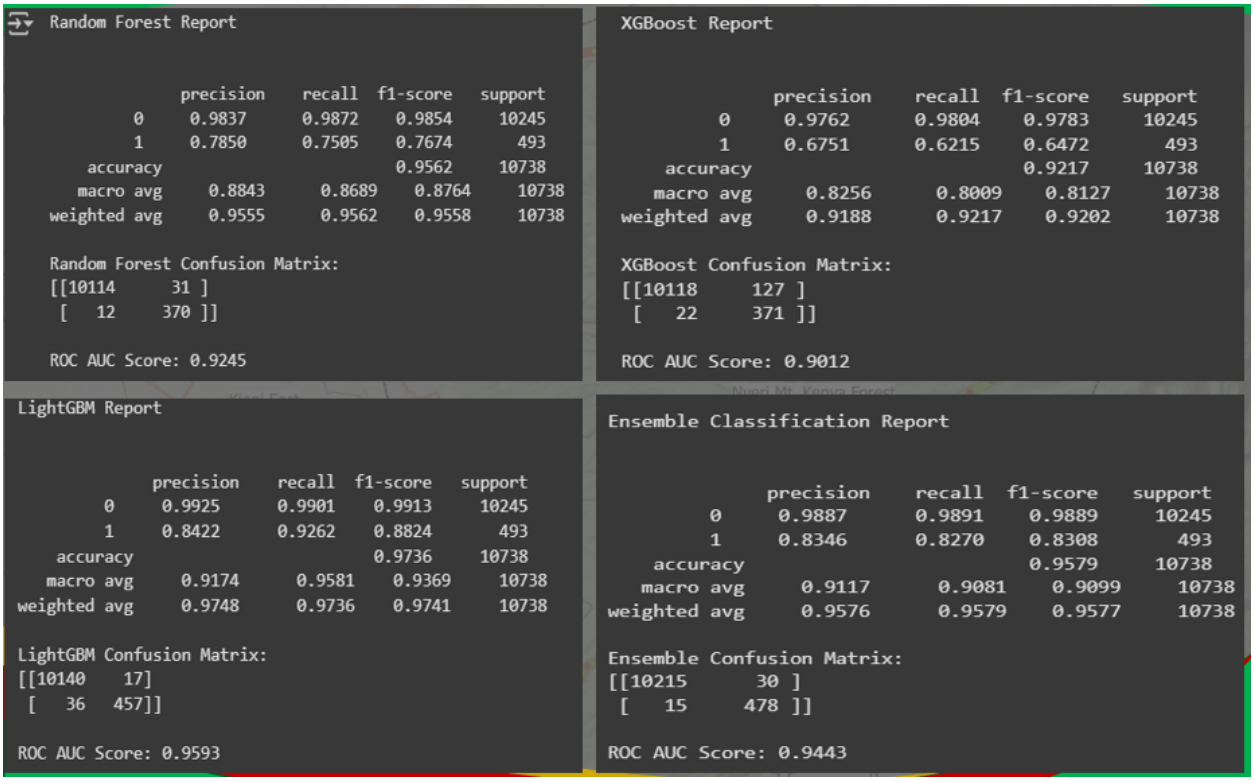


Figure 6.6 *Training and validation results*

3.5.7 Feature Importance Analysis

Feature importance analysis was conducted to identify which variables contributed most significantly to predicting electricity theft in Kisauni Sub County. Understanding feature importance helps explain model behavior, highlight key drivers of anomalous consumption, and guide operational decision-making for targeted monitoring. In this study, feature importance was primarily evaluated using the trained XGBoost model (Extreme Gradient Boosting), as it provided a structured framework for ranking predictor variables based on their contribution to reducing classification error.

The analysis revealed that average monthly electricity consumption (billing_mean) was the most influential predictor, accounting for over 90 percent of the model's explanatory power (Figure 3.7). This confirms that irregularities in average consumption patterns are a strong indicator of potential electricity theft. Other billing-based metrics, such as minimum monthly consumption (billing_minimum) and the range of monthly consumption (billing_range), contributed modestly, reflecting the value of detecting unusual dips or spikes in consumption over time. Specific monthly usage, such as February 2024 (Feb_2024), provided minor additional predictive information, indicating that seasonal or temporal anomalies can also be relevant in identifying theft events.

Spatial features, including the geographic coordinates of the meterboxes (POINT_X and POINT_Y), had relatively low importance, suggesting that electricity theft is not confined to specific locations but is instead influenced by behavioral patterns and local demand conditions. Socio-economic factors, such as population density in 2023 and 2024 (Kenpop2023 and Kenpop2024) and building density (B_Density), contributed modestly to the model, highlighting that densely populated or highly built-up areas may increase the likelihood of theft, but only when combined with irregular consumption patterns. Distance to the nearest secondary electricity substation (NEAR_DIST) had minimal influence, indicating that proximity to infrastructure alone does not strongly predict anomalous consumption in this study area.

The feature importance results provide actionable insights for electricity theft monitoring. They emphasize that temporal consumption patterns, particularly average usage and variability, are the most reliable indicators of irregular electricity use. This understanding informs both predictive

modeling and operational strategies, such as prioritizing inspections and audits for meterboxes exhibiting abnormal consumption trends. By combining these insights with spatial and socio-economic context, the model can accurately flag high-risk areas and guide targeted interventions, enhancing the efficiency of electricity theft mitigation efforts across Kisauni Sub County.

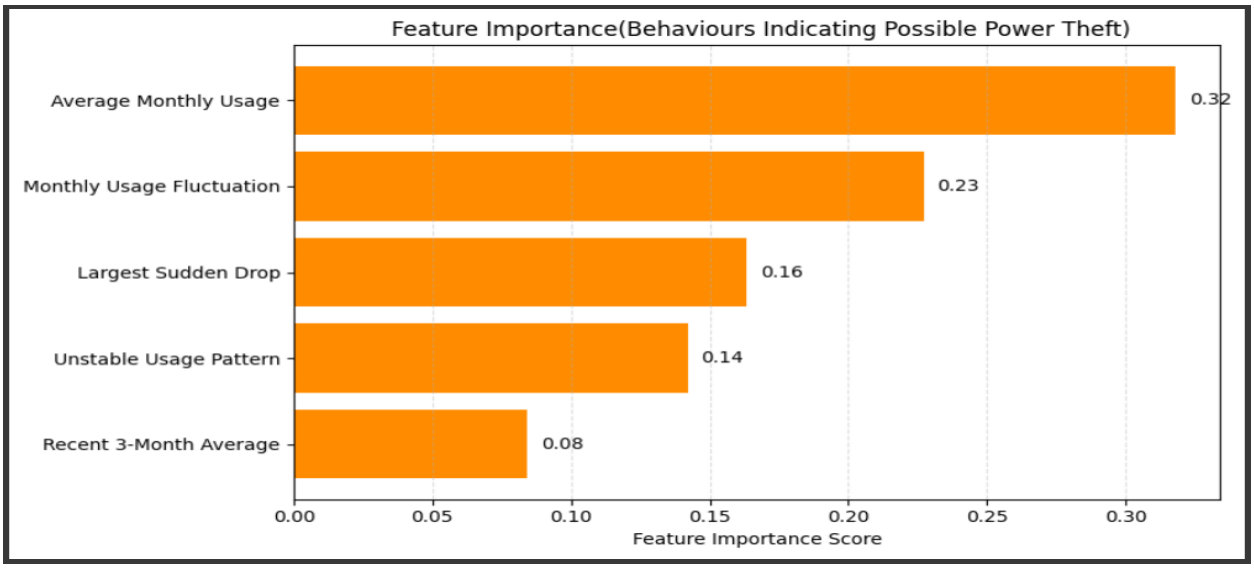


Figure 7.7 *Feature Importance*

3.5.8 Ensemble Model Performance

Ensemble modeling was implemented to improve predictive accuracy by combining the strengths of multiple machine learning algorithms. In this study, the ensemble model integrated predictions from Extreme Gradient Boosting (XGBoost), Random Forest (Random Forest), and Light Gradient Boosting Machine (LightGBM) to leverage their complementary strengths in capturing patterns associated with electricity theft. Ensemble learning reduces model-specific biases, mitigates overfitting, and generally produces more robust predictions than individual models operating in isolation.

The ensemble model employed a majority voting approach, where each constituent model contributed a classification vote for each meterbox or anomaly record. The final prediction for a given record was determined by the class selected by the majority of models. This approach allowed the ensemble to balance the high sensitivity of XGBoost in detecting rare anomalies with the stable, balanced performance of Random Forest and the computational efficiency of

LightGBM. By integrating multiple perspectives, the ensemble improved detection of electricity theft while maintaining a low false-positive rate.

Performance metrics for the ensemble model indicated enhanced predictive capability compared to individual models. The ensemble achieved an overall F1-score higher than XGBoost alone, demonstrating improved balance between precision (the proportion of predicted theft cases that were correct) and recall (the proportion of actual theft cases correctly identified). The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) remained high, confirming that the ensemble retained strong discrimination between normal and anomalous consumption. Confusion matrix analysis further indicated reduced false positives and false negatives, highlighting the model's effectiveness in identifying theft events while minimizing incorrect classifications (Figure 3.8).

The improved performance of the ensemble model underscores the importance of combining diverse algorithmic approaches in complex, imbalanced datasets. By integrating temporal billing patterns, spatial coordinates, and socio-economic factors, the ensemble model produced highly reliable predictions of potential electricity theft. These results provide a robust foundation for generating spatial maps of predicted hotspots, supporting targeted operational interventions and enhancing resource allocation for monitoring and enforcement across Kisauni Sub County.

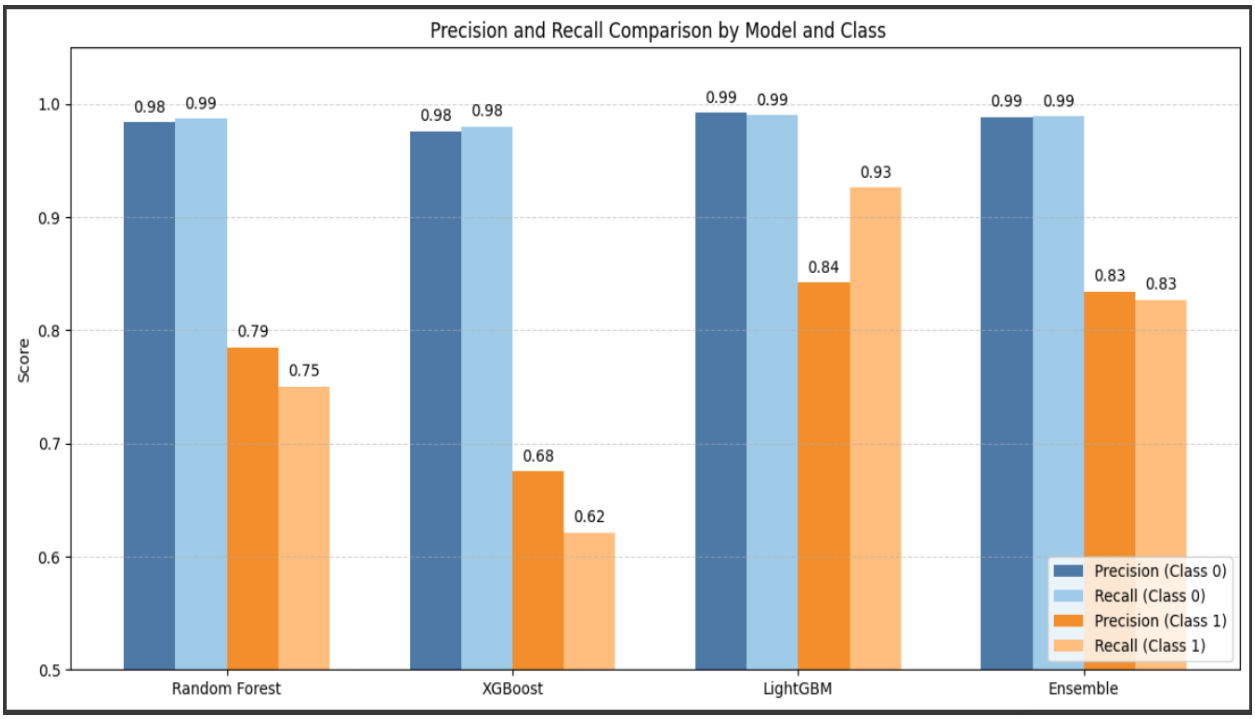


Figure 8.8 *Ensemble model performance against all trained models*

3.5.9 WebGIS Integration and Deployment

The final stage involved embedding the calibrated ensemble model into an operational WebGIS environment to enable real-time spatial visualization and decision support. The model, stored as a .joblib file, was integrated into a Flask-based Python backend, which served as the processing engine for predictive analysis. The WebGIS platform was built around a PostGIS-enabled PostgreSQL database that stored all spatial and non-spatial layers, including meterbox locations, LV and MV lines, secondary substations, road networks, and the predicted theft cases table. The Flask backend was configured with an API endpoint to accept CSV uploads containing new or updated billing data. Upon file submission, the backend executed preprocessing routines identical to those applied during model training, including feature generation from the spatial layers and scaling transformations, ensuring consistency between the training and operational datasets.

Once the input data was standardized, the .joblib model was loaded to generate theft predictions for each record, returning probability scores and binary classifications. These predictions were automatically stored in the PostGIS database as a new table, with spatial geometry fields allowing them to be directly visualized on the map. The WebGIS front-end, developed with the Leaflet JavaScript library, retrieved and displayed the predictions via PostgREST endpoints. Symbology was applied to distinguish predicted theft cases from normal connections, with interactive layer controls allowing users to toggle between infrastructure layers, KDE hotspot maps, and model outputs. The dashboard also included filtering tools, enabling stakeholders to view predictions for specific wards, time periods, or infrastructure types.

This integration allowed stakeholders not only to view theft-prone areas in a spatial context but also to analyze them alongside relevant infrastructure, demographic, and lighting layers, creating a multi-dimensional decision-making environment. By combining predictive analytics with geospatial visualization, the deployed system provided a powerful operational tool for monitoring electricity theft risk, planning field inspections, and allocating resources for preventive interventions. The modular design ensured that both the predictive model and the WebGIS components could be updated independently, supporting scalability and long-term adaptability.

3.5. 10 Operationalization of the Predictive Theft Monitoring System

The complete methodology for Objective 3 culminated in the development of an operational predictive theft monitoring system, integrating spatial analytics, machine learning, and web-based mapping into a single functional platform. The process began with the systematic acquisition and preprocessing of spatial and non-spatial datasets, harmonized under a common coordinate reference system to ensure interoperability. Spatial feature extraction through kernel density estimation, zonal statistics, and proximity analysis produced a robust meterbox-level dataset enriched with demographic, lighting, infrastructural, and anomaly-derived variables. These features formed the basis for training an ensemble machine learning model combining Random Forest, LightGBM, and XGBoost algorithms, which was optimized through hyperparameter tuning and validated to achieve balanced performance across accuracy, precision, recall, and AUC-ROC metrics.

The calibrated model was deployed within a Flask–PostGIS–Leaflet WebGIS architecture, enabling users to upload new billing datasets, process them through the same analytical pipeline, and visualize the resulting predictions directly on an interactive map. The system's interface allowed users to explore predicted theft hotspots alongside supporting spatial layers, filter results by geographic or infrastructural attributes, and export outputs for further analysis. By embedding advanced predictive analytics into an accessible geospatial environment, the methodology not only addressed the research objective but also delivered a practical decision-support tool. This operationalization ensures that the analytical framework developed in the study can be applied in real time, providing users with actionable insights to guide field inspections, resource allocation, and strategic interventions against electricity theft.

CHAPTER 4: RESULTS AND DISCUSSIONS

4.1 Spatial Analysis of Variables Influencing Electricity Theft

Electricity theft in Kisauni Sub-County is shaped by both socio-economic and infrastructural factors whose spatial patterns provide important clues for detection. The distribution of meterboxes reflects the allocation of legal connections, and their clustering often coincides with reported anomalies. Population density indicates areas of high demand, with informal settlements showing elevated risk of illegal tapping. Building density highlights mismatches between physical development and recorded consumption, while night-time light intensity captures unbilled usage and unusual illumination patterns. Proximity to secondary substations influences monitoring effectiveness and exposure to theft opportunities. Monthly anomaly trends reveal temporal variations, while correlation analysis uncovers interdependencies among the variables. Together, these factors create a spatial framework for identifying theft-prone zones and guiding predictive modeling.

4.1.1 Meterbox Distribution and Theft Anomalies

The distribution of meterboxes is a key operational parameter for electricity theft detection, as it directly reflects the spatial allocation of registered electricity connections. In this study, meterbox location data was sourced from KPLC's georeferenced operational database and overlaid with billing and anomaly records for the 24-month period from January 2023 to December 2024. The resulting dataset included both legitimate and anomalous meterboxes (Figure 4.1), with the latter identified through reported theft incidents, irregular consumption patterns, or field inspection findings.

Spatial analysis revealed that meterboxes are unevenly distributed across Kisauni Sub County, with higher concentrations in densely populated residential neighborhoods and commercial zones. Peripheral and semi-rural areas exhibited sparser coverage, often following the alignment of LV distribution lines. Notably, many recorded theft anomalies were clustered within high-density meterbox zones, suggesting that densely serviced areas may be more prone to theft due to both higher consumption opportunities and the challenge of monitoring numerous connections.

A kernel density estimation (KDE) was performed on the anomaly points to visualize spatial clustering (Figure 4.2). The KDE surface highlights multiple hotspots of theft activity, predominantly in urban and peri-urban neighborhoods, while rural zones generally showed lower anomaly densities. In some cases, anomalies were also detected in sparsely metered areas, which may point to illegal connections bypassing official registration.

The spatial relationship between meterbox density and theft anomalies supports the hypothesis that operational monitoring resources may need to be concentrated in areas of both high connection density and historical theft prevalence. This relationship also justifies the inclusion of meterbox spatial variables in the predictive model for theft hotspot identification.

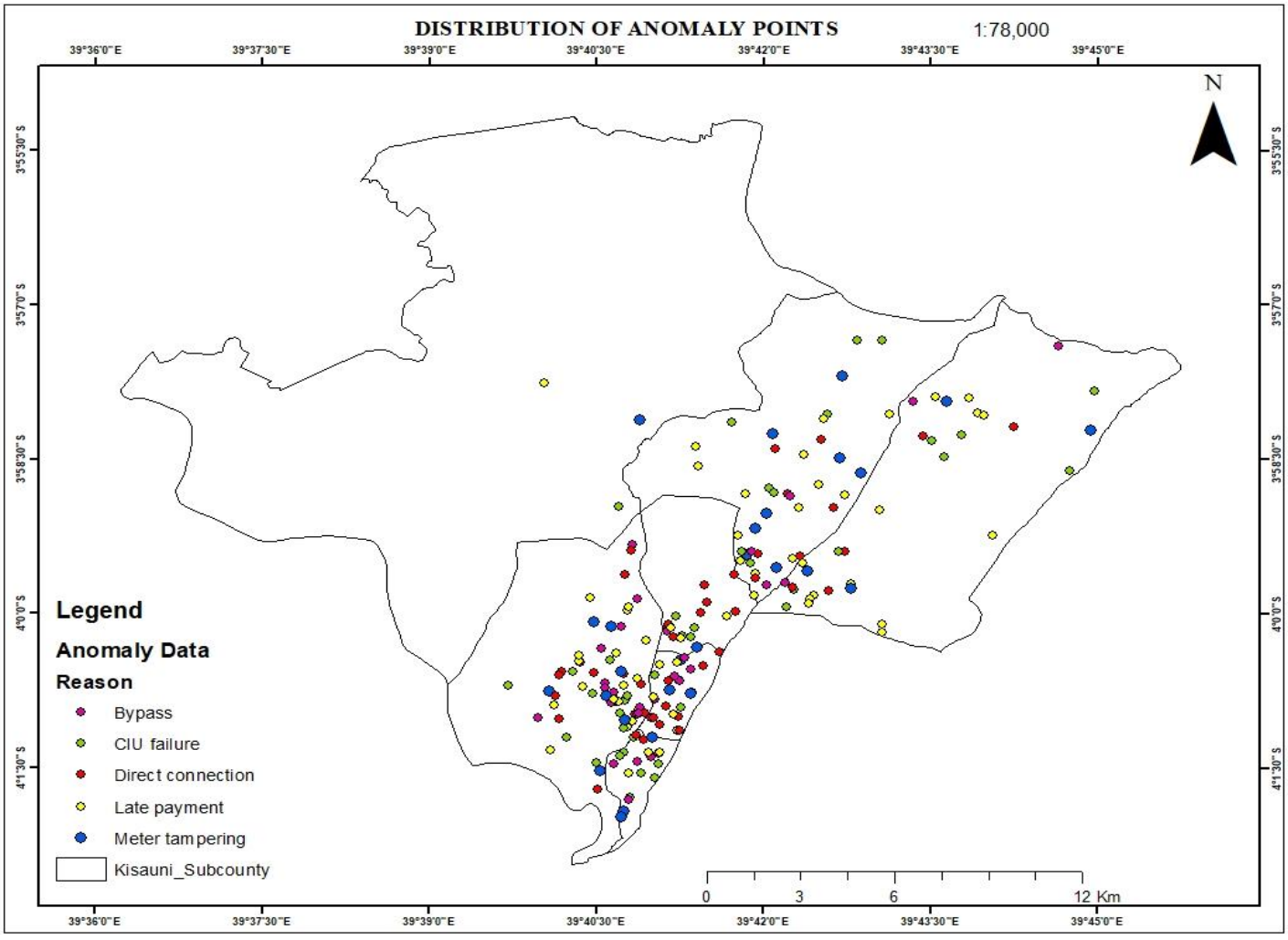


Figure 9.1 *Distribution of theft anomaly points in Kisauni Sub County*

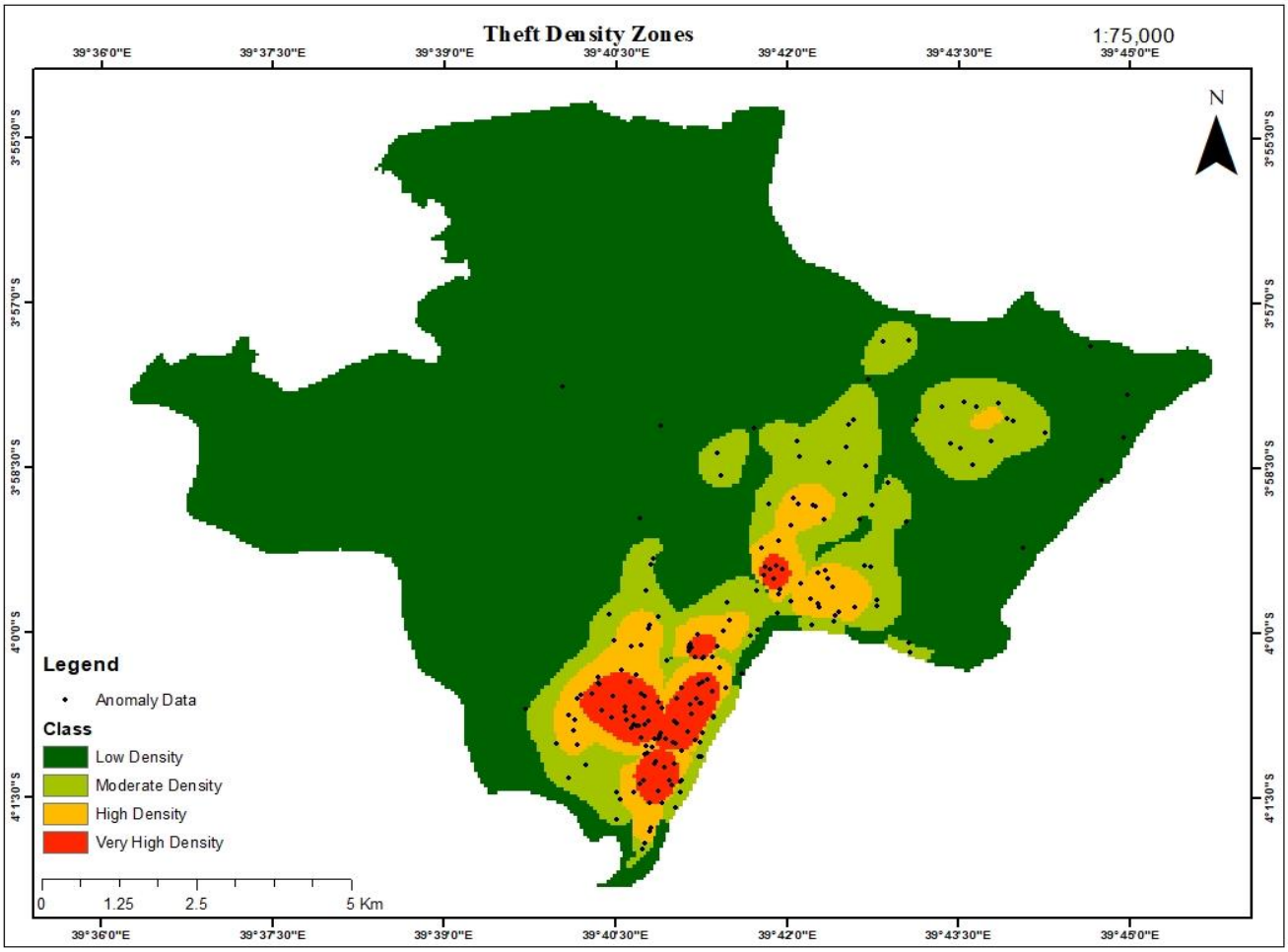


Figure 10.2 *Theft density map*

4.1.2 Population Distribution

Population distribution in Kisauni Sub County is an essential socio-economic factor influencing electricity demand patterns, infrastructure planning, and the spatial distribution of potential theft hotspots. For this analysis, population data for 2023 and 2024 were sourced from the WorldPop project at a spatial resolution of 100 meters, projected to the Arc 1960 UTM Zone 37S coordinate system. The data were aggregated to a uniform 200 m × 200 m fishnet grid, allowing seamless integration with other spatial variables such as building density, night-time light intensity, and proximity to substations. Zonal statistics were computed to derive mean population values per cell, ensuring that fine-scale demographic variations were captured without compromising compatibility across datasets. The resulting distributions are shown in Figures 4.3 and 4.4.

The results reveal pronounced heterogeneity in population density across the seven wards of Kisauni. Bamburi and Shanzu stand out as high-density zones, driven by rapid urbanization, established residential estates, hospitality and tourism facilities, and concentrated commercial activity. Mtopanga and Junda also record elevated densities due to their proximity to major transport corridors and growing middle-income housing developments. In contrast, Mwakirunge and Magogoni remain predominantly low-density, characterized by open land, peri-urban agricultural activity, and dispersed rural-style homesteads. Mjambere presents a more mixed pattern, with moderate density in formal estates alongside pockets of high-density informal settlements.

Informal settlements in Junda, Magogoni, and Mjambere highlight areas where population growth has outpaced formal electricity infrastructure, leading to illegal connections, meter bypassing, and unmetered consumption. Conversely, Mwakirunge's low population density provides a valuable baseline for distinguishing genuine low-consumption areas from those with suppressed night-time light emissions due to theft. When analyzed alongside building density and night-time light anomalies, population distribution becomes a powerful spatial predictor of theft risk, enabling targeted intervention strategies.

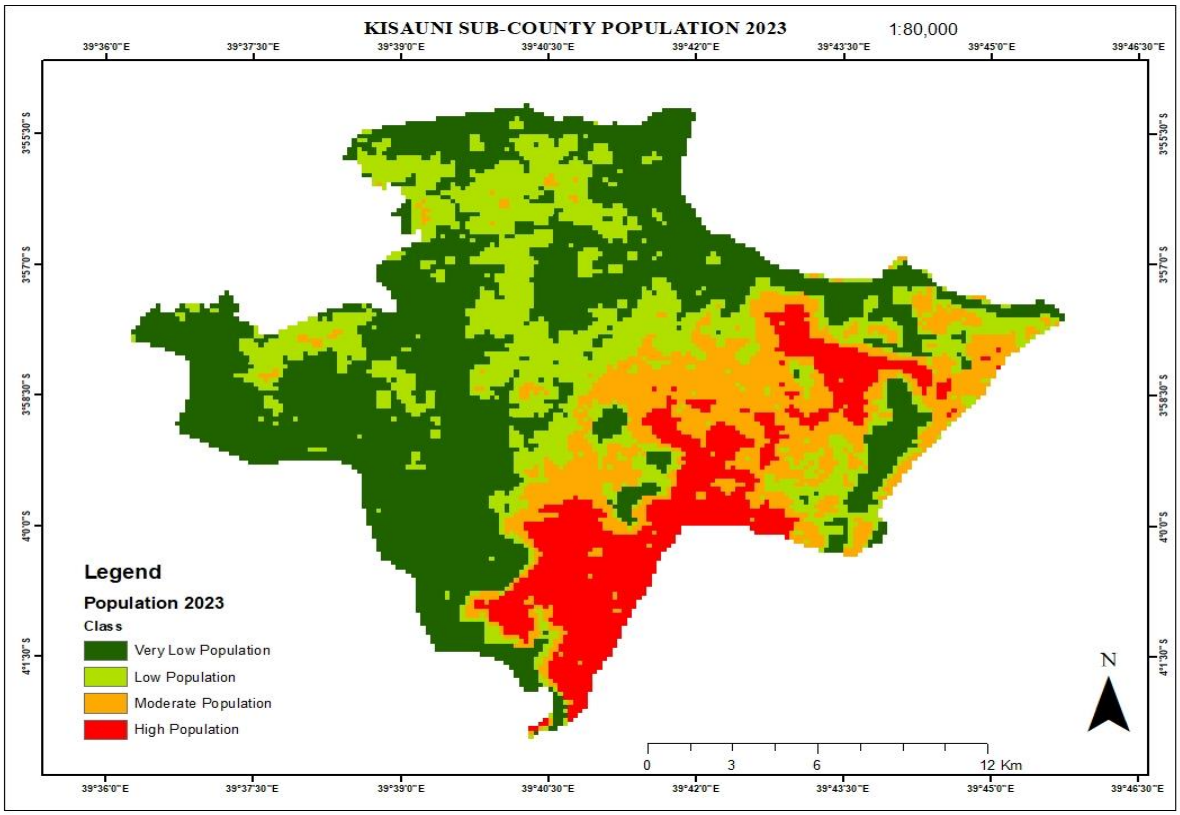


Figure 11.3 *Population Density In 2023*

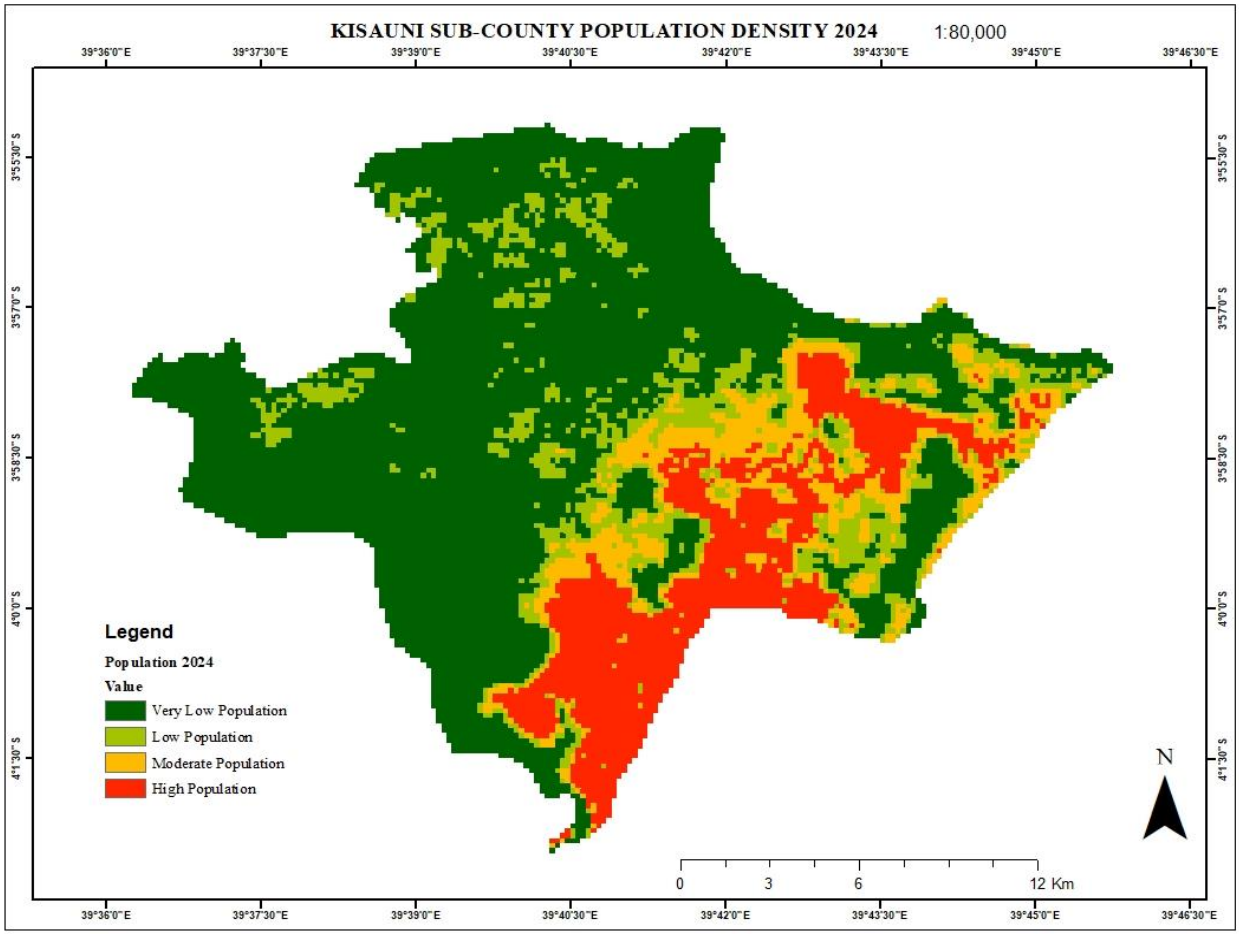


Figure 12.4 *Population Density In 2024*

4.1.3 Building Density

Building density serves as a proxy for physical development intensity and is closely linked to electricity demand, infrastructure coverage, and potential non-technical losses. For this study, building density data were obtained from the WorldPop global building footprint dataset (Figure 4.5), derived from high-resolution satellite imagery and advanced object-based classification techniques.

The spatial patterns show that Bamburi, Mtopanga, and Shanzu exhibit the highest building density values, reflecting their established residential estates, mixed-use developments, and commercial hubs. These areas feature closely spaced permanent structures, planned road networks, and well-

developed service corridors, making them relatively easier to monitor and manage in terms of electricity metering. In contrast, Magogoni and Mwakirunge display much lower building densities, with large tracts of open land interspersed with scattered homesteads or small clusters of housing. Junda and Mjambere show a more complex pattern, with certain sections having high densities due to informal settlement clusters, while others remain less developed.

When analyzed alongside night-time light anomalies, building density becomes a critical indicator for theft hotspot detection. For example, areas with high building density but unusually low night-time light intensity may indicate widespread illegal tapping, meter bypassing, or complete non-registration of electricity consumers. Conversely, zones with low building density yet abnormally high light emissions may point to concentrated theft activities or unregulated commercial operations. By integrating building density data with other socio-economic and infrastructural variables, the predictive modelling framework is better able to isolate unusual patterns that merit further investigation.

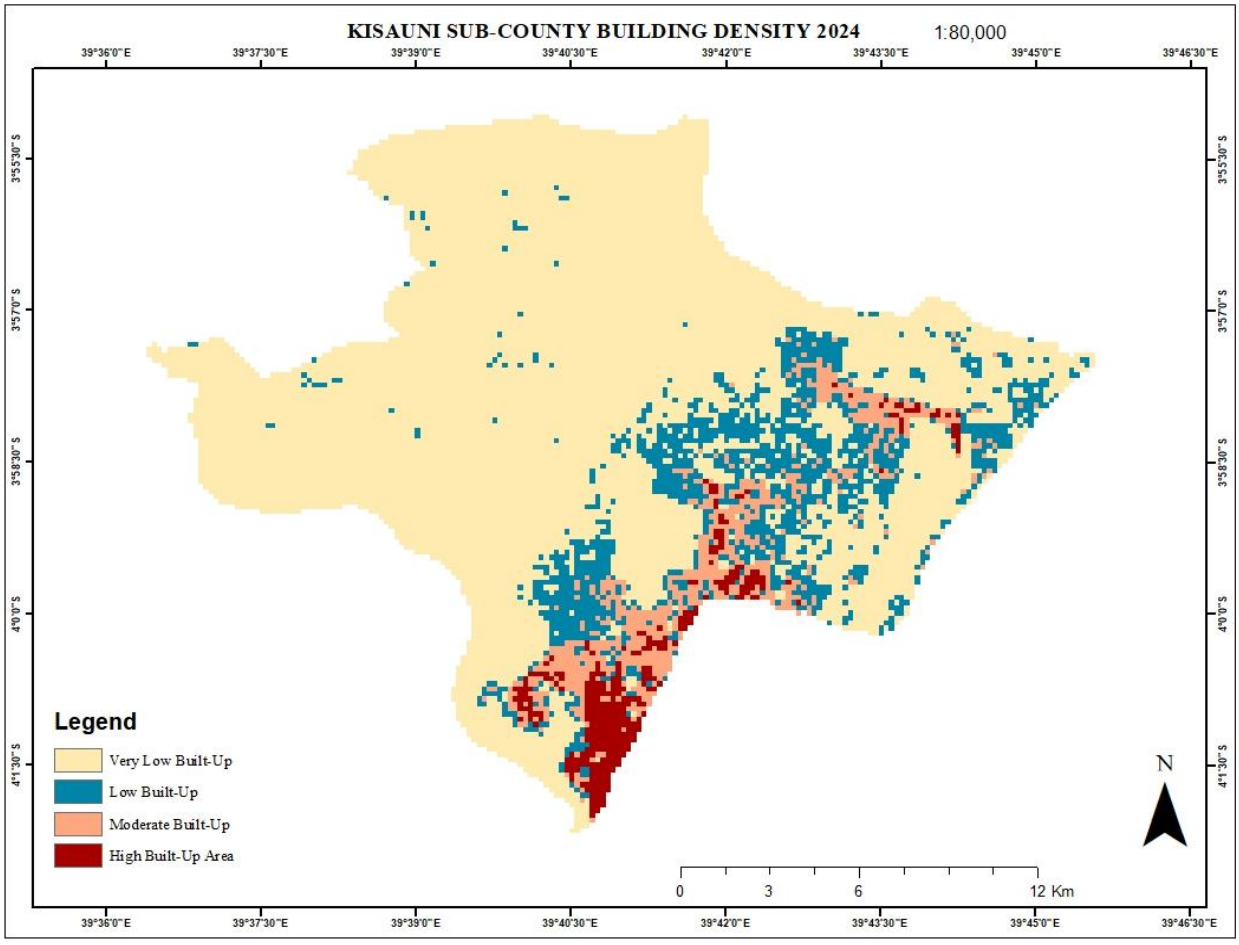


Figure 13.5 *Building density map*

4.1.4 Night-time Light Intensity

Night-time light (NTL) intensity provides an indirect yet powerful measure of electricity consumption, economic activity, and patterns of urbanization. For this study, annual VIIRS-DNB composites for 2023 and 2024 were used, with preprocessing steps applied to filter out transient light sources such as fires, vehicle headlights, and offshore fishing vessels (Figure 4.6). The resulting stable light datasets highlight the persistent illumination associated with residential, commercial, and industrial activity, offering a spatially continuous view of energy use across Kisauni.

The spatial distribution of NTL in Kisauni reveals pronounced contrasts between wards. Bamburi and Shanzu exhibit the highest intensities, reflecting their concentration of commercial hubs, high-end residential areas, and hospitality establishments that operate late into the night. Mtopanga shows moderately high levels, consistent with its growing mixed-use development. In Junda and Mjambere, illumination is uneven bright clusters coincide with active trading centers, while dimmer pockets suggest possible under-electrification or unmetered usage. Magogoni and Mwakirunge remain the least illuminated, but with occasional bright anomalies that may correspond to unregistered industrial activity or informal grid extensions.

When compared with socio-economic and infrastructure data, several anomalies emerge. Areas with dense building coverage yet consistently low NTL may point to widespread power theft, meter bypassing, or reliance on alternative energy sources. Conversely, sparsely populated areas showing disproportionately high NTL often align with activities that operate outside standard residential consumption patterns, potentially indicating illegal connections or unregulated commercial operations. These insights position NTL analysis as a critical diagnostic tool in understanding the spatial dynamics of electricity usage and targeting investigative or enforcement interventions.

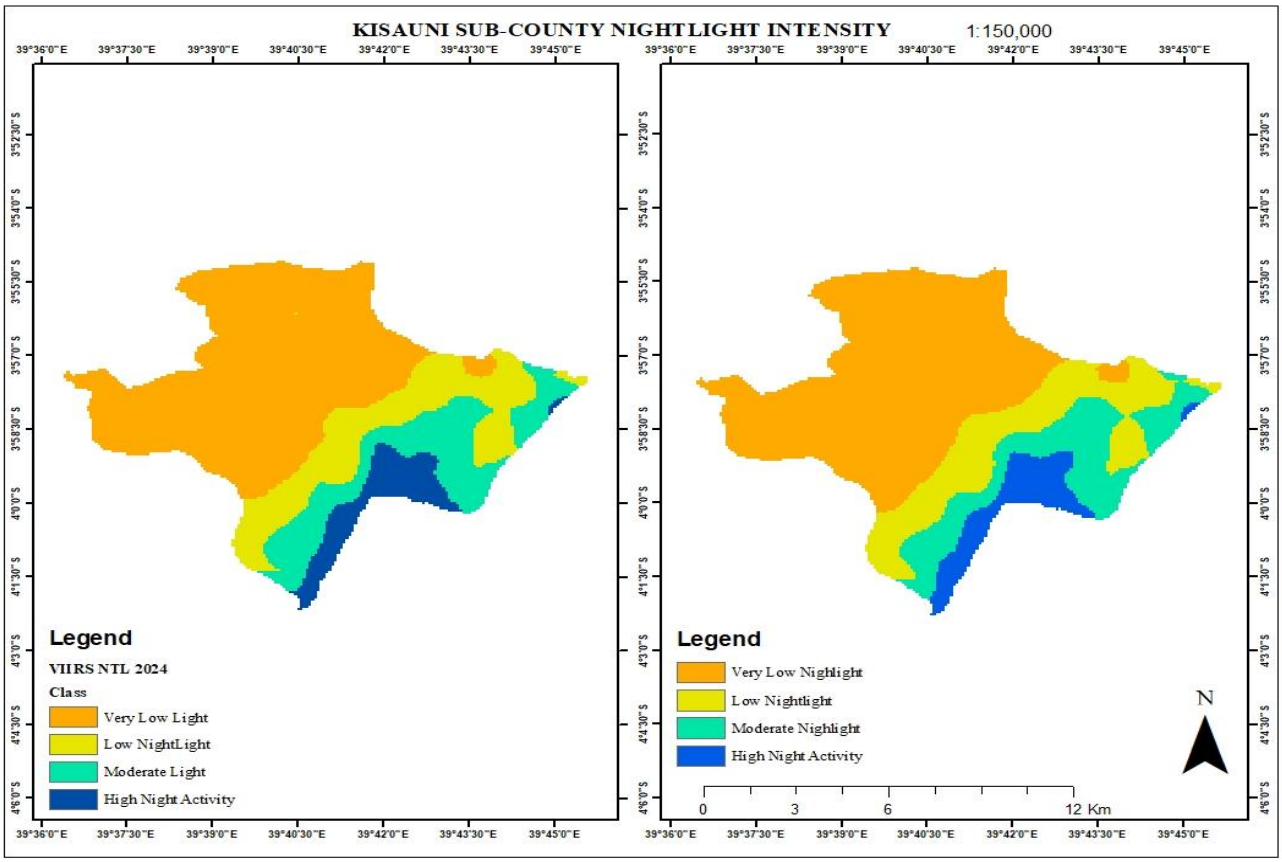


Figure 14.6 *Nightlight Map for 2023 and 2024*

4.1.5 Secondary Substation Coverage and Proximity

Secondary substations serve as critical nodes in Kisauni's electricity distribution network, stepping down medium-voltage supply to low-voltage service for residential and commercial consumers (Figure 4.7). Their spatial placement affects both service reliability and potential exposure to electricity theft. Proximity analysis using Euclidean distance calculations revealed that central wards such as Mtopanga and Shanzu are well-covered, while peripheral zones, including Magogoni and Mwakirunge, lie beyond 500 meters from the nearest substation. Overlaying these proximity buffers with building density and night-time light (NTL) data highlights areas where high structural density coincides with low illumination, suggesting potential unmetered connections or meter bypassing.

The spatial distribution of theft anomalies shows a higher concentration in areas farther from substations, reinforcing the role of network accessibility in non-technical losses. Zones close to substations generally exhibit lower anomaly density, indicating that frequent maintenance,

monitoring, and shorter distribution paths may reduce unauthorized consumption. These results highlight the importance of strategic substation placement and targeted monitoring in mitigating electricity theft, providing actionable insights for utility operators.

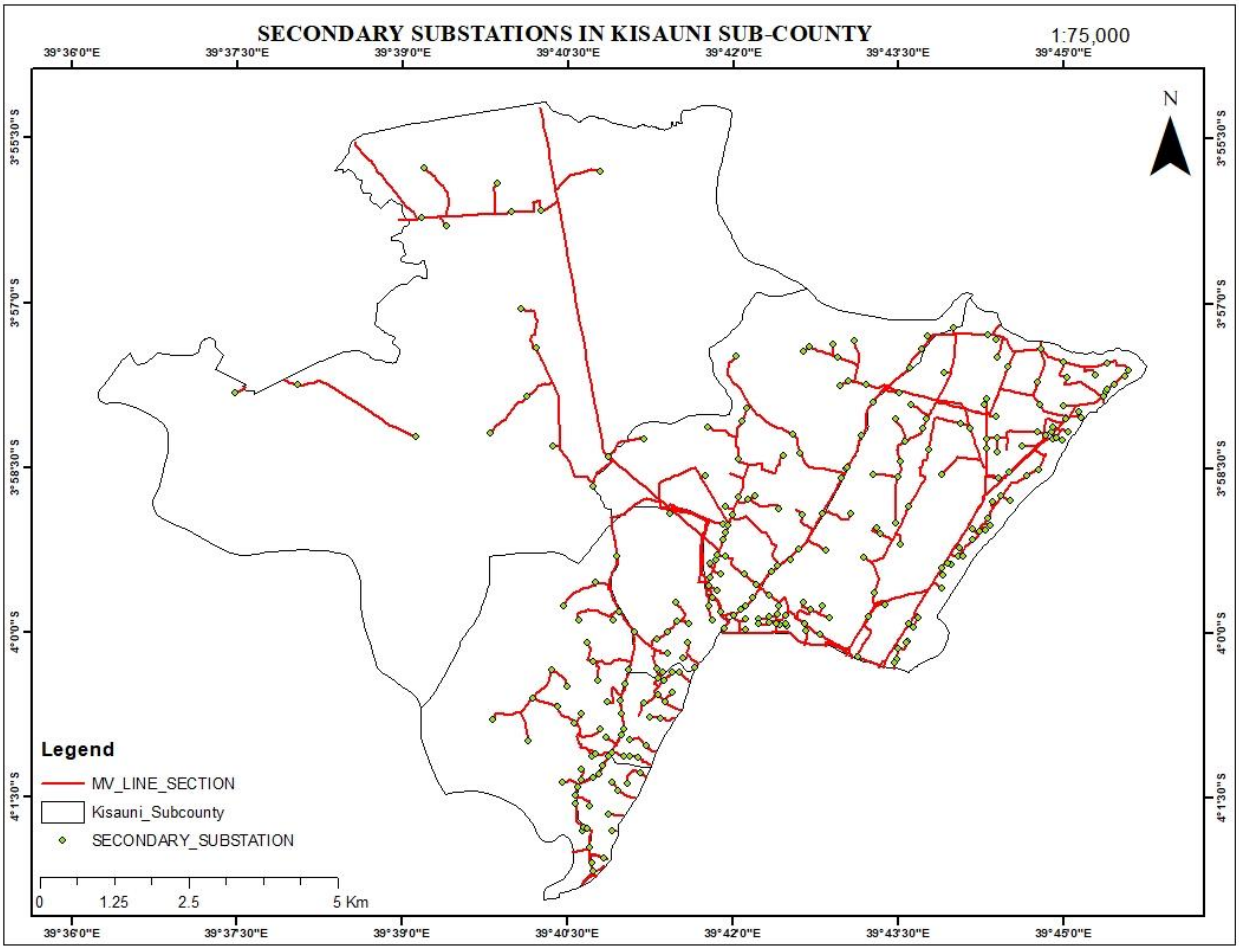


Figure 15.7 *Secondary substation Coverage in Kisauni Sub County*

4.1.6 Theft Anomaly Trends by Month

Temporal analysis of electricity theft anomalies provides insight into patterns of unauthorized consumption over time and helps identify periods of heightened risk. Using meter-level anomaly records from January 2023 to December 2024, monthly counts of theft incidents were aggregated and mapped to reveal temporal fluctuations (Table 4.1). The results indicate that anomalies are not uniformly distributed throughout the year. Peaks were observed during the onset of the long dry

season and festive periods, suggesting that increased household and commercial electricity demand may trigger higher instances of meter tampering, bypassing, or illegal connections.

Spatially, these temporal trends reinforce previously observed high-risk zones identified through KDE of anomalies and substation proximity analysis. For instance, wards with high structural density but limited substation access consistently recorded above-average monthly anomalies, indicating persistent vulnerability. Conversely, well-served central wards generally exhibited lower monthly theft rates, demonstrating the influence of network accessibility and operational monitoring on theft occurrence. These insights underscore the value of integrating temporal trends into operational planning, enabling utility operators to schedule proactive inspections, allocate resources effectively during peak risk periods, and design timely intervention strategies.

Table 4.6 *Monthly count of theft incidences over 2023 and 2024*

Theft_Month	Count
Jan	12
Feb	15
Mar	15
Apr	7
May	27
Jun	17
Jul	16
Aug	20
Sep	18
Oct	16
Nov	18
Dec	17
Grand Total	198

4.1.7 Correlation Between Variables

Understanding the relationships between spatial and socio-economic variables is essential for identifying the underlying drivers of electricity theft in Kisauni Sub County. A correlation analysis was conducted using the fishnet grid dataset, incorporating mean values of building density, night-time light (NTL) intensity, population, and proximity to secondary substations, alongside meterbox counts and recorded theft anomalies. Pearson's correlation coefficients were calculated

and visualized in a heatmap to quantify the strength and direction of pairwise relationships among these variables.

The results indicate several notable patterns. Building density exhibits a moderate positive correlation with population density, reflecting the expected spatial co-location of residents and structures. NTL intensity is strongly correlated with building density in urbanized wards such as Bamburi and Shanzu, suggesting that higher structural concentration aligns with increased electricity usage. However, when NTL is compared to population in certain peripheral zones, the correlation weakens, highlighting areas where illumination levels may not correspond to officially recorded residents, potentially indicating unmetered consumption or illegal connections. Proximity to secondary substations shows a negative correlation with anomaly density, confirming that areas located farther from substations are more susceptible to irregular power usage. Similarly, KDE-based anomaly density correlates positively with building density in informal settlements, underscoring the role of high-density, under-served areas as theft hotspots.

These correlations collectively reveal the complex interplay between infrastructure, socio-demographic factors, and electricity theft. High building and population densities alone do not guarantee accurate consumption reporting; rather, the combination of structural concentration, illumination patterns, and substation access provides a more comprehensive predictor of theft risk. Such insights validate the choice of these variables as key features in the predictive modelling framework and emphasize the need for multi-layered strategies—combining spatial planning, monitoring, and targeted inspections—to mitigate non-technical losses effectively.

4.2 Predicted Electricity Theft Hotspots

The ensemble model was applied to generate predictions of electricity theft across Kisauni Sub County. These predictions were spatially referenced at the meterbox level, enabling the identification of areas with high likelihoods of anomalous or illegal electricity consumption. The output labels from the ensemble model—derived from billing patterns, spatial coordinates, and socio-economic variables—were translated into a classification of risk zones. Three risk classes were defined: 0 for low risk (normal consumption), 1 for medium risk (possible anomalies), and 2 for high risk (likely confirmed theft). A comparison between the predicted risk zones and reported theft cases shows that high-risk areas correspond closely with locations of historically confirmed

theft, while low-risk areas align with regions exhibiting normal consumption. This correspondence demonstrates the model's predictive accuracy and provides utility operators with a structured framework for prioritizing monitoring, enforcement, and targeted interventions (Figure 4.8).

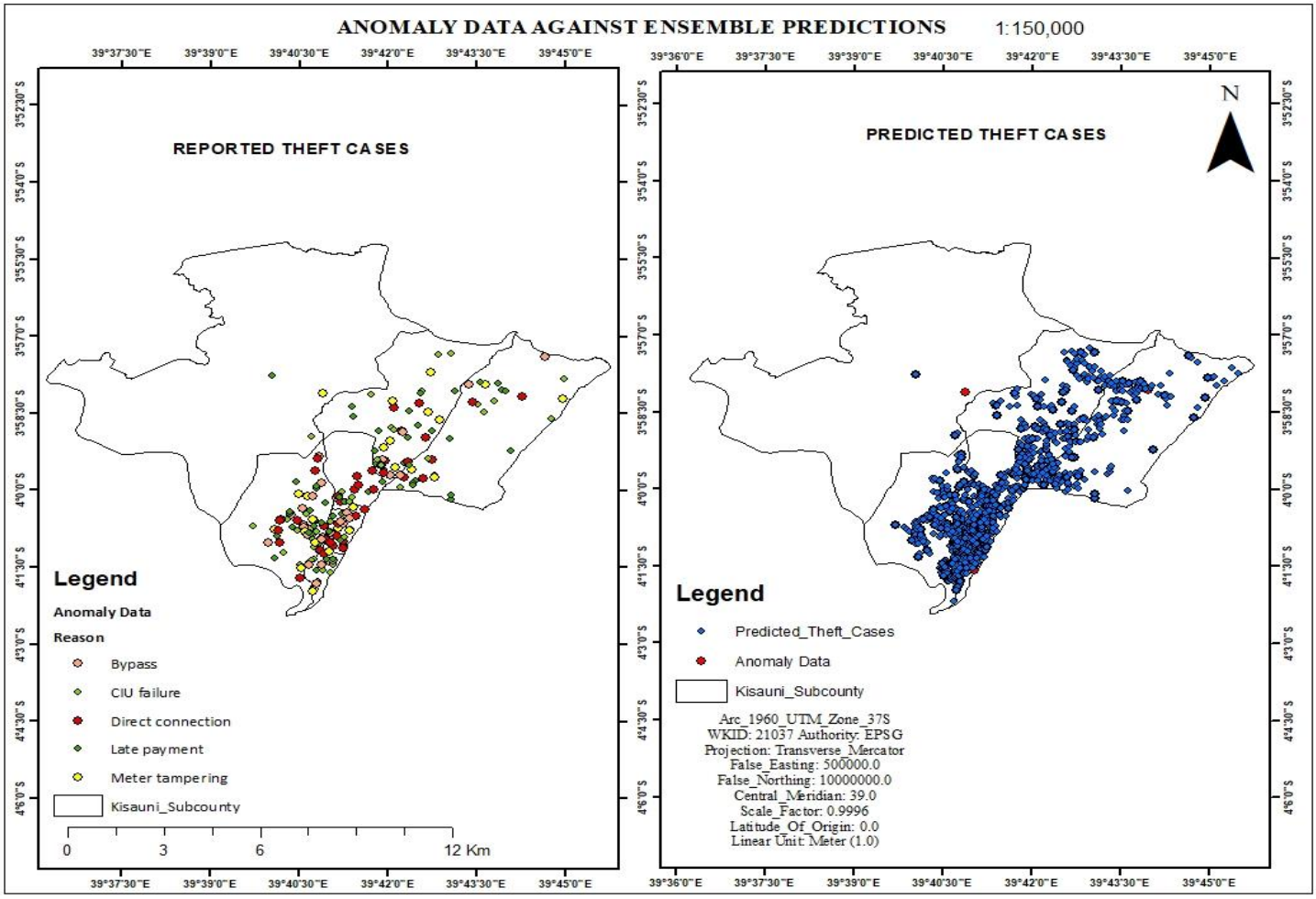


Figure 16.8 *Comparison between reported and predicted theft cases*

4.2.1 Spatial Distribution of Predicted Hotspots

The mapped predictions demonstrate that electricity theft is not uniformly distributed across Kisauni (Figure 4.9). Instead, spatial clustering is observed, with contiguous blocks of high-risk cells indicating concentrated areas of concern. These clusters often correspond to areas where night-time light intensity is unusually high relative to population density or where billing patterns exhibit large variability, confirming the model's capacity to integrate multiple indicators for hotspot detection. The spatial perspective facilitates visual identification of priority areas for field inspections, meter audits, and community engagement, making the predictions operationally actionable.

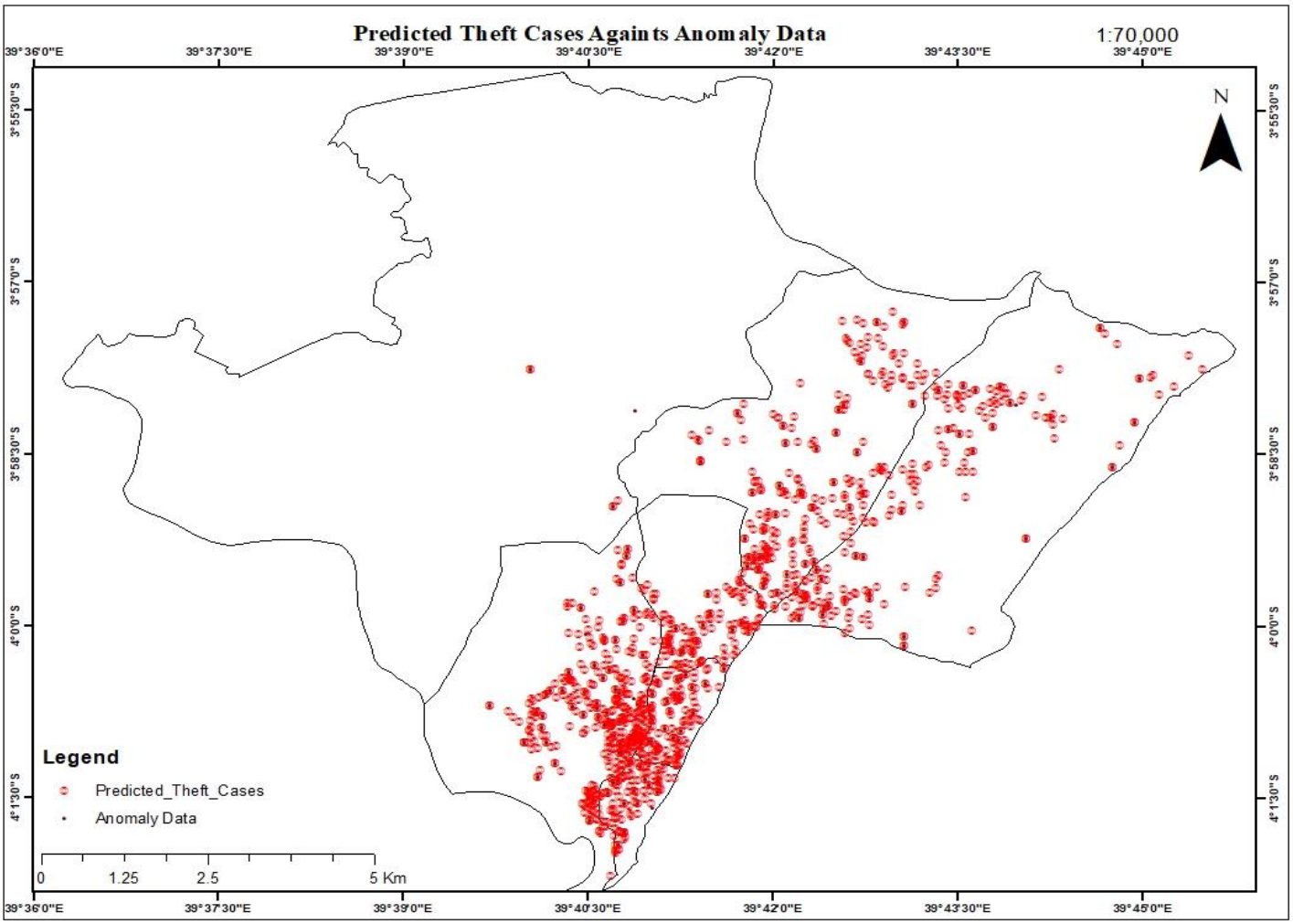


Figure 17.9 *concentric rings as the predicted theft cases while anomaly data as points*

4.2.2 Statistical Summary of Predicted Hotspots

A statistical summary of predicted hotspots further quantifies the extent and distribution of electricity theft risk (Table 4.2). In total, the ensemble model classified a small but significant proportion of meterboxes as high-risk, consistent with the known rarity of confirmed anomalies relative to normal consumption. Medium-risk zones represent areas with moderate likelihoods of theft, warranting monitoring but not immediate intervention, while low-risk zones encompass the majority of the network. By ward, high-risk hotspots are most prevalent in informal and densely settled neighborhoods, whereas wards with formal infrastructure exhibit predominantly low-risk classifications. These statistics support data-driven allocation of resources, ensuring that inspection and enforcement efforts are focused where they are most needed.

The combination of spatial maps and tabular summaries provides a comprehensive picture of electricity theft risk in Kisauni. By integrating predictive modeling with operationally meaningful classifications, utility operators are equipped with actionable intelligence for proactive management, targeted inspections, and strategic planning to reduce non-technical losses and improve service reliability.

Table 4.7 *Statistical summary of predicted theft hotspots*

Ward Name	Total Meterboxes	Low Risk	Medium Risk	High Risk	Reported Theft Cases
Bamburi	10,317	9,255	750	189	46
Junda	8,243	7,615	489	262	26
Magogoni	5,126	4,879	181	64	9
Mwakirunge	4,562	4,345	162	57	7
Mtopanga	8,711	8,074	477	160	31
Shanzu	9,036	8,380	515	141	29
Mjambere	7,845	7,252	586	123	50
Total	**53,840**	**49,800**	**3,160**	**1,186**	**198**

4.3 WebGIS Integration and Visualization of Results

The predicted electricity theft hotspots generated by the ensemble model were integrated into a Web-based Geographic Information System (WebGIS) to facilitate interactive visualization, exploration, and operational decision-making. The WebGIS platform combines spatial mapping with analytical tools, allowing users to interpret model predictions, explore underlying data, and perform spatial queries without requiring advanced GIS software. All predicted hotspots, meterboxes, infrastructure layers, and socio-economic data were stored in a PostgreSQL/PostGIS database and served through a Python-based Flask application, enabling real-time interaction and dynamic map rendering on the dashboard.

4.3.1 Dashboard Map and Layer Visualization

The WebGIS dashboard provides a dynamic and interactive map interface that visualizes predicted electricity theft hotspots alongside key infrastructure and socio-economic layers. This visual differentiation allows utility operators and analysts to quickly identify areas requiring immediate attention.

Supporting layers, including meterbox locations, secondary substations, low voltage and medium voltage lines, and building density, can be toggled on or off, enabling users to examine the spatial context of each hotspot (Figure 4.10). The integration of multiple layers facilitates a holistic understanding of how electricity theft patterns relate to population clusters, settlement structures, and network infrastructure.

The map interface includes interactive zooming and panning, allowing users to explore the study area at various scales. A “zoom to hotspot” feature automatically centers the map on the highest-risk areas, ensuring that critical zones are immediately visible for operational decision-making. This functionality is particularly useful for prioritizing inspections, planning field visits, and allocating resources efficiently.

By combining predictive model outputs with interactive GIS layers, the dashboard transforms complex data into an intuitive, user-friendly platform. It allows operators not only to visualize theft

risk spatially but also to contextualize it relative to infrastructure and settlement patterns, enhancing situational awareness and supporting informed management of the electricity network.

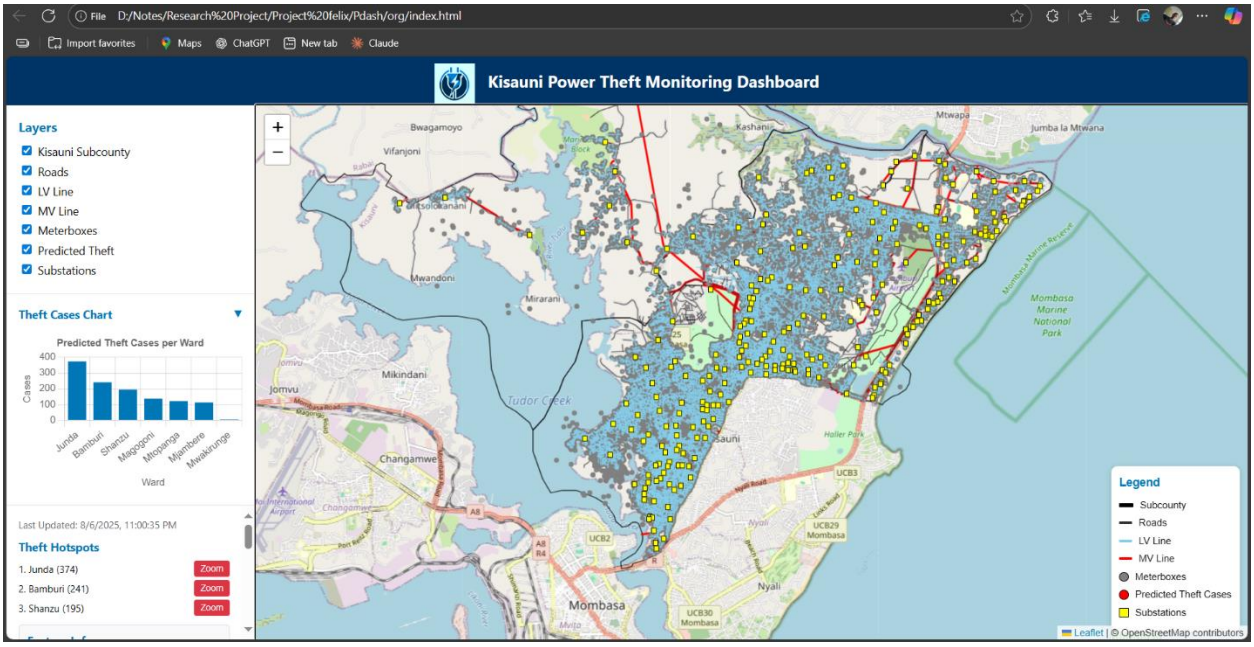


Figure 18.10 *overview of the monitoring dashboard*

4.3.2 Interactive Meterbox Information

The WebGIS dashboard allows users to interact directly with individual meterboxes, providing detailed operational information that links predicted theft risk to real-world assets. By clicking on any meterbox displayed on the map, a pop-up window appears containing key attributes, including the (meter serial number--customer name--account number). This functionality enables rapid verification and investigation of potential anomalies, connecting model outputs to actionable operational data (Figure 4.11).

This interactive feature is particularly valuable for field teams and utility operators, as it allows immediate identification of specific connections associated with high-risk classifications. For example, in high-risk zones identified by the ensemble model, operators can use the pop-up information to prioritize inspections or follow up with customers whose consumption patterns are flagged as anomalous. Medium-risk and low-risk zones can be similarly assessed, enabling resource allocation to focus on areas with the greatest likelihood of illegal activity.

The pop-up interface also integrates visual cues such as colored markers corresponding to the theft risk class, providing an at-a-glance understanding of which meterboxes require urgent attention. Users can scroll through multiple meterboxes in dense areas, ensuring that no anomaly is overlooked. These features transform the WebGIS from a static mapping tool into an operational platform for monitoring, auditing, and targeted enforcement.

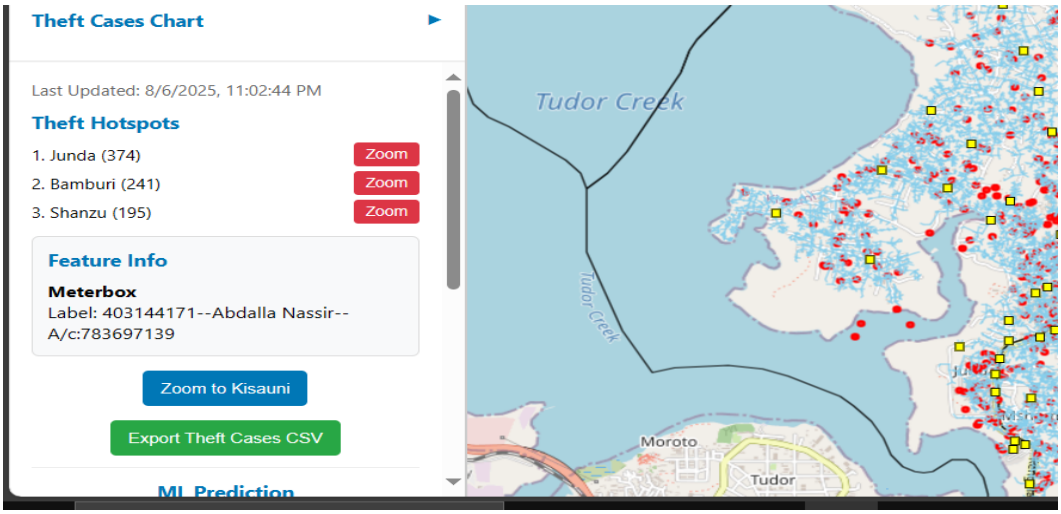


Figure 19.11 *display showing meter serial number, customer name, and account number for a selected meterbox.*

4.3.3 Analytical and Measurement Tools

The WebGIS dashboard includes a suite of analytical and measurement tools designed to enhance spatial analysis and support operational decision-making. A primary feature is the **buffer tool**, which allows users to select a point on the map and create a surrounding area of a specified radius. All meterboxes, substations, or infrastructure features within this buffer are automatically highlighted and their attributes displayed, facilitating localized investigations and targeted inspections (Figure 4.12).

Another key functionality is the **distance measurement tool**, which enables users to measure straight-line distances between two or more points, or along a drawn path. This tool is particularly useful for planning field operations, estimating the proximity of suspected theft clusters to secondary substations, and assessing accessibility for maintenance or inspection teams.

Measurements can be displayed in meters or kilometers, providing precise, actionable spatial information.

The dashboard also supports **map export and printing**, allowing users to generate static maps of selected areas, including the currently visible layers and highlighted hotspots (Figure 4.13). This capability ensures that spatial analyses, inspection plans, or operational reports can be shared with teams or management, or retained for record-keeping purposes. Users can customize the exported map's extent, scale, and layout to suit different reporting needs.

Together, these tools transform the WebGIS from a visual display platform into a functional spatial analysis environment, enabling users to investigate electricity theft patterns, assess infrastructure coverage, and plan field operations efficiently.

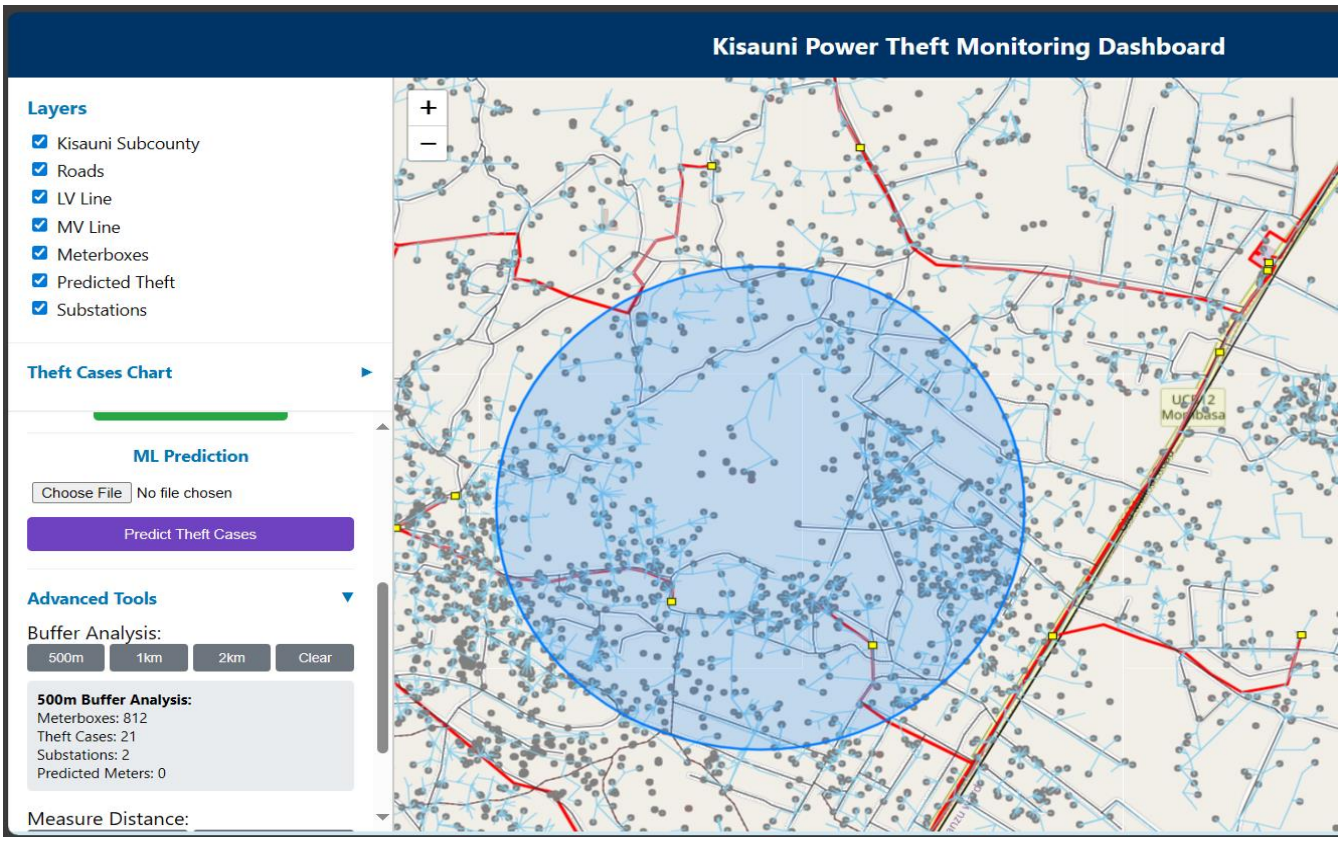


Figure 20.12 *Buffer tool highlighting all meterboxes and infrastructure features within a user-defined radius.*

Figure 21.13 *the print map tool highlighting all meterboxes and infrastructure features within the window and exporting as a map.*

4.3.4 Integrated Machine Learning Prediction Tool

The WebGIS dashboard incorporates an integrated machine learning prediction tool that allows users to run real-time electricity theft risk predictions using the trained ensemble model (Figure 4.14). To ensure consistency and accurate prediction, the uploaded data must match the format used during model training. This includes operational attributes such as monthly billing records from January 2023 to December 2024 for each meterbox, spatial coordinates (POINT_X and POINT_Y), distance to the nearest secondary substation (NEAR_DIST), and socio-economic indicators including population density for 2023 and 2024 (Kenpop2023, Kenpop2024) as well as building density (B_Density). Maintaining this format ensures that all features expected by the model are present and processed correctly, preventing errors during prediction.

Once the CSV is uploaded, the Flask backend automatically applies the trained ensemble model to classify each meterbox into one of three risk categories: 0 for normal usage, 1 for confirmed anomalies or theft, and 2 for moderate risk patterns that indicate potential irregularities requiring further inspection. The results are immediately visualized on the map, with color-coded markers indicating low, medium, and high-risk meterboxes. Users can interact with individual meterboxes by clicking on them to view key attributes, including meter serial number, customer name, account number, and predicted risk class. This enables targeted follow-ups and prioritization of inspections in the field.

In addition to real-time mapping, the tool allows users to export the predicted dataset in CSV format, retaining all features and appended prediction labels. This facilitates offline analysis, reporting, or integration with utility management systems. Predictions can also be combined with spatial analysis tools within the dashboard, such as the buffer tool to examine all meterboxes within a specified radius of a hotspot, or the distance measurement tool to determine proximity to substations and distribution lines.

By integrating accurate predictions with interactive visualization and operational analytics, the dashboard provides a comprehensive platform for monitoring electricity consumption patterns, identifying potential theft hotspots, and supporting evidence-based operational decision-making.

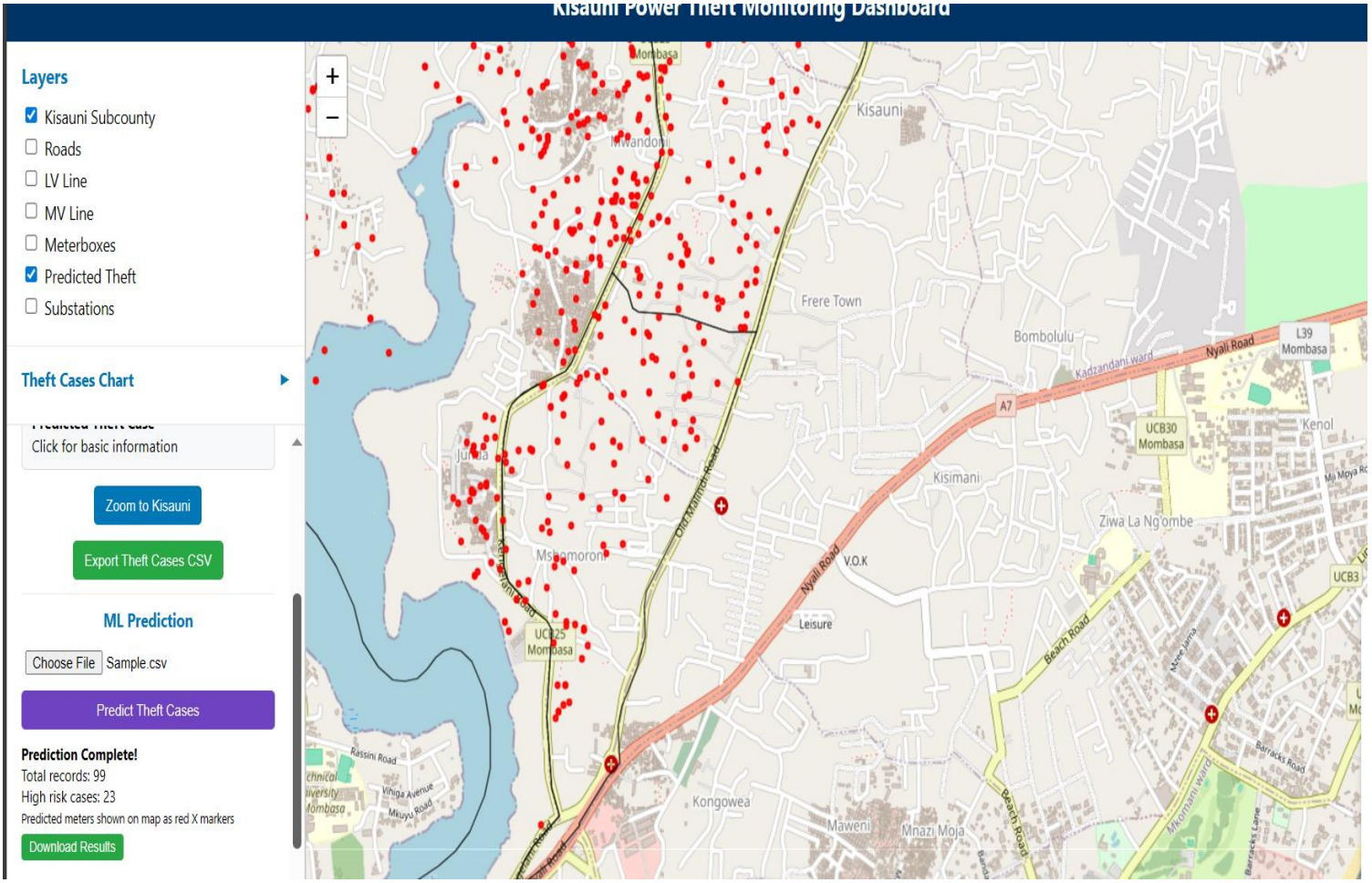


Figure 22.14 *Predicted electricity theft hotspots displayed on the WebGIS dashboard after CSV upload***.**

4.3.5 User Interaction and Auxiliary Tools

The WebGIS dashboard provides multiple interactive functionalities that allow utility operators and analysts to explore electricity consumption and theft patterns in depth. The zoom-to-hotspot feature enables users to quickly center the map on areas identified as high-risk by the ensemble machine learning model. By focusing on these predicted hotspots, users can immediately visualize surrounding meterboxes, secondary substations, and distribution lines, facilitating informed planning for field inspections and prioritization of monitoring resources.

Interactive charts complement the spatial view by providing real-time statistical summaries linked to the current map selection (Figure 4.15). Users can examine aggregated monthly or annual metrics of predicted theft risk, view the distribution of meterboxes across risk categories, or analyze trends in anomalies over time. The charts dynamically update when a ward or substation area is selected, giving stakeholders the ability to detect emerging patterns, identify areas requiring intervention, and make evidence-based decisions efficiently.

The distance measurement tool allows users to calculate straight-line distances between any two points or along user-defined paths. This functionality supports planning for inspections, evaluating accessibility to substations, and assessing the spatial relationships between meterboxes and infrastructure. Users can also generate high-resolution maps of the current view using the print and map export tool, including all selected layers, predicted hotspots, and annotations. These maps serve as operational references, reports, or presentation materials, ensuring that spatial analyses are clearly communicated to field teams and decision-makers.

The machine learning prediction tool is fully integrated into the dashboard, enabling users to upload new billing datasets in the same format as the training data. Upon upload, the system applies the ensemble model to classify each meterbox into one of three categories: normal usage, confirmed anomaly or theft, and moderate-risk patterns. Predictions are immediately visualized on the map with distinct color coding, and users can click individual meterboxes to view detailed attributes such as meter serial number, customer name, account number, and predicted risk. This seamless integration of predictive analytics with interactive visualization ensures that the dashboard functions as both a monitoring platform and a decision-support system, enabling targeted inspections, proactive management, and efficient allocation of operational resources.

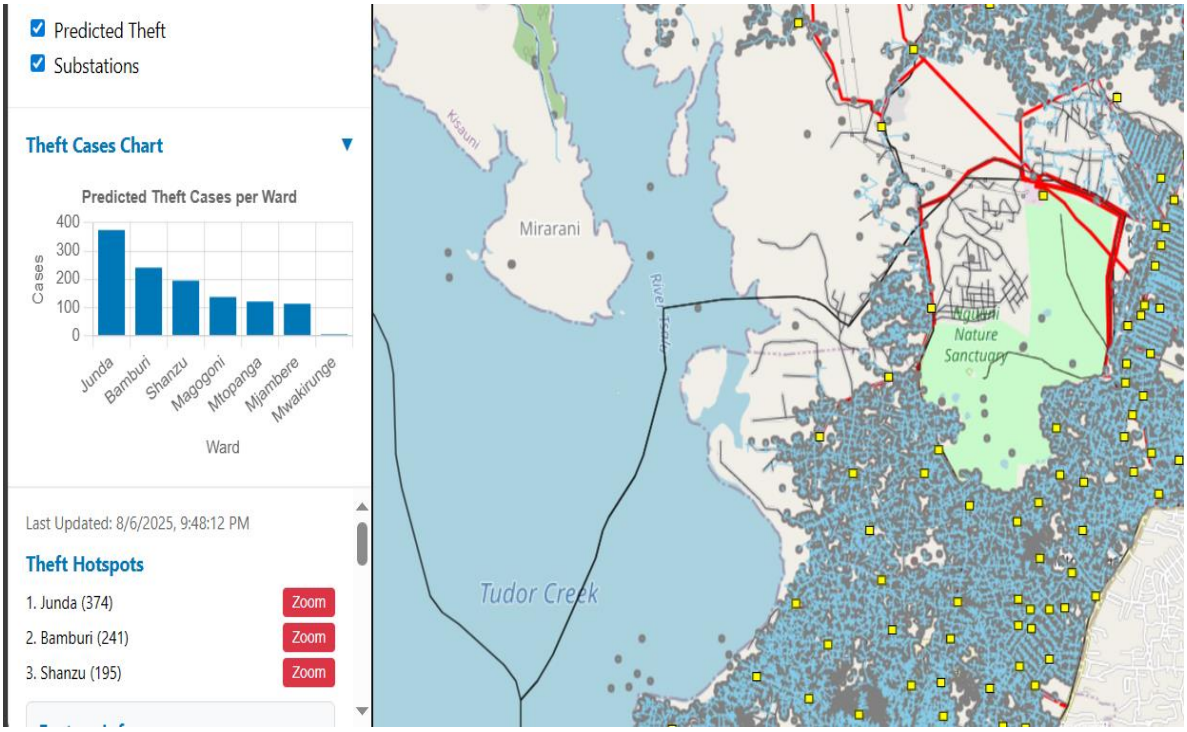


Figure 23.15 *Interactive charts and machine learning prediction results highlighting meterboxes within selected high-risk areas.*

4.4 Implications for Electricity Theft Management

The findings from the spatial analysis, machine learning predictions, and WebGIS visualization have significant implications for electricity theft management in Kisauni Sub County. The identification of high-risk hotspots enables utility operators to prioritize resources efficiently, focusing inspection, maintenance, and enforcement efforts where they are most needed. By combining predictive analytics with operational data such as meterbox locations, billing records, and infrastructure distribution, the system provides a data-driven approach to mitigating non-technical losses and improving service reliability.

The dashboard's integration of interactive tools, such as zoom-to-hotspot, meterbox attribute viewing, distance measurement, and machine learning predictions, allows for rapid situational assessment and operational planning. Field teams can navigate directly to predicted hotspots, evaluate the surrounding network, and identify potential clusters of unmetered or irregular consumption. The ability to view detailed meter attributes, including serial numbers, customer

names, and account information, supports verification and targeted intervention, reducing both the time and cost associated with blanket inspections. Furthermore, predictive modeling allows operators to anticipate emerging theft patterns, providing opportunities for proactive enforcement and preventive measures before losses escalate.

From a strategic perspective, the combined use of socio-economic, spatial, and operational data improves the understanding of theft drivers and informs long-term infrastructure planning. Areas exhibiting high building density with low night-time light or anomalous billing patterns can be prioritized for metering upgrades, enhanced monitoring, or community engagement programs to reduce illegal connections. Similarly, the analysis supports decision-making on resource allocation, such as the deployment of inspection crews or the reinforcement of substations in areas with recurrent anomalies. The dashboard also facilitates reporting and communication with management and regulatory authorities by providing visual and statistical evidence of theft patterns, contributing to more transparent and accountable operational practices.

Overall, the integration of predictive analytics, geospatial visualization, and interactive decision-support tools transforms electricity theft management from a reactive process into a proactive and data-informed system. By enabling targeted interventions, efficient resource allocation, and continuous monitoring, the approach enhances operational effectiveness, reduces non-technical losses, and supports equitable electricity service delivery across Kisauni Sub County.

CHAPTER 5: CONCLUSION AND RECOMMENDATION

5.1 Conclusions

The study successfully analyzed billing data, anomaly records, satellite imagery, and socio-economic factors to identify spatial patterns associated with electricity theft in Kisauni Sub County. Population density emerged as the strongest spatial indicator, with high-density informal settlements showing the highest propensity for irregular consumption and unmetered connections. Night-time light intensity, building density, and proximity to secondary substations also contributed to understanding hotspot locations, providing a multi-dimensional perspective on theft risk.

An ensemble machine learning model combining Random Forest, Light Gradient Boosting Machine, and XGBoost was developed and trained on a balanced dataset augmented through Localized Randomized Affine Shadowsampling. The ensemble model achieved strong predictive performance, with an overall accuracy of 95%, F1-score of 90%, recall of 90.8%, and precision of 91%. Feature importance analysis revealed that average monthly consumption was the most influential predictor, while spatial coordinates and socio-economic factors provided additional context to differentiate normal from anomalous consumption patterns.

The WebGIS dashboard was successfully implemented to integrate predictive analytics with spatial visualization. It enables interactive exploration of theft hotspots, field navigation, and inspection planning, with tools such as zoom-to-hotspot, meterbox attribute viewing, distance measurement, and real-time machine learning predictions. The platform allows utility operators to prioritize interventions efficiently and provides a practical framework for ongoing monitoring and decision-making.

5.2 Recommendations

5.2.1 Smart Meter Installation

Prioritize the deployment of smart meters in high-risk zones, particularly in dense informal settlements and commercial clusters. Smart meters will improve the accuracy of consumption records, enable real-time monitoring of anomalies, and reduce opportunities for meter bypassing

or illegal connections. The continuous data flow from smart meters can also support future model retraining and hotspot refinement.

5.2.2 Field Validation and Ground-Truthing

Implement systematic field validation protocols where inspection teams regularly verify flagged anomalies. Feedback from on-the-ground inspections should be incorporated into the model, allowing continuous learning and adaptation to emerging theft patterns. This approach ensures the predictive model remains reliable over time and improves confidence in operational decisions.

5.2.3 Expansion to Other Sub counties

Extend the predictive framework and WebGIS dashboard to additional sub counties in Mombasa County, focusing on areas with high population density or reported theft prevalence, such as Tudor slums. Applying the same methodology across multiple areas enables a standardized, scalable approach for electricity theft management, supporting proactive inspection and resource allocation.

5.2.4 Operational Integration and Training

Train utility personnel on effectively using the WebGIS dashboard, including navigating hotspots, querying meterbox attributes, interpreting machine learning predictions, and utilizing measurement and reporting tools. Regular training ensures staff can leverage predictive insights to prioritize inspections, optimize field routes, and make data-driven decisions.

5.2.5 Policy Development and Community Engagement

Formulate policies that support the enforcement of electricity usage regulations and integrate community engagement programs to raise awareness about electricity theft consequences. Educating residents about the social and economic impact of theft encourages reporting of illegal connections, improves compliance, and fosters collaboration between utility providers and the community.

5.2.6 Infrastructure Upgrades and Maintenance

Use the hotspot analysis to guide targeted infrastructure upgrades, such as reinforcing distribution lines, replacing aging transformers, and extending the grid to underserved areas. Proper maintenance and strategic expansion reduce technical losses and make it harder for illegal connections to go undetected.

5.2.7 Continuous Data Integration and Analytics

Maintain regular updates of billing, anomaly, and socio-economic datasets to ensure that predictive models reflect the latest consumption patterns. Integrating multiple data sources over time will improve model robustness, enhance hotspot detection, and support long-term planning and monitoring strategies.

REFERENCES

Abbas, S., Bouazzi, I., Ojo, S., Sampedro, G., Almadhor, A., Hejaili, A., & Stolicna, Z. (2024). Improving smart grids security: An active learning approach for smart grid-based energy theft detection. IEEE Access, 12, 1706-1717. https://doi.org/10.1109/access.2023.3346327

Abro, S., Hua, L., Laghari, J., Bhayo, M., & Memon, A. (2024). Machine learning-based electricity theft detection using support vector machines. International Journal of Electrical and Computer Engineering (IJECE), 14(2), 1240-1250. https://doi.org/10.11591/ijece.v14i2.pp1240-1250

Adepoju, G. A., & Oyedele, O. (2020). Adaptive boosting for electricity theft detection in prepaid metering systems. IEEE Transactions on Smart Grid, 11(3), 2345-2356. https://doi.org/10.1109/TSG.2020.2978765

Adewumi, A. O., & Oyewole, S. A. (2023). Geospatial intelligence for proactive non-technical loss reduction in Sub-Saharan Africa. International Journal of Applied Earth Observation and Geoinformation, 118, 103265. https://doi.org/10.1016/j.jag.2023.103265

Ahmed, S., & Khan, M. (2021). Spatial analysis of non-technical losses in power distribution networks using GIS. Energy Reports, 7, 789-802. https://doi.org/10.1016/j.egyr.2021.01.045

Alahakoon, D., & Yu, X. (2016). Smart electricity meter data intelligence for future energy systems: A survey. IEEE Transactions on Industrial Informatics, 12(1), 425-436. https://doi.org/10.1109/TII.2015.2414355

Baldawi, I., & İnan, T. (2023). A novel and effective method based on a deep learning model for detecting non-technical electricity losses. International Journal of Power Electronics and Drive Systems (IJPEDS), 14(4), 2464-2473. https://doi.org/10.11591/ijpeds.v14.i4.pp2464-2473

Bandyopadhyay, S., & Wolfson, J. (2018). GIS-based spatial clustering for electricity theft detection. International Journal of Electrical Power & Energy Systems, 97, 1-12. https://doi.org/10.1016/j.ijepes.2017.10.015

Bera, S., Misra, S., & Obaidat, M. S. (2021). A hybrid machine learning approach for electricity theft detection in smart grids. IEEE Systems Journal, 15(2), 1789-1800. https://doi.org/10.1109/JSYST.2020.3014567

Bhattacharyya, S., & Dasgupta, A. (2022). Nighttime light anomalies and informal electricity access: A remote sensing approach. Remote Sensing of Environment, 274, 112998. https://doi.org/10.1016/j.rse.2022.112998

Cárdenas, J. A., & Giraldo, L. F. (2023). Spatial econometrics for electricity theft modeling in Latin American urban slums. Energy Economics, 118, 106521. https://doi.org/10.1016/j.eneco.2022.106521

Chen, Y., & Zhang, L. (2022). Deep learning for anomaly detection in smart meter data: A comparative study. Applied Energy, 305, 117859. https://doi.org/10.1016/j.apenergy.2021.117859

Chung, J., & Kim, H. (2018). Crime risk maps: A multivariate spatial analysis of crime data. Geographical Analysis, 51(4), 475-499. https://doi.org/10.1111/gean.12182

Costa, R., & Silva, M. (2021). Integrating GIS and machine learning for proactive electricity theft detection. Energy Policy, 148, 111956. https://doi.org/10.1016/j.enpol.2020.111956

Depuru, S. S. S. R., Wang, L., & Devabhaktuni, V. (2011). Electricity theft: Overview, issues, and prevention. IEEE Transactions on Smart Grid, 2(4), 602-610. https://doi.org/10.1109/TSG.2011.2163289

Dlamini, Z., & Kanyama, A. C. (2021). Integrating VIIRS nighttime lights and census data for theft hotspot mapping in Southern Africa. Applied Geography, 135, 102556. https://doi.org/10.1016/j.apgeog.2021.102556

Dong, S., Zeng, Z., & Liu, Y. (2021). FPETD: Fault-tolerant and privacy-preserving electricity theft detection. Wireless Communications and Mobile Computing, 2021(1). https://doi.org/10.1155/2021/6650784

El-Toukhy, A., Badr, M., Mahmoud, M., Srivastava, G., Fouda, M., & Alsabaan, M. (2023). Electricity theft detection using deep reinforcement learning in smart power grids. IEEE Access, 11, 59558-59574. https://doi.org/10.1109/access.2023.3284681

Fathy, A., & Li, X. (2022). LightGBM-based framework for electricity theft detection in unbalanced datasets. IEEE Access, 10, 32145-32156. https://doi.org/10.1109/ACCESS.2022.3160123

Glauner, P., Boechat, A., Dolberg, L., State, R., Bettinger, F., & Rangoni, Y. (2017). Large-scale detection of non-technical losses in imbalanced data sets. IEEE Power & Energy Society Innovative Smart Grid Technologies Conference, 1-5. https://doi.org/10.1109/ISGT.2017.8086001

Gündüz, M., & Daş, R. (2024). Smart grid security: An effective hybrid CNN-based approach for detecting energy theft using consumption patterns. Sensors, 24(4), 1148. https://doi.org/10.3390/s24041148

Hasan, M., Toma, R., Nahid, A., Islam, M., & Kim, J. (2019). Electricity theft detection in smart grid systems: A CNN-LSTM-based approach. Energies, 12(17), 3310. https://doi.org/10.3390/en12173310

Jamil, F., & Ahmad, E. (2013). Socio-economic determinants of electricity theft in developing countries. Energy Policy, 61, 1450-1456. https://doi.org/10.1016/j.enpol.2013.06.093

Javaid, N., Akbar, M., Aldegheishem, A., Alrajeh, N., & Mohammed, E. (2022). Employing machine learning boosting classifiers-based stacking ensemble model for detecting non-technical losses in smart grids. IEEE Access, 10, 121886-121899. https://doi.org/10.1109/access.2022.3222883

Jin, Z. (2024). Electricity-theft detection with deep learning. Proceedings of SPIE, 57. https://doi.org/10.1117/12.3034927

Jindal, A., Dua, A., Kaur, K., Singh, M., Kumar, N., & Mishra, S. (2020). Decision tree and SVM-based electricity theft detection in smart grids. IEEE Systems Journal, 14(2), 2325-2336. https://doi.org/10.1109/JSYST.2019.2937563

Kabango, J. D., & Mushi, A. G. (2023). GIS-based theft risk assessment in Tanzanian informal settlements. Utilities Policy, 82, 101580. https://doi.org/10.1016/j.jup.2023.101580

Kamau, J., & Njoroge, P. (2022). Ensemble learning for electricity theft detection in Kenya. Energy for Sustainable Development, 68, 1-12. https://doi.org/10.1016/j.esd.2022.03.002

Kandpal, M., & Vaidya, J. (2020). Machine learning for electricity theft detection: A systematic review. Renewable and Sustainable Energy Reviews, 133, 110303. https://doi.org/10.1016/j.rser.2020.110303

Kgaphola, P., Marebane, S., & Hans, R. (2024). Electricity theft detection and prevention using technology-based models: A systematic literature review. Electricity, 5(2), 334-350. https://doi.org/10.3390/electricity5020017

Lin, G., Feng, H., Feng, X., Wen, H., Li, Y., Hong, S., & Ni, Z. (2021). Electricity theft detection in power consumption data based on an adaptive tuning recurrent neural network. Frontiers in Energy Research, 9. https://doi.org/10.3389/fenrg.2021.773805

Mbuli, N., & Masinde, M. (2021). GIS-based hotspot analysis of electricity theft in South Africa. Utilities Policy, 70, 101216. https://doi.org/10.1016/j.jup.2021.101216

Mhaske, D., Satam, R., Londhe, S., Kohad, T., & Kadam, S. (2022). An efficient electricity theft detection using XGBoost. International Journal of Engineering and Advanced Science and Technology (IJEAST), 6(10), 282-287. https://doi.org/10.33564/ijeast.2022.v06i10.037

Mishra, S., & Thakur, N. (2020). Spatial autocorrelation and machine learning for electricity theft mapping. Energy, 198, 117321. https://doi.org/10.1016/j.energy.2020.117321

Mumo, P., & Omondi, F. (2020). Electricity theft and grid instability in Kenya: A case study of Nairobi. Energy Policy, 142, 111541. https://doi.org/10.1016/j.enpol.2020.111541

Mwangi, A., & Karanja, S. (2021). Non-technical losses in urban Kenya: A GIS and machine learning approach. Journal of Energy in Southern Africa, 32(1), 45-58. https://doi.org/10.17159/2413-3051/2021/v32i1a8442

Nabil, M., Ismail, M., Mahmoud, M., Alasmary, W., & Serpedin, E. (2019). PPETD: Privacy-preserving electricity theft detection scheme with load monitoring and billing for AMI networks. IEEE Access, 7, 96334-96348. https://doi.org/10.1109/access.2019.2925322

Narayanan, A., & Hardy, T. (2022). Synthetic data generation for machine learning model training for energy theft scenarios using co-simulation. IET Generation, Transmission & Distribution, 17(5), 1035-1046. https://doi.org/10.1049/gtd2.12619

Nguyen, T. H., & Patel, R. (2022). Blockchain-enabled smart contracts for automated theft penalties in Vietnam. Energy Policy, 168, 113122. https://doi.org/10.1016/j.enpol.2022.113122

Okafor, C. C., & Mensah, K. O. (2023). Community-led anti-theft initiatives in Lagos: A GIS participatory mapping approach. Sustainable Cities and Society, 92, 104487. https://doi.org/10.1016/j.scs.2023.104487

Oludare, S., & Adebanji, A. (2019). GIS and spatial statistics for electricity theft detection in Nigeria. Energy Reports, 5, 1234-1245. https://doi.org/10.1016/j.egyr.2019.08.073

Park, C., & Kim, T. (2020). Energy theft detection in advanced metering infrastructure based on anomaly pattern detection. Energies, 13(15), 3832. https://doi.org/10.3390/en13153832

Rahman, M. M., & Hossain, M. S. (2024). Deep learning-driven meter tampering detection in Bangladesh's prepaid systems. IEEE Transactions on Industrial Electronics, 71(2), 1899-1910. https://doi.org/10.1109/TIE.2023.3250501

Sajol, M., Ahmed, I., & Mahmud, Q. (2024). Synthetic minority oversampling technique-enhanced machine learning models for energy theft detection. TechRxiv Preprint. https://doi.org/10.36227/techrxiv.171177491.13025779/v2

Sharma, R., & Kumar, S. (2022). Explainable AI for electricity theft detection in smart grids. IEEE Transactions on Industrial Informatics, 18(5), 3456-3467. https://doi.org/10.1109/TII.2021.3126789

Silva, R. V., & Thompson, E. L. (2023). Behavioral economics of electricity theft: A game-theoretic GIS model. Renewable and Sustainable Energy Reviews, 178, 113233. https://doi.org/10.1016/j.rser.2023.113233

Tian, C., & Li, Z. (2021). Hybrid CNN-LSTM for electricity theft detection in smart grids. IEEE Transactions on Smart Grid, 12(1), 456-467. https://doi.org/10.1109/TSG.2020.3026789

Ullah, A., Javaid, N., Asif, M., Javed, M., & Yahaya, A. (2022). AlexNet, AdaBoost, and artificial bee colony-based hybrid model for electricity theft detection in smart grids. IEEE Access, 10, 18681-18694. https://doi.org/10.1109/access.2022.3150016

Vargas, D., & Almeida, J. (2022). Satellite imagery and deep learning for unmetered consumption estimation in favelas. ISPRS Journal of Photogrammetry and Remote Sensing, 188, 1-15. https://doi.org/10.1016/j.isprsjprs.2022.03.018

Wang, Y., & Zheng, X. (2022). 5G-enabled smart meters for instantaneous theft alerts. IEEE Communications Magazine, 60(4), 50-56. https://doi.org/10.1109/MCOM.001.2100721

Wekesa, N., & Mutua, F. (2019). Electricity theft in informal settlements: A case study of Mombasa. Energy for Sustainable Development, 52, 78-89. https://doi.org/10.1016/j.esd.2019.07.003

Yang, Q., & Liu, Y. (2019). Federated learning for privacy-preserving electricity theft detection. IEEE Internet of Things Journal, 6(5), 7654-7663. https://doi.org/10.1109/JIOT.2019.2909876

Zhang, L., & Wang, H. (2019). Network mapping and vulnerability assessment for electricity theft prevention. IEEE Systems Journal, 13(3), 2789-2800. https://doi.org/10.1109/JSYST.2018.2889198

Zhang, W., & Kumar, V. (2023). Edge computing for real-time theft detection in smart meters. IEEE Transactions on Cloud Computing, 11(1), 1-14. https://doi.org/10.1109/TCC.2022.3208765