energies

MDPI

Article

Electricity Theft Detection from Electricity and Gas Measurements Using Machine Learning

Fayiz Alfaverh [1,*], Hock Gan [1], Volodymyr Miroshnyk [2], Zaid Bin Saeed [3], Ihor Blinov [2,4], Pavlo Shymaniuk [2], Pouya Tarassodi [3] and Iosif Mporas [1,*]

[1] School of Physics, Engineering and Computer Science, University of Hertfordshire, Hatfield AL10 9AB, UK; h.c.gan@herts.ac.uk
[2] Department of Modelling of Electrical Power Objects and Systems, Institute of Electrodynamics NASU, 03057 Kyiv, Ukraine; miroshnyk.volodymyr@gmail.com (V.M.); blinovihor@gmail.com (I.B.)
[3] Innvotek Ltd., Birmingham B25 8DW, UK; zaid.bin-saeed@innvotek.com (Z.B.S.)
[4] Department of Electrical Power Systems and Networks, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", 03056 Kyiv, Ukraine
* Correspondence: f.alfaverh@herts.ac.uk (F.A.); i.mporas@herts.ac.uk (I.M.)

Abstract

Electricity theft is a critical source of non-technical losses in modern power systems, causing substantial financial and operational challenges for utilities. Traditional detection methods, such as manual inspections, are inadequate to detect advanced theft techniques, including meter tampering and cyberattacks on smart grids. This study introduces a machine learning-based framework for electricity theft detection using the TDD2022 dataset (derived from OEDI) and evaluates multiple algorithms—Random Forest, Decision Tree, XGBoost, LightGBM, CatBoost, Extra Trees, and Logistic Regression. To address class imbalance, SMOTE is applied, while feature selection leverages LASSO and ReliefF. Experiments compare electricity-only data with multi-utility inputs (electricity and gas) under balanced and imbalanced conditions. Results show that tree-based ensembles, particularly Extra Trees combined with SMOTE and ReliefF, achieve superior performance (accuracy $> 95\%$, AUC ≈ 0.99). Consumer-specific models outperform global models, with commercial classes yielding near-perfect detection, while residential profiles remain challenging. The findings highlight the importance of tailored modeling and feature selection for scalable, accurate theft detection in smart grid environments.

Keywords: electricity theft; non-technical loss; machine learning; SMOTE; feature selection; TDD2022; OEDI

Received: 4 February 2026
Revised: 10 April 2026
Accepted: 10 April 2026
Published: 23 April 2026

1. Introduction

Electricity theft, a form of non-technical loss, poses severe economic and safety risks to power utilities. High electricity tariffs with unaffordable bills have become a common cause. Opportunities present themselves where there is a lack of strict enforcement and monitoring. Corruption and collusion within utility systems are not uncommon. Global estimates suggest annual losses exceeding EUR 96 billion [1], with countries like India reporting transmission and distribution (T&D) [2] losses of up to 20% [3] and Brazil incurring losses of BRL 8.15 billion (approx. EUR 2.3 billion) in 2017 [4]. Beyond financial implications, electricity theft is linked to illegal activities such as cryptocurrency mining [5,6] and cannabis cultivation [7,8], both of which demand high energy consumption and often involve bypassing meters or tapping distribution lines.

The consequences of theft extend beyond revenue loss for power companies and increased tariffs for honest customers. Incorrect demand estimation can lead to voltage sags [9], power disruptions, and even fatalities due to unsafe connections. Extreme weather conditions exacerbate these risks, causing fires and electrocution hazards. In general, damage to electrical infrastructure and safety hazards are additional consequences [10].

The pursuit of detecting electricity theft often necessitates monitoring consumer behavior at an extremely granular level. While this granularity helps utilities identify irregularities, it creates significant privacy risks by allowing for the detailed profiling of individuals [11].

Contributions:

This study addresses a binary classification task, namely, electricity theft detection (Normal vs. Theft). Although the dataset includes six synthetic theft types, these are aggregated into a single "theft" category for model training and evaluation. This study is part of the Artificial Intelligence (AI) enhanced microgrid architecture Optimised Microgrid Management—Ukraine (OMMU) [12], and it investigates three distinct scenarios: the utilization of electricity measurements in isolation, the integration of multi-utility data (electricity and gas), and the application of Synthetic Minority Oversampling Technique (SMOTE) for data preprocessing. Additionally, this research includes a granular performance analysis across specific consumer categories.

Overview:

The remainder of this article is organized as follows. In Section 2, a literature review is performed on research in the detection of electricity theft centered around the use of machine learning techniques. In Section 3 the materials and methods used are described. Section 4 presents the results, and Section 5 discusses the results. Finally, in Section 6, conclusions are provided.

2. Literature Review

Electricity theft detection has evolved significantly over the past decade, transitioning from manual inspection methods to advanced data-driven approaches. Historically, utilities relied on periodic meter readings and on-site inspections to identify anomalies. These methods were labor-intensive, prone to human error, and ineffective against sophisticated theft techniques such as meter bypassing [13] or cyberattacks on smart infrastructure [14].

The introduction of smart meters and Advanced Metering Infrastructure (AMI) [15] has enabled real-time monitoring and granular data collection, creating opportunities for machine learning-based detection. Smart meters provide high-frequency consumption data, which can be analyzed to identify irregular patterns indicative of theft.

Recent research emphasizes the role of Machine Learning (ML) algorithms in detecting non-technical losses. Commonly used models include Support Vector Machines (SVMs) [16], Decision Trees (DTs), Random Forests (RFs) [17], and eXtreme Gradient Boosting (XGB) [18]. These algorithms excel in handling large datasets and capturing complex consumption behaviors. For instance, Ref. [5] demonstrated that fraudulent profiles associated with Bitcoin mining exhibit a "flat curve with large power consumption," which can be detected through pattern analysis. Similarly, Ref. [7] proposed a state estimation model to identify partial meter bypass scenarios, highlighting the growing sophistication of theft techniques.

A critical challenge in electricity theft detection is class imbalance, as fraudulent cases represent a small fraction of overall consumption data. Techniques such as SMOTE are widely adopted to address this imbalance, ensuring robust model training. Furthermore, feature selection plays a pivotal role in improving model accuracy and interpretability.

Methods like Least Absolute Shrinkage and Selection Operator (LASSO) in LASSO-SVM and the use of the ReliefF feature selection algorithm in conjunction with Logistic Regression (LR) (LR-ReliefF) [19] have been employed to rank and select the most relevant features, reducing dimensionality and enhancing performance.

Popular datasets such as Theft Detection Dataset 2022 (TDD2022) (which is a modification of Open Energy Data Initiative (OEDI) data) [17] provide real-world consumption profiles, enabling researchers to benchmark algorithms under realistic conditions. Experimental protocols typically involve data cleaning, balancing, feature selection, and rigorous evaluation using metrics like accuracy, precision, recall, and F1-Score. The use of gas usage readings in addition to the electricity usage pattern can improve theft detection accuracy due to the strong correlations that exist between them [20].

Beyond traditional theft detection studies, recent research in 2024–2025 has expanded the broader energy analytics landscape with innovations in imputation, optimization, and hybrid modeling that are relevant to the methodology of this work. For example, Ref. [21] highlights the increasing importance of hybrid ML frameworks for handling noisy or incomplete data—issues directly applicable to the preprocessing of multi-utility consumption profiles. Similarly, Ref. [22] demonstrates the benefits of combining classical ML with optimization-driven approaches to improve system-level decision making, reinforcing the need for robust and interpretable models within modern microgrid architectures. Advances in forecasting, such as [23], further illustrate the trend toward hybrid computational pipelines that integrate feature engineering, non-linear modeling, and optimization—principles shared with feature selection and ensemble modeling strategies explored in the present work.

Despite these advancements, gaps remain in scalability, adaptability to evolving theft patterns, and integration with real-time grid operations. Future research directions include hybrid models combining statistical and machine learning approaches, deployment of federated learning for privacy-preserving detection, and leveraging graph-based techniques [24] to analyze network-level anomalies.

2.1. Overview of Benchmark Datasets

Table 1 summarizes widely used datasets for electricity theft research, covering their origin, labeling strategy (real vs. synthetic theft), scale, and data modalities. The TDD2022 (OEDI-based) corpus provides hourly consumption profiles with six synthetic theft types across sixteen consumer classes and ten meter channels, enabling controlled benchmarking. In contrast, the State Grid Corporation of China (SGCC) dataset offers real smart-meter records with expert-verified theft labels, while Irish Smart Energy Trial (ISET)-, Ausgrid-, and Reference Energy Disaggregation Data Set (REDD)-derived sets vary in resolution and scope (from household-level hourly data to appliance-level high-frequency traces) and are used per se in anomaly detection or modified to contain synthetic data. This diversity highlights trade-offs between realism, granularity, and experimental control that should be considered when selecting datasets for training and evaluation. Generally, the choice for machine learning of theft is between using real theft samples [25] or synthetically generating the theft samples from real data [17]. There are techniques that depend on training using the consumption data of honest users to reconstruct typical usage patterns. When theft data is input, the reconstruction error becomes significantly larger [26,27].

Table 1. Summary of bibliography of electricity theft datasets (Part 1).

[REF]	Dataset Name	Origin	Labels	Theft Data	# Samples	# Consumers	Data Types
[17]	OEDI-based	US	6 types of theft [1] + normal	Synthetic	560,640	16 different classes [2]	10 m types [3]
[25]	SGCC	China	Theft/Normal	Real	1034 × 42,372	42,372	Electricity only
[16]	ISET-based	Ireland	6 types of theft [1] + normal	Synthetic	535 × 24	5000+	Hourly electricity
[28]	Ausgrid-based	Australia	4 types of theft	Synthetic	31 × 3 × 24	31	Solar + load
[29]	REDD	US	Normal	Real	1 TB raw	10	Appliance-level consumption

[1] Types of theft: Types 1 and 2 reduce electricity consumption by a certain proportion—(1) by a constant amount and (2) by a time-varying amount—while Type 3 subtracts a constant fixed value from electricity consumption. Types 4 and 5 set the user's electricity consumption to zero and to the recent average value, respectively. Type 6 reverses the order of electricity consumption to simulate consistent use during low-price periods. [2] Class of consumer: (1) FullServiceRestaurant; (2) Hospital; (3) LargeHotel; (4) LargeOffice; (5) MediumOffice; (6) MidriseApartment; (7) PrimarySchool; (8) OutPatient; (9) Warehouse; (10) SecondarySchool; (11) SmallHotel; (12) SmallOffice; (13) StandaloneRetail; (14) StripMall; (15) SuperMarket; (16) QuickServiceRestaurant. [3] Type of data (10 types): (1) Electricity:Facility, (2) Fans:Electricity, (3) Cooling:Electricity, (4) Heating:Electricity, (5) InteriorLights:Electricity, (6) InteriorEquipment:Electricity, (7) Gas:Facility, (8) Heating:Gas, (9) InteriorEquipment:Gas, (10) WaterHeater:WaterSystems:Gas.

2.2. Extended Dataset Characteristics

Table 2 expands on the summary in Table 1 by detailing the scope and structure of widely used electricity theft datasets. TDD2022 (OEDI-based) offers hourly consumption profiles for one year, including synthetic theft patterns across sixteen consumer classes and ten meter channels. SGCC provides large-scale real smart meter data from over 42,000 consumers with expert-labeled theft cases, enabling high-fidelity benchmarking. ISET-based datasets incorporate synthetic theft scenarios into smart meter trials, while Ausgrid combines solar and load data for residential homes. REDD delivers appliance-level consumption traces at a high resolution, primarily for Non-Intrusive Load Monitoring (NILM) research but adapted for theft detection studies. These datasets collectively illustrate the trade-offs between realism, granularity, and experimental control in electricity theft detection research.

Table 2. Extended dataset descriptions for electricity theft research (Part 2).

[REF]	Dataset	Cited by	Description
[17]	OEDI-based (TDD2022)	[17,18,26,27,30–38]	Hourly kW consumption for one year, 35 k samples/user, 12 features (10 m types, user class, theft type), 16 consumer classes, synthetic theft patterns applied.
[25]	SGCC (China State Grid)	[25,30,33,36,39–41]	Large-scale utility smart meter records from 42 k consumers, spanning 2 years. Theft labels are assigned by onsite inspections and expert review.
[16]	ISET-based (Irish SEAI Study)	[16,42]	Smart meter trial with 12,840 hourly samples per consumer, electricity-only, with synthetic theft scenarios injected using time-shift, scaling, and zero-load patterns.
[43]	Ausgrid-based Dataset	[43]	Combined solar and load data for 31 homes, 3 years at 30 min resolution, converted to hourly consumption; multiple theft simulation techniques applied.
[29]	REDD (Building-level Disaggregation)	[27,29,30]	High-resolution appliance-level energy dataset (1 TB), real recordings, primarily for Non-Intrusive Load Monitoring (NILM) research but adapted by several studies to simulate theft patterns.

2.3. Theft Detection Study Method

Table 3 provides a comparative summary of electricity theft detection studies published between 2023 and 2025. It lists the reference, year, datasets used (including TDD2022, SGCC, Ausgrid, and others), input data types (such as hourly profiles, 1D/2D time-series, and multimodal measurements), and whether theft data is synthetic or real. The table details the machine learning and deep learning methods applied, ranging from classical models (K-Nearest Neighbours (KNN), Decision Tree, Random Forest, XGBoost, Light

Gradient Boosting Machine (LightGBM), Gradient Boosting using Categorical features (CatBoost)) to advanced architectures (Recurrent Neural Network (RNN), Bidirectional Long Short-Term Memory (BiLSTM)-Conditional Random Field (CRF), Convolutional Neural Network (CNN) hybrids, MobileNet, Transformer–Generative Adversarial Network (GAN), and Reinforcement Learning (RL)). It also reports the best performance metrics achieved in each study, including accuracy, F1-Score, precision, recall, Area Under the (Receiver Operating Characteristic) Curve (AUC), and other indicators. Reported results vary widely, with synthetic datasets achieving up to 97.7% accuracy, while real-world datasets typically yield lower scores (e.g., F1 around 85%). Emerging approaches such as quantum machine learning, federated learning, and anomaly detection using GANs and reinforcement learning are also highlighted.

Table 3. Summary of electricity theft detection studies.

[REF]	Year	Dataset	Input Data	Theft Data	Method Used	Best Reported Results
[17]	2023	TDD2022 (560 k samples, 16 consumers, OEDI-based)	10 m types × 6 theft types + normal	Synthetic	KNN–DT–RF classifier. 4 validation protocols (P7C, P7U, P6C, P6U)	Random Forest up to 94.71% accuracy, 94.56% F1 (P6C).
[18]	2023	TDD2022 (560 k samples)	Same as [17]	Synthetic	RF + XGBoost + MLP ensemble	Accuracy 94.75%, F1 94.87%, precision 94.90%, Recall 94.88%.
[30]	2024	TDD2022 + SGCC + Ausgrid	1D/2D time-series load profiles. Use of SMOTE for data balancing	(1) Syn (2) Real	RNN–BiLSTM–CRF hybrid. Temporal patterns from 1D data and contextual/spatial correlations from 2D learnt	Accuracy 93.05%, precision 90.02%, Recall 84.18%, F1 86.62%.
[31]	2025	TDD2022 + Ausgrid	Hourly profiles with synthetic attacks generated in both TDD2022 and Ausgrid data	Synthetic	Classical ML (LSTM, XGBoost, LightGBM, CatBoost) compared with Quantum ML. Results for consumption and net metering domain	Accuracy: 0.977 (net meter); 0.87 (consumption)
[32]	2025	TDD2022	Same as [17]. Use of SMOTE	Synthetic	KNN, DT, RF, Bagging, Ensemble Learning	Ensemble Learning Accuracy: 97.7%.
[33]	2024	SGCC + TDD2022 + Ausgrid	Hourly electricity consumption. Use of SMOTE.	(1) Syn (2) Real	RoGRUT and pre-trained DNN compared with CNN and XGBoost hybrid	Accuracy: 91.75%, precision: 89.02%, Recall: 83.88%, F1: 85.12%, AUC: 91%, MCC: 0.82
[34]	2024	TDD2022 (22,330 samples)	Same as [17]	Synthetic	RF, XGBoost, DT, Gradient Boosting, KNN, CatBoost, LightGBM	Accuracy: 70.61%
[35]	2025	TDD2022	Same as [17]	Synthetic	CNN-Random Forest hybrid	Accuracy: 88.84%, precision: 86.55, Recall: 88.84%, F1: 86.60%
[36]	2025	TDD2022 + SGCC	Same as [17]	(1) Syn (2) Real	MobileNet and CNN for IoT deployment	SGCC Accuracy: 93%; Computation time
[37]	2025	TDD2022 + SGCC	Same as [17]	(1) Syn (2) Real	MobileNet and Deep CNN (2D load images) for IoT deployment	Attack reduction and Privacy preservation rates
[38]	2024	TDD2022	Same as [17]	Synthetic	XGBoost + Optuna tuning	Accuracy: 87.03%, AUC: 93.74%, F1: 84.46%, Kappa: 78.37%
[27]	2025	TDD2022 + Smart Meter + Pecan Street + REDD	Grid, intermediate and household level patterns	Raw real	CNN + RNN and Reinforcement Learning for anomaly detection	Accuracy: 95.83%, AUC: 98.27%, Recall: 96.21%
[26]	2024	TDD2022 + AMI + Smart Meter + Pecan Street	Grid, intermediate and household level patterns	Raw real	Transformer—GAN for anomaly detection	Accuracy: 92.51%, AUC: 96.64%, Recall: 93.17%, F1: 92.51%

3. Materials and Methods

3.1. OMMU Microgrid Architecture

The architecture of the microgrid [44] developed in the OMMU project is illustrated in Figure 1. The core parts of OMMU are the microgrid controller and the data platform. The microgrid controller is interconnected with clean energy sources (PV solar panels and optionally, wind turbines), with an energy storage system (ESS) and optionally with other green energy generation units (waste, crop residue). Alternating Current (AC)/Direct Current (DC) inverters are used for the generator, the solar panel and the Energy Storage System (ESS) serviceable battery, which produce and store DC power while the electrical appliances are AC loads. A smart meter is used to measure the aggregated energy consumption from the OMMU microgrid. The microgrid can operate in both islanded (off-grid) and grid-connected (on-grid) mode. When the microgrid is in on-grid operational mode, the microgrid controller is also connected to the utility grid. The microgrid controller switches the load demand to the ESS, the energy generator or to the grid and optionally controls the operation and scheduling of electrical appliances and other loads.

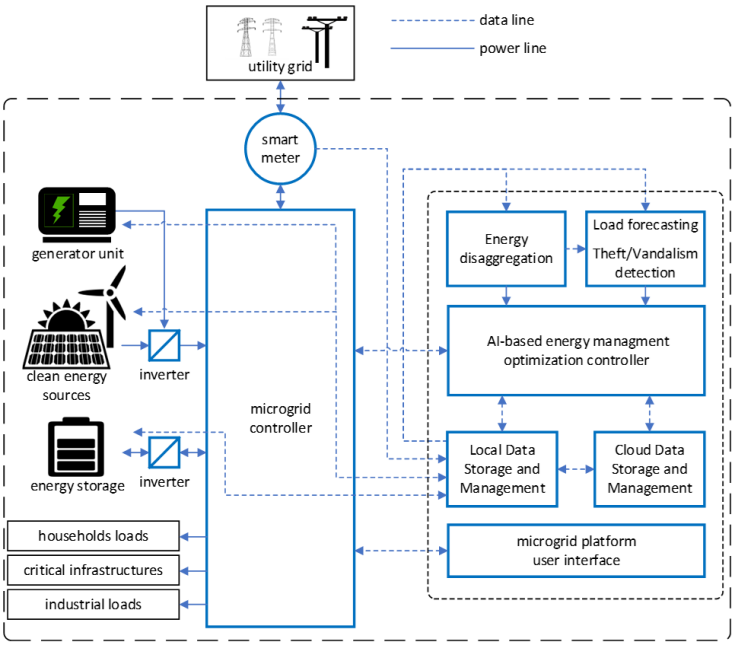

Figure 1. Block diagram of the OMMU microgrid architecture.

The aggregated energy consumption acquired by the smart meter is further used to optimize the energy management of the microgrid using a NILM setup in order to split the aggregated energy consumption signal on device level [45]. Moreover, there are AI models to predict the load demand in the microgrid [46], optimize the scheduling and the management within the microgrid [47], and, in parallel, to minimize the distortion to the utility grid [48]. A cornerstone of the OMMU architecture is its AI-driven model designed to identify electricity theft and vandalism. This paper specifically focuses on the electricity theft detection module, which monitors and evaluates consumption patterns to pinpoint anomalies. By processing data harvested from smart meters, the system analyzes usage trends and generates automated alerts upon detecting suspicious behavior. This study provides a comparative evaluation of various machine learning models to determine their efficacy in this context.

The remainder of this paper focuses exclusively on the design, evaluation, and optimization of the electricity-theft detection module depicted in Figure 1, which constitutes a key analytical component of the OMMU microgrid.

3.2. Experimental Setup

3.2.1. Dataset

The experiments use the TDD2022 dataset derived from the Open Energy Data Initiative (OEDI). The dataset consists of hourly consumption records for multiple consumer types over a full year. Each training instance corresponds to a single hourly observation containing 10 measurement channels (6 electricity, 4 gas), the consumer class label, and a binary target (Normal or Theft). The dataset contains 10 types of meter readings (6 electricity, 4 gas), consumer class, and theft type labels.

The theft synthesis is not performed in this study, but is the original dataset construction described in [17]. In that dataset, synthetic theft is generated at the daily level by applying transformations to 24 h consumption vectors. The dataset is then provided in hourly format, where each hour is treated as an individual sample. All hourly observations within a manipulated daily block inherit the same theft label.

The total number of instances is approximately 560k across 16 consumer classes. Details of the consumer class are given in the table, and those of the meter channels are shown in Table 1 [17].

3.2.2. Dataset Composition

Table 4 presents the distribution of samples in the TDD2022 dataset across sixteen consumer classes. Each class contributes 35,040 instances, split between normal and theft scenarios, resulting in a total of 560,740 samples. The dataset includes ten meter channels (six electricity, four gas), providing multimodal input for model training. Specifically, the six electricity channels are Electricity:Facility, Fans:Electricity, Cooling:Electricity, Heating:Electricity, InteriorLights:Electricity, and InteriorEquipment:Electricity, while the four gas channels are Gas:Facility, Heating:Gas, InteriorEquipment:Gas, and WaterHeater:WaterSystems:Gas (shown in Table 1). This balanced representation of consumer types ensures diversity in consumption patterns, while synthetic theft injections enable controlled evaluation of detection algorithms.

Table 4. Data distribution in the TDD2022 [17] dataset.

Consumer Type	# Normal Samples	# Theft Samples	# All Samples
FullServiceRestaurant	20,778	14,262	35,040
Hospital	20,319	14,721	35,040
LargeHotel	20,196	14,844	35,040
LargeOffice	20,977	14,063	35,040
MediumOffice	21,234	13,806	35,040
MidriseApartment	20,663	14,377	35,040
OutPatient	20,361	14,679	35,040
PrimarySchool	21,213	13,827	35,040
QuickServiceRestaurant	20,539	14,501	35,040
SecondarySchool	21,048	13,992	35,040
Small Hotel	20,503	14,537	35,040
SmallOffice	21,326	13,714	35,040
StandaloneRetail	21,117	13,923	35,040
StripMall	20,742	14,298	35,040
SuperMarket	20,539	14,501	35,040
Warehouse	20,951	14,089	35,040
Total	330,517	230,223	560,740

The dataset includes 10 types (Table 1 [17]) of input meter readings. Each hourly instance contains six electricity-related measurement channels. This does not imply an increase in sample size; rather, each observation includes multiple feature channels corresponding to different electricity subsystems.

3.2.3. Theft Synthesis Equations

Let the daily electricity consumption vector be

$$\mathbf{X} = \{x_1, x_2, \ldots, x_{24}\},$$

where x_i denotes the hourly consumption at hour $i \in \{1, \ldots, 24\}$. Let $\overline{x} = \frac{1}{24}\sum_{i=1}^{24} x_i$ denote the daily mean.

Theft1 (Uniform proportional reduction).

$$x'_i = a\, x_i, \qquad a \sim \mathsf{Uniform}(0.1,\, 0.8), \quad \forall i.$$

Theft2 (Random zeroing over a contiguous interval).

Choose a start time $t_{\text{start}} \in \{0, \ldots, 23\}$ and a duration $\text{duration}_{\text{off}} \in \{1, \ldots, 24\}$, then let $t_{\text{end}} = t_{\text{start}} + \text{duration}_{\text{off}}$. Define

$$x'_i = \begin{cases} 0, & t_{\text{start}} < i < t_{\text{end}}, \\ x_i, & \text{otherwise.} \end{cases}$$

Theft3 (Hour-wise random proportional reduction).

$$x'_i = c_i\, x_i, \qquad c_i \sim \mathsf{Uniform}(0.1,\, 0.8) \text{ i.i.d.}, \quad \forall i.$$

Theft4 (Hour-wise random fraction of daily mean).

$$x'_i = d_i\, \overline{x}, \qquad d_i \sim \mathsf{Uniform}(0.1,\, 0.8) \text{ i.i.d.}, \quad \forall i.$$

Theft5 (Constant daily mean reporting).

$$x'_i = \overline{x}, \qquad \forall i.$$

Theft6 (Time-order reversal).

$$x'_i = x_{24-i}, \qquad i = 1, \ldots, 24.$$

3.3. Data Visualization

The distribution of the theft data is completed in blocks throughout the samples of the data for each consumer class.

Figure 2 shows a stacked bar chart of the distribution for a specific example of a consumer class. The boundaries of the distribution bins split the boundaries of the blocks of theft between bins. The sample IDs are hourly samples of the electricity consumption at a quick-service restaurant facility.

Methodology

The methodology for electricity theft detection in this study is structured into a series of systematic steps, encompassing data preprocessing, feature selection, model training, and evaluation. The workflow ensures robust handling of imbalanced data and optimal feature selection for improved detection accuracy.

Although the TDD2022 dataset contains year-long hourly profiles, the current study models each hour as an independent sample, in alignment with previously reported study on the same dataset [17]. No temporal, sequential, or seasonal features were used, with the classifiers operating entirely in a static feature space, to isolate the effects of feature selection, SMOTE, and consumer-specific modeling. As the data were evaluated using a random train–test split, temporally adjacent observations may appear in both training and testing sets, which may affect model generalizability in real-world deployment.

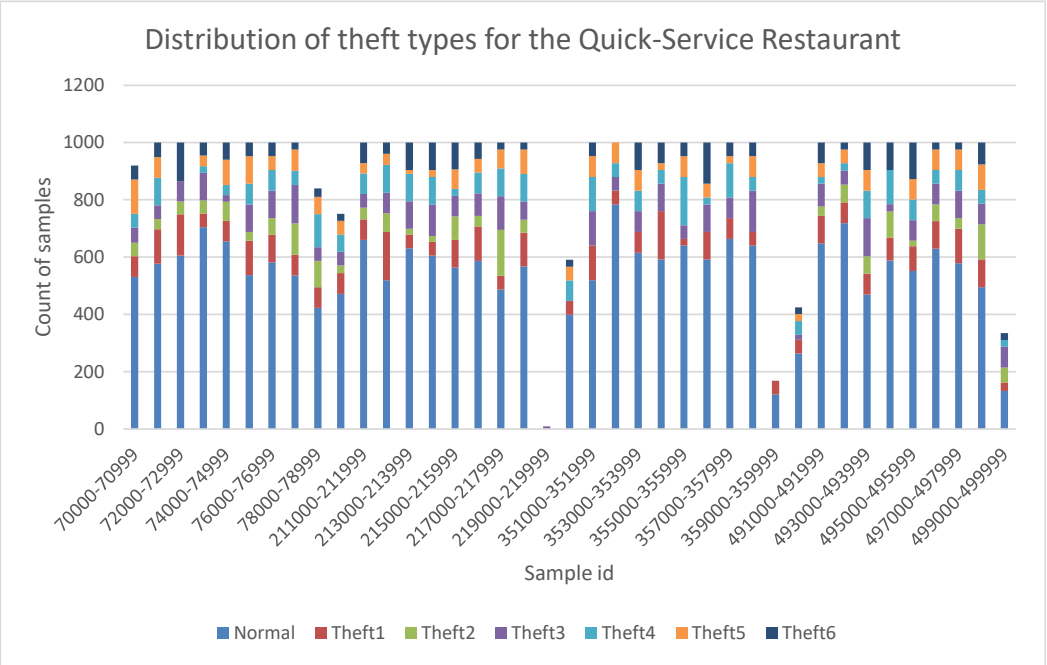

Figure 2. Stacked chart of theft data for a quick-service restaurant.

Data loading:

The raw dataset is imported from a CSV file, and unnecessary columns (e.g., index column "0" with sample ids) are removed. Data manipulation is performed using Pandas.

Label Encoding:

Categorical labels for theft and consumer types are converted into numerical values using LabelEncoder from scikit-learn, enabling compatibility with machine learning algorithms.

Train-test Split:

The dataset is partitioned into training and testing subsets with an 80:20 random split using the train_test_split from scikit-learn.

Feature Scaling:

To normalize feature values, Min–Max scaling is applied. The scaler is fitted on the training set and subsequently applied to both training and test sets.

Feature Selection:

Two feature selection techniques are employed:

1. LASSO (L1-regularized logistic regression).
2. Relief F algorithm (ReliefF).

Each method is tested with two configurations (k = 3 and k = 4 features), resulting in four settings: LASSO (3), LASSO (4), ReliefF (3), and ReliefF (4). While our study compares LASSO and ReliefF independently, we recognize that combining linear and non-linear selectors or applying stability selection frameworks may yield more reliable feature subsets, particularly in the presence of correlated or heterogeneous consumption features. Future work will incorporate such approaches.

Cross-validation for Feature Selection:

Feature selection settings are evaluated using 10-fold stratified cross-validation on the training data, with AUC used to select the best-performing feature subset.

To reduce fold dependence, feature selection is performed independently within each fold, and the frequency of selected features across folds is computed. Features consistently selected across folds are considered stable and are reported as the final subsets for each scenario.

Variations across consumer classes are expected due to differing consumption patterns, while within-scenario stability is ensured through this procedure.

Data Balancing with SMOTE:

In this study, electricity theft detection is formulated as a binary classification problem (normal vs. theft). Therefore, SMOTE was applied strictly in binary mode to balance the minority (theft) class. The oversampling procedure was performed exclusively on the training data after feature selection, while the test set remained untouched to ensure an unbiased evaluation.

Although the dataset exhibits a moderate class imbalance (approximately 60% normal and 40% theft samples, as shown in Table 4), SMOTE was employed to investigate its impact on improving minority class detection, particularly recall. This is important in practical electricity theft scenarios, where failing to detect fraudulent behavior can have significant economic consequences.

Furthermore, by applying SMOTE after feature selection, the dimensionality of the feature space is reduced, which helps mitigate the risk of generating unrealistic or noisy synthetic samples. It is also noted that concerns associated with multi-class SMOTE do not apply in this work, as oversampling is performed in a binary setting.

Model Training:

A wide range of prediction and anomaly-detection methods exist for smart-grid analytics, including deep neural architectures, hybrid metaheuristics, and temporal sequence models. However, the aim of this study is not to propose a novel prediction algorithm but to establish a controlled and interpretable benchmark for evaluating feature-selection strategies, class balancing, and consumer-specific versus global modeling scenarios on the TDD2022 dataset. For this reason, we adopt the following seven well-established tree-based classifiers as baseline models. A fixed random seed (42) is used consistently across all experiments, and the reported results correspond to a single experimental run. The models and their corresponding hyperparameter settings are as follows:

1. RF [49]: n_estimators = 300, max_depth = None, n_jobs = -1
2. DT [17,50]: default settings with random_state = 42
3. XGB [51]: n_estimators = 300, learning_rate = 0.1, max_depth = 8, subsample = 0.9, colsample_bytree = 0.9
4. LightGBM [52]:
   - `n_estimators` = 700, `learning_rate` = 0.05
   - `num_leaves` = 127, `min_child_samples` = 15
5. CatBoost [31,53]: iterations = 500, learning_rate = 0.08, depth = 8
6. Extra Trees/Extremely randomized trees (ET) [54,55]: n_estimators = 400, max_depth = None, n_jobs = -1
7. LR: penalty = L2, C = 1.0, solver = “lbfgs”, max_iter = 2000

These methods are widely used in electricity theft detection and in energy systems machine learning due to their strong performance on tabular, non-temporal data; their robustness to non-linear feature interactions; and their relatively low computational cost. Using standard, reproducible models allows us to isolate the effects of feature selection (LASSO vs. ReliefF) and SMOTE without confounding them with architectural complexity or additional hyperparameter dimensions. More advanced architectures—such as recurrent neural networks, temporal CNNs, transformer-based anomaly detectors, or hybrid optimization-driven methods—are acknowledged as important future directions but lie beyond the scope of this baseline methodological comparison.

Libraries include scikit-learn, XGBoost, LightGBM, and CatBoost. To ensure a fair and consistent comparison across all evaluated machine learning models, no systematic hyperparameter optimization (e.g., grid search or Bayesian optimization) was performed.

Model Evaluation:

Models were assessed on the test set using metrics such as accuracy, precision, recall, F1-Score, and Area Under the Curve (AUC). All experiments were conducted on a Windows 10 (64-bit) system with an Intel Core i7-1365U processor and 32 GB of RAM.

To ensure a fair and unbiased evaluation, all model selection and preprocessing steps, including feature selection (LASSO and ReliefF), class balancing using SMOTE, and algorithm configuration, were performed exclusively on the training data. A stratified 10-fold cross-validation scheme was applied within the training set to identify the optimal feature subsets and model configurations based on AUC performance.

The test set was strictly held out and used only once for the final evaluation of the selected models. No tuning or model selection decisions were based on test set performance, thereby preventing data leakage and ensuring the validity of the reported results. While nested cross-validation provides a more rigorous framework for joint model selection and performance estimation, its application in this study is constrained by the computational cost associated with the large-scale dataset (approximately 560 k samples across 16 consumer classes) and the extensive set of evaluated models. Therefore, nested cross-validation is identified as a valuable direction for future work to further strengthen model generalization assessment. As the evaluation relies on a random split rather than a temporally ordered holdout, the reported results focus on classification performance of the evaluated models, rather than time series predictive capacity.

4. Results

All experiments reported in this section correspond to the theft-detection block within the OMMU architecture.

4.1. Theft Detection Using Electricity Measurements

Table 5 displays the performance of seven machine learning algorithms (RF, DT, XGB, LightGBM, CatBoost, ET, LR) using only electricity measurements and the LASSO feature selection method. The top four features selected out of the six (Table 1 [17]) are as follows:

1. Electricity:Facility [kW] (Hourly)
2. Fans:Electricity [kW] (Hourly);
3. InteriorLights:Electricity [kW] (Hourly);
4. InteriorEquipment:Electricity [kW] (Hourly)electricity-related.

This table shows:

- Top Performers: Extra Trees (ETs) and Random Forest (RF) achieved the highest accuracy levels, at 94.76% and 94.73%, respectively.

- Other Models: Decision Tree (DT) and LightGBM also performed well, with accuracies of 93.41% and 92.38%.
- Lower Performance: Logistic Regression (LR) showed significantly lower results, with an accuracy of only 20.12% and an F1-Score of 20.52%.
- AUC Scores: Most models (except LR) demonstrated high Area Under the Curve (AUC) values, with RF and ET both reaching approximately 0.993.

Table 5. Performance of ML models for binary electricity theft detection using electricity-only measurements without SMOTE.

ML Algorithm	Best Feature Selection Method (N)	Accuracy	F1-Score	Precision	Recall	AUC	AUC-PR
RF	LASSO (4)	0.94728	0.90009	0.89920	0.90171	0.99295	0.94071
DT	LASSO (4)	0.93405	0.87556	0.87614	0.87559	0.93163	0.79428
XGB	LASSO (4)	0.89671	0.80176	0.79480	0.81734	0.98219	0.84551
LightGBM	LASSO (4)	0.92380	0.85568	0.85184	0.86071	0.98819	0.89025
CatBoost	LASSO (4)	0.86193	0.74385	0.73283	0.77356	0.97460	0.79367
ET	LASSO (4)	0.94761	0.90079	0.89969	0.90252	0.99288	0.94000
LR	LASSO (4)	0.20121	0.20517	0.26511	0.34790	0.67232	0.36221

Table 6 presents results using the same algorithms and feature selection but applies the Synthetic Minority Oversampling Technique (SMOTE) to balance the dataset:

- Best Algorithm: Extra Trees (ET) emerged as the overall best-performing model with SMOTE, improving its accuracy to 95.03% and its F1-Score to 90.46%.
- Impact of SMOTE: For most high-performing algorithms like RF and DT, the application of SMOTE resulted in very similar or slightly improved accuracy and AUC scores.
- Logistic Regression Improvement: SMOTE significantly increased the accuracy of Logistic Regression from 20.12% to 63.18%, although its F1-Score remained very low (12.91%).

Table 6. Performance of ML models for binary electricity theft detection using electricity-only measurements with SMOTE applied to the training data.

ML Algorithm	Best Feature Selection Method (N)	Accuracy	F1-Score	Precision	Recall	AUC	AUC-PR
RF	LASSO (4)	0.94733	0.89863	0.90099	0.89995	0.99352	0.94307
DT	LASSO (4)	0.93179	0.86982	0.87153	0.86958	0.92831	0.78599
XGB	LASSO (4)	0.89875	0.80109	0.80488	0.80745	0.98139	0.83941
LightGBM	LASSO (4)	0.92371	0.85288	0.85412	0.85613	0.98801	0.88842
CatBoost	LASSO (4)	0.87023	0.74145	0.75700	0.74986	0.97417	0.78431
ET	LASSO (4)	0.95028	0.90457	0.90571	0.90611	0.99421	0.95009
LR	LASSO (4)	0.63175	0.12905	0.10529	0.16667	0.69011	0.36643

Figure 3 highlights the Extra Trees (ETs) algorithm as the best performer, particularly when paired with SMOTE preprocessing.

The figure serves to visually confirm that the combination of the Extra Trees classifier, LASSO feature selection (N = 4), and SMOTE balancing provides the most effective framework for detecting electricity theft in the study's scenarios.

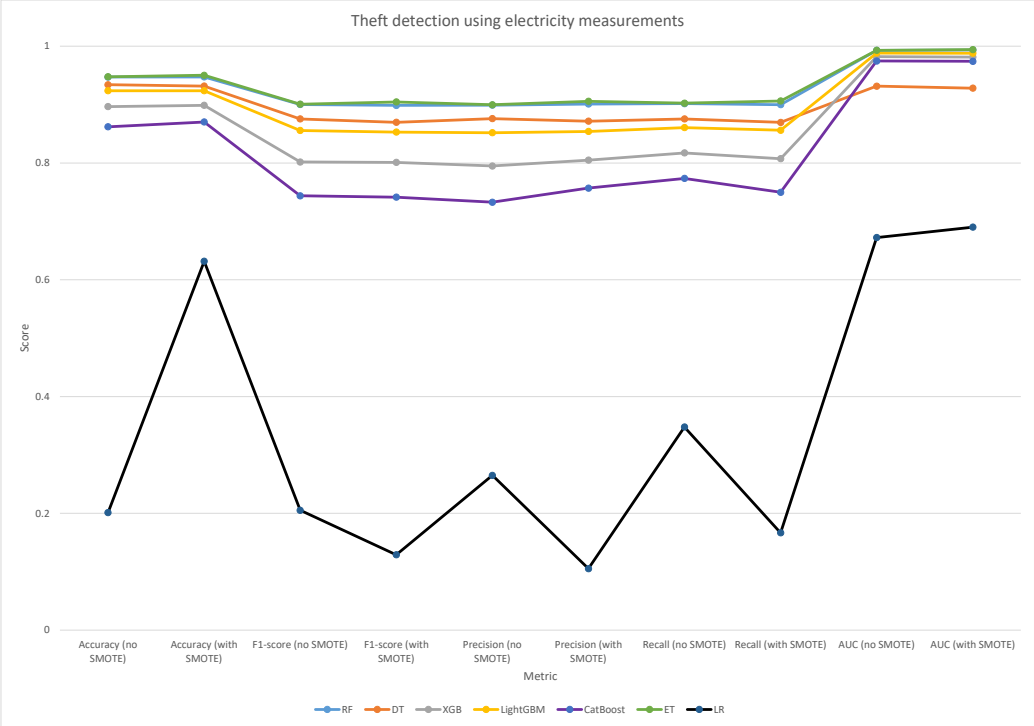

Figure 3. Comparison of model performance (accuracy, F1-Score, precision, recall, and AUC) for binary electricity theft detection using electricity-only measurements, with and without SMOTE.

4.2. Theft Detection Using Electricity and Gas Measurements

Table 7 presents the performance of seven machine learning algorithms—Random Forest (RF), Decision Tree (DT), XGBoost (XGB), LightGBM, CatBoost, Extra Trees (ET), and Logistic Regression (LR)—using a combination of electricity and gas measurements without synthetic balancing:

- Feature Selection: For most models (RF, DT, XGB, LightGBM, and CatBoost), the ReliefF algorithm was used to select the top four out of ten features, while ET and LR utilized LASSO. ReliefF selected:

  1. Electricity:Facility [kW] (Hourly)
  2. InteriorLights:Electricity [kW] (Hourly);
  3. InteriorEquipment:Electricity [kW] (Hourly);
  4. InteriorEquipment:Gas [kW] (Hourly).

  LASSO selected:

  1. Electricity:Facility [kW] (Hourly);
  2. Cooling:Electricity [kW] (Hourly);
  3. InteriorLights:Electricity [kW] (Hourly);
  4. InteriorEquipment:Electricity [kW] (Hourly).

  Comparing the selections made by LASSO with the previous results for just electricity measurements (Section 4.1), there is a small discrepancy with one item that is not uncommon if the selection list was expanded [56–58]. ReliefF chose gas measurements over cooling, but its selection is possibly based on a non-linear relationship of the samples, whereas LASSO is based on a linear relationship [59]. However, while ReliefF selected a gas-related feature among its top four, this selection alone does not establish statistical significance or prove the presence of meaningful non-linear cross-modal interactions. Further analysis is required to determine whether gas-related channels genuinely contribute predictive information.

- Performance Metrics:
  - Extra Trees (ET) achieved the highest accuracy in this scenario, 92.90%, with an F1-Score of 0.867.
  - Random Forest (RF) followed closely, with 92.82% accuracy and a slightly higher F1-Score of 0.869.
  - Logistic Regression (LR) was the poorest performer, with a very low accuracy of 25.05%.

Table 7. Performance of machine learning models for binary electricity theft detection using multi-utility measurements (electricity and gas) without SMOTE.

ML Algorithm	Best Feature Selection Method (N)	Accuracy	F1-Score	Precision	Recall	AUC	AUC-PR
RF	RELIEFF (4)	0.92820	0.86944	0.86230	0.87839	0.99022	0.91887
DT	RELIEFF (4)	0.91830	0.85114	0.84525	0.85864	0.92162	0.75666
XGB	RELIEFF (4)	0.88503	0.78483	0.77282	0.80486	0.98066	0.83515
LightGBM	RELIEFF (4)	0.90421	0.82515	0.81370	0.83965	0.98628	0.87349
CatBoost	RELIEFF (4)	0.85628	0.73580	0.72405	0.76452	0.97372	0.79135
ET	LASSO (4)	0.92895	0.86697	0.86205	0.87287	0.98859	0.90685
LR	LASSO (4)	0.25045	0.21893	0.25899	0.34967	0.66542	0.36428

Table 8 evaluates the same algorithms and multi-utility data but applies the Synthetic Minority Oversampling Technique (SMOTE) to address data imbalance.

- Impact of SMOTE: The application of SMOTE generally improved the performance of the models.
  - Extra Trees (ET) remained the top performer, with accuracy increasing to 93.64
  - Random Forest (RF) showed similar gains, reaching 93.63% accuracy.
  - For Logistic Regression, SMOTE produced a notable increase in overall accuracy (from 20.12% to 63.18%), yet the theft-class F1-score remained extremely low (0.12). This behavior reflects a substantial rise in false-positive predictions following minority-class oversampling. Unlike tree-based models, Logistic Regression enforces a single linear decision boundary and is therefore highly sensitive to the synthetic samples generated by SMOTE. The interpolation-based oversampling shifts the minority-class centroid in ways that LR cannot model effectively, leading the classifier to overextend the "theft" decision region and misclassify many normal samples as theft. This pattern reveals that linear models are poorly aligned with the non-linear structure of the synthetic theft patterns in TDD2022, and that SMOTE can exacerbate this mismatch by amplifying overlapping regions in the feature space. These results confirm that LR serves as a weak baseline for this dataset, whereas ensemble tree models remain robust under both balanced and imbalanced conditions.

Figure 4 is used to visually highlight the Extra Trees (ETs) algorithm as the most effective model for electricity theft detection in this study.

- Visualization: It confirms that the combination of the Extra Trees classifier, LASSO feature selection (using four features), and SMOTE data balancing provides the best overall framework for the detection scenarios analyzed.
- Interpretation: LR's poor performance stems from its linear nature, sensitivity to imbalance, and inability to capture complex, non-linear patterns inherent in electricity theft data. Ensemble methods thrive because they model non-linearities, handle feature interactions, and are robust to noise and imbalance.

Table 8. Performance of machine learning models for binary electricity theft detection using multi-utility measurements (electricity and gas) with SMOTE.

ML Algorithm	Best Feature Selection Method (N)	Accuracy	F1-Score	Precision	Recall	AUC	AUC-PR
RF	RELIEFF (4)	0.93628	0.87965	0.88165	0.88098	0.99070	0.92937
DT	RELIEFF (4)	0.92659	0.86216	0.86384	0.86171	0.92364	0.77295
XGB	RELIEFF (4)	0.89048	0.78698	0.79377	0.78914	0.98010	0.83079
LightGBM	RELIEFF (4)	0.91210	0.83194	0.83632	0.83164	0.98623	0.87331
CatBoost	RELIEFF (4)	0.86552	0.73343	0.75236	0.73885	0.97350	0.78370
ET	LASSO (4)	0.93642	0.87820	0.87916	0.87989	0.99024	0.92686
LR	LASSO (4)	0.63175	0.12905	0.10529	0.16667	0.68921	0.37025

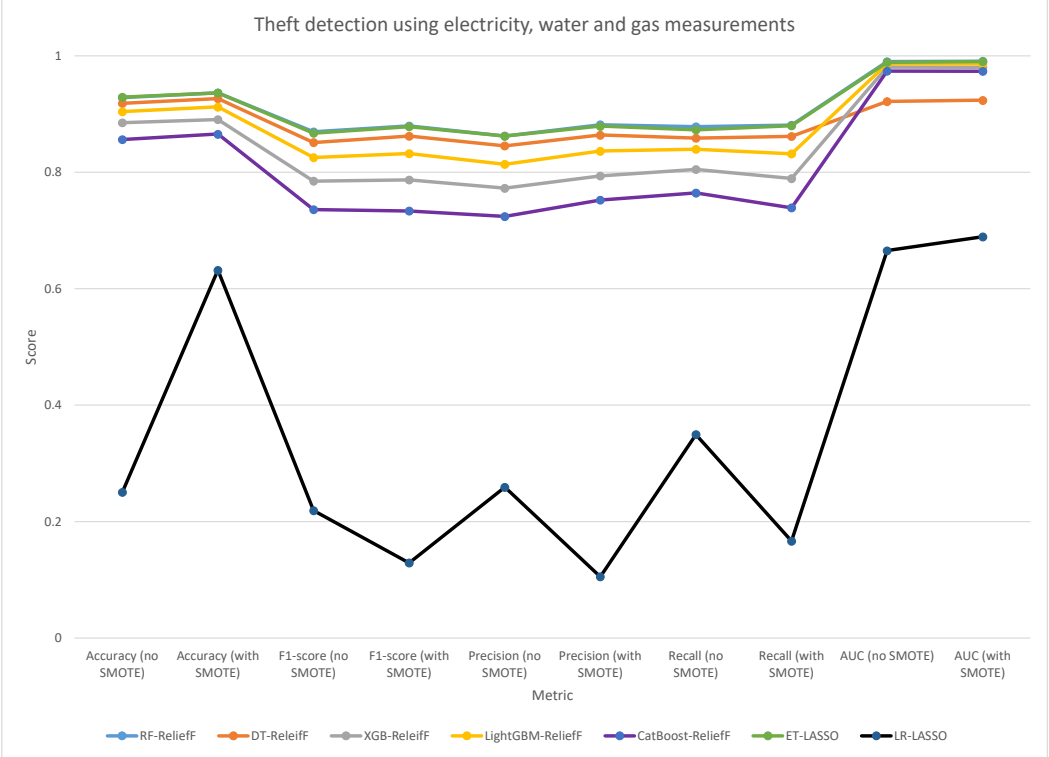

Figure 4. Comparison of model performance (accuracy, F1-Score, precision, recall, and AUC) for binary electricity theft detection using multi-utility measurements (electricity and gas), with and without SMOTE.

5. Discussion

For this section, the comparative analysis of theft performance per consumer type separately is carried out.

5.1. Using a Global Theft Detection Model

In this section, the Extra Trees algorithm is used with LASSO using four features (see Section 4.1) with electricity measurements. The performance of measurements for individual consumer classes applied to the detection model are evaluated.

Table 9 evaluates the performance of a global machine learning model (Extra Trees with LASSO feature selection) across 16 consumer classes using only electricity measurements. The metrics used are the accuracy, F1-Score, precision, recall, and AUC. The key observations are as follows:

- Highest Accuracy: quick-service restaurant (97.59%) and full-service restaurant (97.14%) performed best.
- Lowest Accuracy: midrise apartment (89.98%) and small hotel (91.07%) showed weaker performance.
- Overall Performance: average accuracy across all classes was 94.76%, with an F1-Score of 0.9006 and AUC of 0.9922.

Consumer types with more predictable load patterns (restaurants, offices) achieved higher detection accuracy, while residential-like patterns (apartments) were harder to classify.

It is important to note that although the global model achieves AUC values above 0.99 for most consumer types, the midrise apartment and a small number of other residential-like classes show lower accuracy (≈89%). This discrepancy does not reflect an evaluation failure but indicates that the classifier's ranking ability (AUC) is strong while its default decision threshold is suboptimal due to score-distribution overlap and high behavioral variability (AUC is threshold-independent, but accuracy is threshold-dependent). Threshold tuning and calibration are required for deployment in residential segments.

Table 9. Performance of the global Extra Trees model across consumer classes for binary electricity theft detection using electricity-only measurements without SMOTE.

Consumer Type	Accuracy	F1-Score	Precision	Recall	AUC	AUC-PR
FullServiceRestaurant	0.97139	0.94573	0.94686	0.94522	0.99689	0.97134
Hospital	0.96056	0.92514	0.92678	0.92476	0.99583	0.96152
LargeHotel	0.92228	0.86040	0.85978	0.86110	0.98751	0.88664
LargeOffice	0.94436	0.89355	0.89223	0.89594	0.99259	0.93717
MediumOffice	0.93915	0.88861	0.88667	0.89095	0.99054	0.93003
MidriseApartment	0.89983	0.80494	0.80150	0.81527	0.98087	0.83232
OutPatient	0.94196	0.88920	0.89056	0.89087	0.99178	0.92567
PrimarySchool	0.94964	0.90538	0.90250	0.90952	0.99432	0.94472
QuickServiceRestaurant	0.97587	0.95475	0.95760	0.95385	0.99753	0.97862
SecondarySchool	0.94277	0.89370	0.89035	0.89932	0.99195	0.93251
SmallHotel	0.91073	0.82687	0.82333	0.83187	0.98406	0.85264
SmallOffice	0.97060	0.94237	0.94312	0.94212	0.99491	0.96517
Stand-aloneRetail	0.96493	0.93059	0.93046	0.93121	0.99567	0.96316
StripMall	0.96524	0.93245	0.93179	0.93335	0.99452	0.96100
SuperMarket	0.95672	0.91727	0.92010	0.91800	0.99419	0.94953
Warehouse	0.94642	0.89987	0.89774	0.90252	0.99251	0.94174
ALL	0.94760	0.90059	0.89999	0.90278	0.99222	0.93324

Table 10 has the same setup as Table 9, but SMOTE was applied to balance the dataset. The impact of SMOTE:

- Overall Accuracy Improved: From 94.76% to 95.03%.
- Best-Performing Classes: quick-service restaurant (97.80%) and full-service restaurant (97.67%) remained top performers.
- Worst-Performing Classes: midrise apartment still lagged (89.74%).

SMOTE improved detection slightly for most classes by increasing theft samples but did not fully resolve challenges for residential categories.

Figure 5 provides a grouped chart comparing accuracy, F1-Score, precision, recall, and AUC for each consumer class before and after SMOTE. The insights gained:

- Clear Gains: most metrics improved marginally with SMOTE, especially recall for underrepresented classes.

- Top Classes: quick-service restaurant and full-service restaurant consistently dominate across all metrics.
- Challenging Classes: midrise apartment and small hotel remain the lowest performers, even after balancing.

Table 10. Performance of the global Extra Trees model across consumer classes for binary electricity theft detection using electricity-only measurements with SMOTE.

Consumer Type	Accuracy	F1-Score	Precision	Recall	AUC	AUC-PR
FullServiceRestaurant	0.97672	0.95507	0.95983	0.95373	0.99832	0.98428
Hospital	0.96363	0.92978	0.93470	0.92927	0.99705	0.97196
LargeHotel	0.92948	0.87100	0.87067	0.87278	0.98934	0.90093
LargeOffice	0.95003	0.90407	0.90622	0.90427	0.99465	0.95057
MediumOffice	0.94357	0.89633	0.89542	0.89855	0.99277	0.94278
MidriseApartment	0.89742	0.79645	0.79494	0.81111	0.98199	0.83746
OutPatient	0.94735	0.89827	0.90113	0.90134	0.99371	0.94120
PrimarySchool	0.95162	0.90857	0.91167	0.90956	0.99259	0.95012
QuickServiceRestaurant	0.97797	0.95835	0.96367	0.95735	0.99874	0.98873
SecondarySchool	0.94720	0.90020	0.90231	0.90083	0.99218	0.94271
SmallHotel	0.91073	0.82188	0.81750	0.83208	0.98481	0.85850
SmallOffice	0.96969	0.93989	0.94198	0.93954	0.99714	0.97414
Stand-aloneRetail	0.96570	0.93098	0.93203	0.93201	0.99613	0.96533
StripMall	0.96493	0.93047	0.93116	0.93135	0.99661	0.96829
SuperMarket	0.96232	0.92714	0.93055	0.92837	0.99623	0.96606
Warehouse	0.94688	0.89898	0.90153	0.89927	0.99429	0.95047
ALL	0.95028	0.90412	0.90586	0.90626	0.99352	0.94323

This figure confirms that SMOTE enhances overall robustness but consumer-specific variability persists.

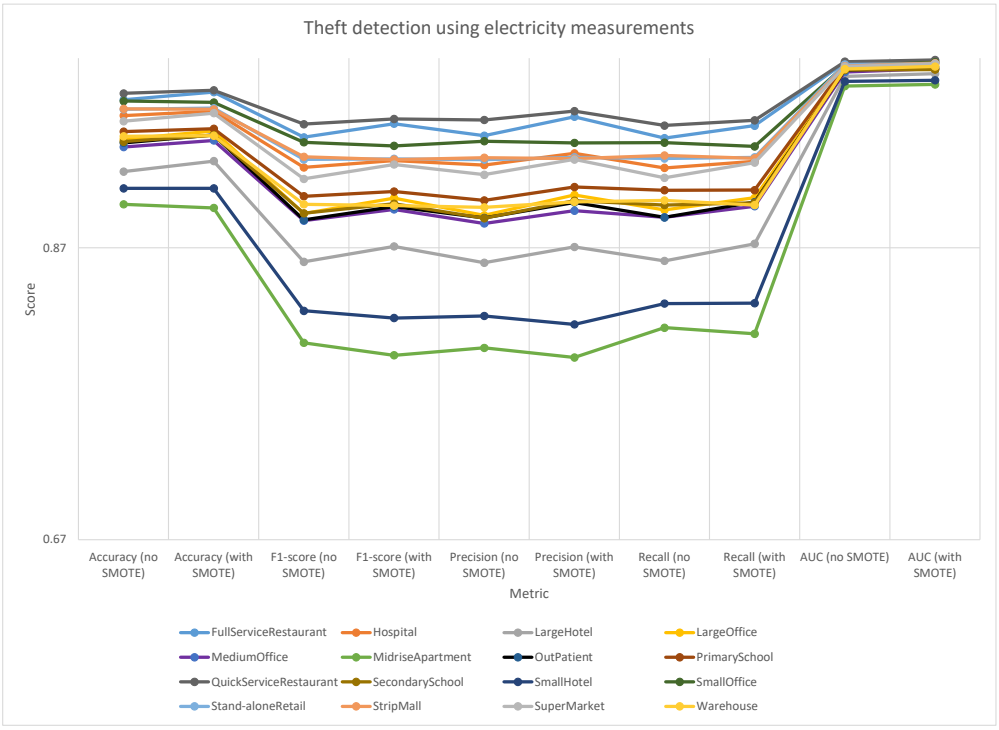

Figure 5. Performance comparison across consumer classes (accuracy, F1-Score, precision, recall, and AUC) using the global Extra Trees model for binary electricity theft detection with electricity-only measurements, with and without SMOTE.

Table 11 evaluates a global machine learning model (Extra Trees with LASSO feature selection) across 16 consumer classes using multi-utility data (electricity, gas) without applying SMOTE. The key observations are as follows:

- Overall Accuracy: 92.90% (average across all classes).
- Top Performers: quick-service restaurant (96.03%), full-service restaurant (94.98%), and strip mall (95.25%).
- Lowest Accuracy: small hotel (88.75%) and midrise apartment (89.84%).
- F1-Score: average of 0.8669; highest for quick-service restaurant (0.9245).
- AUC: very high across all classes ($\geq$0.98), indicating strong discrimination capability.

The multi-utility data performs worse than just using electricity data compared with Table 9. This behavior can be attributed to several factors. In the TDD2022 dataset, theft is injected only into electricity channels, while gas measurement remains unaffected, thus providing no direct anomaly signal. In addition, electricity features exhibit stronger intra-channel correlations, whereas cross-utility correlations are weaker.

Differences in scale and distribution across modalities may introduce noise, and some gas channels have low variability, limiting their discriminative power. Potential temporal misalignment between utilities may further reduce effectiveness. These factors collectively explain why multi-utility inputs do not consistently improve performance in this dataset.

Table 12 is the same setup as Table 11, but applies SMOTE to balance theft vs. normal samples. The key observations are as follows:

- Overall Accuracy: increased to 93.64
- Top Performers: quick-service restaurant (96.89%), full-service restaurant (96.08%), and strip mall (95.69%).
- Improvement: most classes show marginal gains in accuracy and recall; F1-Score improved to 0.877 overall.
- AUC: slightly higher than Table 11 (up to 0.997 for the top classes).

The insight offered is that SMOTE enhances robustness, especially for underrepresented theft cases, but does not fully resolve performance gaps for residential classes.

Table 11. Performance of the global Extra Trees model across consumer classes for binary electricity theft detection using multi-utility measurements (electricity and gas) without SMOTE.

Consumer Type	Accuracy	F1-Score	Precision	Recall	AUC	AUC-PR
FullServiceRestaurant	0.94978	0.90411	0.90395	0.90474	0.99273	0.93659
Hospital	0.94323	0.88914	0.89062	0.88936	0.99180	0.92974
LargeHotel	0.90758	0.83052	0.82846	0.83331	0.98331	0.86017
LargeOffice	0.92811	0.86232	0.85858	0.86759	0.98890	0.90741
MediumOffice	0.90621	0.82918	0.82955	0.83199	0.98356	0.87411
MidriseApartment	0.89848	0.80099	0.79918	0.81341	0.98082	0.82828
OutPatient	0.93134	0.86619	0.86564	0.87002	0.98994	0.91198
PrimarySchool	0.93645	0.88266	0.87743	0.88989	0.98910	0.91877
QuickServiceRestaurant	0.96029	0.92453	0.92806	0.92462	0.99460	0.95261
SecondarySchool	0.90676	0.84081	0.82825	0.86110	0.98316	0.87184
SmallHotel	0.88751	0.77597	0.77296	0.78532	0.97693	0.80335
SmallOffice	0.94897	0.89998	0.89938	0.90072	0.99189	0.93563
Stand-aloneRetail	0.95048	0.90240	0.90276	0.90263	0.99316	0.94651
StripMall	0.95247	0.90623	0.90513	0.90807	0.99314	0.94714
SuperMarket	0.94794	0.90083	0.90138	0.90192	0.99257	0.93198
Warehouse	0.90793	0.85538	0.83939	0.88173	0.98260	0.88030
ALL	0.92895	0.86689	0.86437	0.87283	0.98800	0.90218

Table 12. Performance of the global Extra Trees model across consumer classes for binary electricity theft detection using multi-utility measurements (electricity and gas) with SMOTE.

Consumer Type	Accuracy	F1-Score	Precision	Recall	AUC	AUC-PR
FullServiceRestaurant	0.96089	0.92275	0.92863	0.92292	0.99596	0.96103
Hospital	0.94353	0.88801	0.89278	0.88842	0.99302	0.93820
LargeHotel	0.91358	0.83756	0.83688	0.84334	0.98607	0.87895
LargeOffice	0.93256	0.87061	0.87521	0.86830	0.99158	0.92361
MediumOffice	0.91475	0.84357	0.84860	0.84253	0.98500	0.89138
MidriseApartment	0.89742	0.79583	0.79655	0.81161	0.98259	0.83856
OutPatient	0.93812	0.87837	0.88083	0.88283	0.99236	0.93264
PrimarySchool	0.94100	0.88856	0.89199	0.88759	0.98664	0.92993
QuickServiceRestaurant	0.96898	0.94033	0.94563	0.94045	0.99712	0.97414
SecondarySchool	0.92736	0.86441	0.86731	0.86392	0.98426	0.91448
SmallHotel	0.88720	0.77158	0.77054	0.78514	0.97791	0.80649
SmallOffice	0.95217	0.90556	0.90775	0.90513	0.99406	0.94577
Stand-aloneRetail	0.95555	0.91104	0.91121	0.91249	0.99481	0.95508
StripMall	0.95693	0.91352	0.91481	0.91589	0.99503	0.95622
SuperMarket	0.95869	0.92011	0.92412	0.92171	0.99540	0.95773
Warehouse	0.93457	0.88232	0.88351	0.88387	0.98231	0.91846
ALL	0.93642	0.87706	0.87970	0.87970	0.98963	0.92006

Figure 6 shows a grouped line chart comparing accuracy, F1-Score, precision, recall, and AUC for each consumer class before and after SMOTE. The highlights are as follows:

- Clear Gains: SMOTE improves recall and F1-Score for most classes by increasing theft samples.
- Consistent Leaders: quick-service restaurant and full-service restaurant dominate across all metrics.
- Persistent Challenges: midrise apartment and small hotel remain the lowest performers even after balancing.
- The contrast between near-perfect performance for commercial classes (e.g., restaurants achieving $\approx$ 98% accuracy) and substantially lower performance for residential-like classes (e.g., apartments achieving $\approx$ 89%) reflects a fundamental difference in the underlying statistical structure of consumption behavior rather than a limitation of the modeling framework itself. Commercial buildings exhibit highly regular, operationally constrained load profiles, making theft-related perturbations more detectable even under simple feature subsets and classical ensemble models. In contrast, residential consumption is inherently stochastic, influenced by occupant behavior, irregular appliance usage, seasonal routines, and variable occupancy patterns. This variability amplifies intra-class noise and produces overlapping distributions between normal and theft samples, particularly under synthetic theft injections that do not account for real residential masking strategies. Consequently, the consumer-specific modeling approach does not "fail" in these classes; rather, it exposes the domain where conventional tabular classifiers and uniformly applied preprocessing pipelines are least effective. This result highlights an important direction for future work: residential detection requires temporally-aware models, richer behavioral features, and consumer-tailored balancing or augmentation strategies. The observed gap between commercial and residential classes therefore provides valuable diagnostic insight into where methodological refinements are most needed for real-world deployment.

The interpretation offered is that SMOTE improves overall detection reliability, but consumer-specific variability persists.

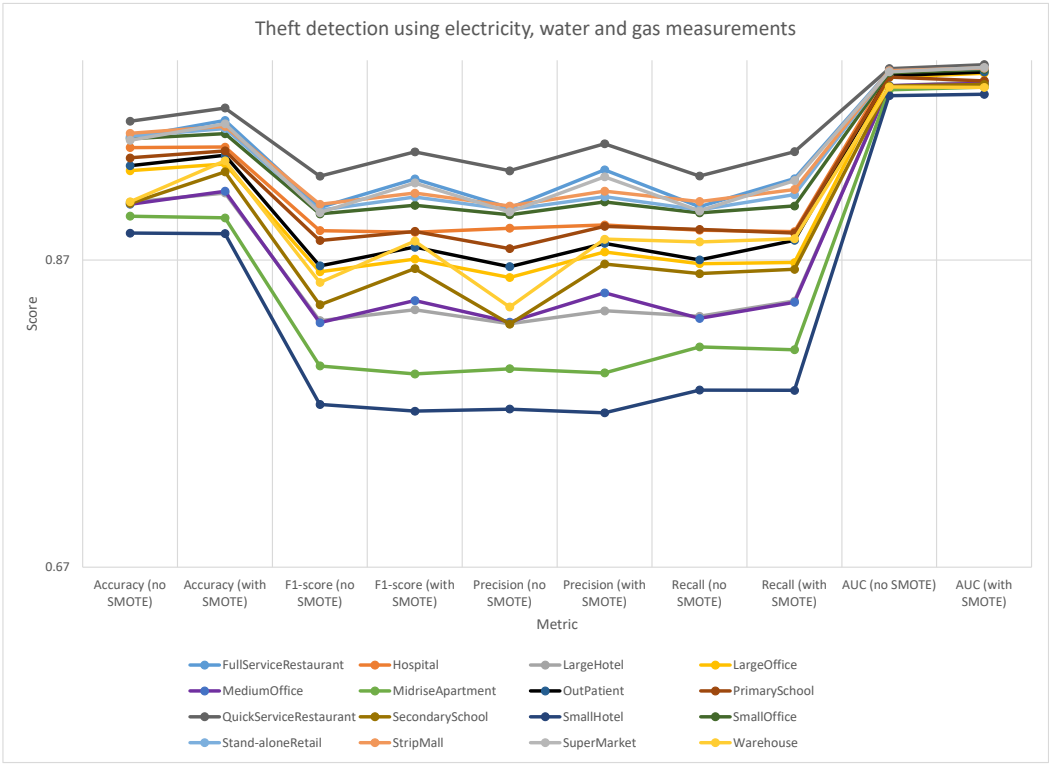

Figure 6. Performance comparison across consumer classes (accuracy, F1-Score, precision, recall, and AUC) using the global Extra Trees model for binary electricity theft detection with multi-utility measurements (electricity and gas), with and without SMOTE.

5.2. Using Consumer Specific Model

Having established the best consumer class using a globally good combination of feature selector and algorithm (LASSO-ET), this section looks at trying to find the best feature selector, feature selection and algorithm for each individual consumer class. Because each consumer-specific model uses a feature subset optimized for that class, cross-class comparisons of accuracy, F1-Score, or AUC cannot be interpreted as statistically controlled comparisons. The results should be understood as within-class improvements rather than evidence that certain classes are intrinsically easier or harder in an absolute sense. The two algorithms that were found to be most effective, namely, ET and RF, have separate compatibilities with SMOTE due to differences in the way the trees are evaluated and the SMOTE synthesis process, providing combination pairs that come into play for best performance. Table 13 reports the best model configuration per consumer class when training on electricity measurements only, without class balancing. For each class, we select the top-performing combination of algorithm and feature selector (with the number of selected features, N), and report accuracy, F1-score, precision, recall, and AUC:

- Overall performance: averaged over all 16 classes, the configuration attains accuracy = 0.9515, F1 = 0.9082, precision = 0.9078, recall = 0.9097, and AUC = 0.9931.
- Best vs. hardest classes: The quick-service restaurant (RF–ReliefF, N = 4) and full-service restaurant (ET–ReliefF, N = 4) are the easiest to detect (accuracy of 0.9778 and 0.9780, respectively; AUC $\geq$ 0.997). The midrise apartment class (ET–LASSO, N = 4) is the most challenging, with accuracy = 0.9118 and F1 = 0.8282.
- Algorithms and feature selection: Random Forest (RF) is the best choice in 10/16 classes (predominantly with ReliefF, N = 4), while Extra Trees (ET) leads in 6/16 classes (with ReliefF or LASSO). This split suggests that both tree ensembles are strong baselines,

with ReliefF typically selecting four electricity features; LASSO is preferred in a few classes (e.g., hospital, midrise apartment, small office, warehouse).

Without rebalancing, ensemble trees already deliver high discrimination (AUC ≈ 0.99) across segments. Classes with regular, business-like load profiles (restaurants, offices, retail) achieve top scores.

Table 13. Performance of consumer-specific models across consumer classes for binary electricity theft detection using electricity-only measurements without SMOTE.

Consumer Type	Best ML (FS) (N)	Accuracy	F1-Score	Precision	Recall	AUC	AUC-PR
FullServiceRestaurant	ET-RELIEFF (4)	0.97796	0.95871	0.96067	0.95808	0.99714	0.97988
Hospital	ET-LASSO (4)	0.95016	0.90257	0.90309	0.90346	0.99317	0.93802
LargeHotel	ET-RELIEFF (4)	0.93330	0.87521	0.87402	0.87721	0.98964	0.90107
LargeOffice	RF-RELIEFF (4)	0.94869	0.90445	0.90240	0.90805	0.99307	0.94345
MediumOffice	RF-RELIEFF (4)	0.94627	0.90138	0.90136	0.90231	0.99223	0.94150
MidriseApartment	ET-LASSO (4)	0.91179	0.82818	0.82623	0.83075	0.98378	0.85064
OutPatient	RF-RELIEFF (3)	0.94895	0.90490	0.90574	0.90538	0.99247	0.94130
PrimarySchool	RF-RELIEFF (4)	0.94826	0.90380	0.90104	0.90716	0.99413	0.94878
QuickServiceRestaurant	RF-RELIEFF (4)	0.97782	0.95816	0.96057	0.95735	0.99742	0.97883
SecondarySchool	ET-RELIEFF (4)	0.94756	0.90222	0.89860	0.90763	0.99391	0.93931
SmallHotel	ET-RELIEFF (4)	0.91800	0.84314	0.84141	0.84539	0.98556	0.86664
SmallOffice	RF-LASSO (4)	0.96697	0.93696	0.93757	0.93703	0.99663	0.97225
Stand-aloneRetail	RF-RELIEFF (4)	0.97069	0.94463	0.94419	0.94515	0.99599	0.97162
StripMall	RF-RELIEFF (4)	0.96916	0.94114	0.94218	0.94120	0.99666	0.97377
SuperMarket	RF-RELIEFF (4)	0.95985	0.92053	0.92163	0.92053	0.99519	0.95722
Warehouse	RF-LASSO (4)	0.94883	0.90542	0.90339	0.90831	0.99325	0.94998
ALL	–	0.95152	0.90821	0.90776	0.90969	0.99314	0.94089

Table 14 repeats the consumer-specific study after balancing theft vs. normal samples with SMOTE:

- Overall performance uplift: The average performance improves to accuracy = 0.9546 (+0.31 pp), F1 = 0.9127 (+0.45 pp), precision = 0.9145, recall = 0.9138, and AUC = 0.9942. Gains are modest but consistent, driven mainly by better recall on underrepresented theft patterns.
- Best vs. hardest classes: full-service restaurant (ET–ReliefF, N = 4) reaches accuracy = 0.9821, F1 = 0.9662, AUC = 0.9984; quick-service restaurant (ET–ReliefF, N = 4) reaches accuracy = 0.9804, F1 = 0.9628. Midrise apartment remains the hardest (accuracy = 0.9132, F1 = 0.8262), indicating that class balancing alone does not resolve residential variability.
- Algorithms and feature selection: With SMOTE, Extra Trees (ETs) becomes the best model in all 16 classes, most often paired with ReliefF (N = 4); a few classes use N = 3 (e.g., OutPatient, Stand-alone Retail). LASSO remains optimal for a subset (hospital, midrise apartment, small offfice, and warehouse), but ET remains the classifier of choice across the board.

SMOTE helps by increasing the number of theft samples, lifting recall and F1 while keeping AUC extremely high. The shift to ET as the universal winner suggests that, once the minority class is better represented, ET's stronger randomization and split diversity generalize more robustly across consumer profiles.

Table 14. Performance of consumer-specific models across consumer classes for binary electricity theft detection using electricity-only measurements with SMOTE.

Consumer Type	Best ML (FS) (N)	Accuracy	F1-Score	Precision	Recall	AUC	AUC-PR
FullServiceRestaurant	ET-RELIEFF (4)	0.98210	0.96621	0.96974	0.96530	0.99840	0.97988
Hospital	ET-LASSO (4)	0.95382	0.90799	0.91108	0.90881	0.99437	0.93802
LargeHotel	ET-RELIEFF (4)	0.93923	0.88518	0.88512	0.88745	0.99122	0.90107
LargeOffice	ET-RELIEFF (4)	0.95051	0.90749	0.90886	0.90769	0.99493	0.94245
MediumOffice	ET-RELIEFF (4)	0.94658	0.90130	0.90177	0.90317	0.99311	0.93612
MidriseApartment	ET-LASSO (4)	0.91316	0.82619	0.82357	0.83235	0.98485	0.85064
OutPatient	ET-RELIEFF (3)	0.95399	0.91319	0.91448	0.91476	0.99441	0.93780
PrimarySchool	ET-RELIEFF (4)	0.95450	0.91430	0.91726	0.91368	0.99444	0.94646
QuickServiceRestaurant	ET-RELIEFF (4)	0.98040	0.96280	0.96683	0.96193	0.99784	0.97476
SecondarySchool	ET-RELIEFF (4)	0.95247	0.91000	0.91412	0.90910	0.99407	0.93931
SmallHotel	ET-RELIEFF (4)	0.92088	0.84588	0.84456	0.85118	0.98698	0.86664
SmallOffice	ET-LASSO (4)	0.96865	0.93996	0.94124	0.93991	0.99728	0.97234
Stand-aloneRetail	ET-RELIEFF (3)	0.97251	0.94776	0.94827	0.94813	0.99715	0.97554
StripMall	ET-RELIEFF (4)	0.96658	0.93529	0.93954	0.93544	0.99712	0.97279
SuperMarket	ET-RELIEFF (4)	0.96606	0.93128	0.93600	0.93158	0.99630	0.95502
Warehouse	ET-LASSO (4)	0.95173	0.90895	0.90990	0.91096	0.99474	0.94821
ALL	–	0.95458	0.91274	0.91452	0.91384	0.99420	0.93981

Figure 7 visualizes, per consumer class, the accuracy, F1, precision, recall, and AUC achieved by the best tree ensemble—comparing RF and ET configurations under the consumer-specific setting. This figure highlights the cases where RF is dominant without SMOTE, but the opposite ensues with SMOTE, where ET is dominant. In the rest of the cases, either RF or ET is dominant for both with and without SMOTE. The plot emphasizes two complementary findings:

1. Without SMOTE: (Table 13), RF slightly edges ET in 10/16 classes; ET dominates the remaining 6/16. Both deliver AUC $\geq$ 0.99, but class-wise leaders flip depending on profile regularity and the selected feature subset (ReliefF vs. LASSO).
2. With SMOTE: (Table 14), ET consistently surpasses RF and becomes the top performer in all classes, with visible recall/F1 gains in previously underrepresented segments. The curves/bars for ET rise above RF across accuracy and F1, while AUC remains very high (often $\geq$ 0.997) for restaurant and retail classes.

The figure corroborates the tables: class balancing + ET provides the most reliable consumer-specific detector on electricity-only inputs, while RF is highly competitive when data remain imbalanced. The superior performance of Extra Trees (ET) combined with ReliefF compared to Random Forest (RF) with ReliefF when SMOTE is applied can be attributed to the inherent differences in tree construction and their interaction with synthetic data. SMOTE introduces interpolated minority samples that may form localized clusters. RF, which optimizes split points based on impurity reduction, tends to overfit these artificial clusters, reducing generalization. In contrast, ET selects split points randomly and uses the entire dataset without bootstrapping, creating more diverse decision boundaries and perhaps mitigating overfitting to synthetic patterns. This randomness may complement the diversity introduced by SMOTE, resulting in better bias-variance trade-off and improved detection performance. Additionally, ReliefF's instance-based feature ranking based on nearest neighbors benefits from the increased minority sample diversity, and ET's randomized splits leverage this effectively, leading to higher accuracy and F1-Scores across most consumer classes. It is emphasized that the underlying mechanism driving this improvement is not conclusively established in this study. We provide one possible interpretation

based on the known algorithmic differences between ET and RF, but further ablation and feature importance analyses are required to validate this hypothesis.

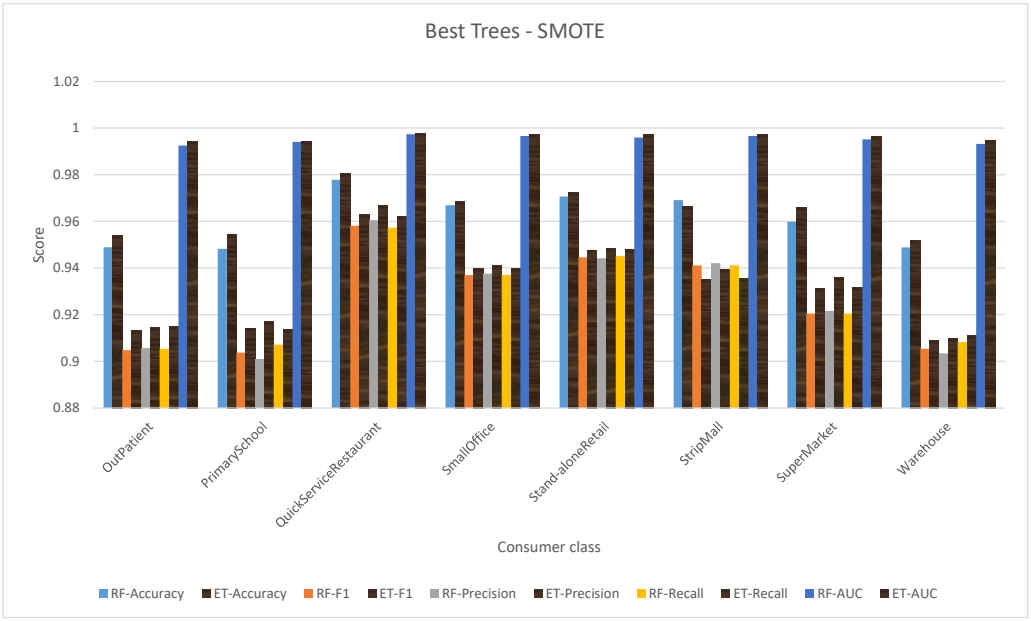

Figure 7. Comparison of Extra Trees and Random Forest Performance with and without SMOTE across consumer classes.

Table 15 shows the best-performing machine learning (ML) configuration for each consumer type when using electricity and gas measurements without applying SMOTE (data balancing). Each row corresponds to one of the 16 consumer classes (e.g., full-service restaurant, hospital, and large hotel). Columns include the following:

- Best ML (FS) (N): The optimal algorithm and feature selection method (e.g., RF–ReliefF or ET–LASSO) and number of selected features.
- Accuracy, F1-Score, precision, recall, AUC: Performance metrics for that configuration.

The key observations are:

- Full-service restaurant achieved the highest accuracy (99.07%) and F1-Score (0.9823) using RF–ReliefF.
- Midrise apartment was the hardest to detect (accuracy ≈ 91.18%, F1 ≈ 0.8282).
- Overall average accuracy across all classes: 94.97%; AUC values were very high ($\geq$0.98), indicating strong discrimination.

Table 16 has the same setup as Table 15 but with SMOTE applied to balance theft vs. normal samples. The key observations are:

- SMOTE generally improved performance slightly across most classes.
- Extra Trees (ET) became dominant for most consumer types when SMOTE was applied.
- The top performers are:
  1. Quick-service restaurant: accuracy = 98.04%, F1 = 0.9628.
  2. Full-service restaurant: accuracy = 98.41%, F1 = 0.9699.
- Overall average: accuracy = 95.35%, F1 = 0.9116, AUC ≈ 0.9938.

SMOTE improves recall and F1 for underrepresented theft cases, but residential classes (e.g., midrise apartment) remain challenging.

Table 15. Performance of consumer-specific models across consumer classes for binary electricity theft detection using multi-utility measurements (electricity and gas) without SMOTE.

Consumer Type	Best ML (FS) (N)	Accuracy	F1-Score	Precision	Recall	AUC	AUC-PR
FullServiceRestaurant	RF-RELIEFF (3)	0.99067	0.98232	0.98402	0.98168	0.99731	0.98980
Hospital	RF-RELIEFF (4)	0.93553	0.87789	0.87840	0.87815	0.98986	0.90997
LargeHotel	RF-LASSO (4)	0.92555	0.86166	0.86064	0.86287	0.98730	0.88356
LargeOffice	RF-RELIEFF (4)	0.94869	0.90445	0.90240	0.90805	0.99307	0.94345
MediumOffice	RF-RELIEFF (4)	0.93951	0.88859	0.88977	0.89052	0.99113	0.92948
MidriseApartment	ET-LASSO (4)	0.91179	0.82818	0.82623	0.83075	0.98378	0.85064
OutPatient	RF-RELIEFF (4)	0.95858	0.92383	0.92590	0.92371	0.99455	0.95548
PrimarySchool	RF-LASSO (4)	0.94050	0.89394	0.88690	0.90266	0.99212	0.93780
QuickServiceRestaurant	RF-RELIEFF (4)	0.97782	0.95816	0.96057	0.95735	0.99742	0.97883
SecondarySchool	RF-LASSO (4)	0.91536	0.85748	0.84264	0.88244	0.98982	0.90881
SmallHotel	ET-RELIEFF (4)	0.93812	0.88731	0.88764	0.88714	0.99077	0.90824
SmallOffice	RF-LASSO (4)	0.96164	0.92763	0.92847	0.92752	0.99478	0.95719
Stand-aloneRetail	RF-RELIEFF (4)	0.97069	0.94463	0.94419	0.94515	0.99599	0.97162
StripMall	ET-RELIEFF (3)	0.97509	0.95262	0.95401	0.95211	0.99718	0.98049
SuperMarket	ET-LASSO (4)	0.95591	0.91337	0.91344	0.91365	0.99446	0.95031
Warehouse	RF-LASSO (4)	0.94883	0.90542	0.90339	0.90831	0.99325	0.94998
ALL	–	0.94966	0.90675	0.90558	0.90953	0.99268	0.93788

Table 16. Performance of consumer-specific models across consumer classes for binary electricity theft detection using multi-utility measurements (electricity and gas) with SMOTE.

Consumer Type	Best ML (FS) (N)	Accuracy	F1-Score	Precision	Recall	AUC	AUC-PR
FullServiceRestaurant	ET-RELIEFF (3)	0.98409	0.96990	0.97456	0.96875	0.99784	0.98927
Hospital	ET-RELIEFF (4)	0.94117	0.88575	0.88828	0.88806	0.99248	0.90601
LargeHotel	RF-LASSO (4)	0.93270	0.87109	0.87270	0.87436	0.98996	0.88356
LargeOffice	ET-RELIEFF (4)	0.95051	0.90749	0.90886	0.90769	0.99493	0.94245
MediumOffice	ET-RELIEFF (4)	0.94827	0.90420	0.90612	0.90674	0.99352	0.92900
MidriseApartment	ET-LASSO (4)	0.91316	0.82619	0.82357	0.83235	0.98485	0.85064
OutPatient	ET-RELIEFF (4)	0.96545	0.93617	0.93871	0.93633	0.99671	0.95022
PrimarySchool	RF-LASSO (4)	0.94400	0.89644	0.90002	0.89588	0.99219	0.93780
QuickServiceRestaurant	ET-RELIEFF (4)	0.98040	0.96280	0.96683	0.96193	0.99784	0.97476
SecondarySchool	RF-LASSO (4)	0.93208	0.87338	0.87822	0.87289	0.98695	0.90881
SmallHotel	ET-RELIEFF (4)	0.93843	0.88642	0.88856	0.88647	0.99182	0.90824
SmallOffice	ET-LASSO (4)	0.96362	0.93118	0.93200	0.93116	0.99665	0.95393
Stand-aloneRetail	ET-RELIEFF (3)	0.97251	0.94776	0.94827	0.94813	0.99715	0.97554
StripMall	ET-RELIEFF (3)	0.97159	0.94529	0.94912	0.94489	0.99707	0.98049
SuperMarket	ET-LASSO (4)	0.96606	0.93155	0.93648	0.93145	0.99671	0.95031
Warehouse	ET-LASSO (4)	0.95173	0.90895	0.90990	0.91096	0.99474	0.94821
ALL	–	0.95350	0.91156	0.91390	0.91240	0.99384	0.93685

Figure 8 shows the best performance achievable with models tested in this study. Table 14, which holds the best results for electricity data (using SMOTE), is compared with Table 16, which holds the best results for electricity and gas data (using SMOTE). The metrics used are accuracy, F1-Score, precision, recall, AUC. Some key notes are:

- Electricity-only models generally outperform multi-utility models for most consumer types.
- Restaurant and retail classes (e.g., quick-service restaurant, full-service restaurant) consistently achieve the highest scores across all metrics.
- Residential-like classes (e.g., midrise apartment) remain the lowest performers in both scenarios.

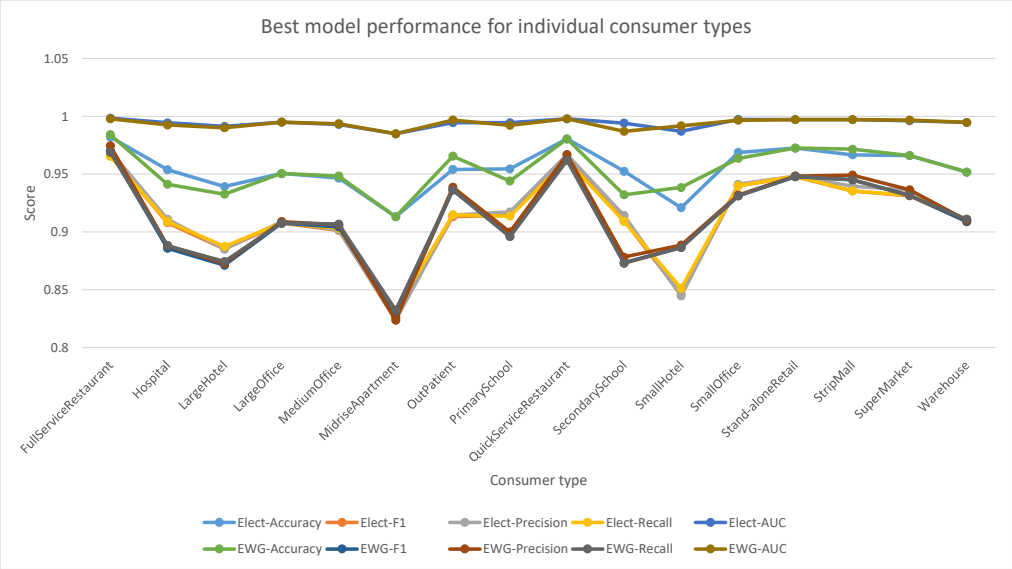

Figure 8. Best model performance per consumer class for electricity-only vs. multi-utility configurations.

Although multi-utility inputs are conceptually valuable, the current experimental results do not demonstrate a performance gain. Adding gas data does not always improve detection; electricity-only models with SMOTE often yield better results. This is likely due to limitations of the TDD2022 dataset's synthetic theft patterns and the linear/non-linear feature selectors used. Future work must explore cross-modal correlation modeling, real theft data, and advanced fusion architectures before drawing conclusions about the potential of multi-utility sensing. We cite prior work showing that multi-source data helps in other contexts. Studies such as [20] have already demonstrated that integrating electricity and gas can improve theft detection when the modalities are fully exploited with specialized cross-modal architectures (e.g., graph-based and correlation-aware models). Our manuscript positions our multi-utility experiments as baseline tests rather than as confirmation of cross-modal benefits.

It should be noted that the evaluation in this study is based on stratified random splitting of the dataset. While this ensures balanced class representation, it may introduce optimistic performance estimates in time-series data due to potential temporal dependencies between training and testing samples.

Given the synthetic and block-based nature of theft patterns in the TDD2022 dataset, this approach enables controlled and consistent comparison across models. However, more realistic evaluation strategies, such as time-based splitting (training on earlier periods and testing on later periods) or grouped splitting per consumer, would better reflect real-world deployment scenarios. Exploring these evaluation settings is therefore identified as an important direction for future work.

6. Conclusions

This study demonstrates that machine learning models, particularly tree-based ensembles such as Extra Trees (ET) and Random Forest (RF), are highly effective for electricity theft detection in smart grid environments. Using the TDD2022 dataset, this research evaluated scenarios involving electricity-only data and multi-utility data (electricity and gas), with and without SMOTE for class balancing. Since TDD2022 uses algorithmically generated theft patterns that are structurally simpler than real-world theft behaviors, the high accuracy values in this study should not be interpreted as estimates of real deployment performance. Real theft detection requires datasets that include authentic behavioral anomalies, partial bypassing, AMI manipulation, and load camouflage, which synthetic

datasets do not fully capture. The evaluated models will be incorporated into the OMMU microgrid's real-time decision-support system as the theft-detection component shown in Figure 1. Key findings include:

- Electricity-only models generally outperform multi-utility models, achieving higher accuracy and F1-Scores.
- Across the global theft detection task, the application of SMOTE improved overall minority class performance, yielding higher recall and F1 scores by increasing the density of theft samples during training. These gains demonstrate that class balancing enhances the model's ability to identify theft events at the aggregate level. However, the improvements were not uniform across all consumer categories. Residential-like classes, characterized by high behavioral variability, showed limited or negligible benefit from resampling, indicating that SMOTE alone is insufficient for these segments. This highlights the need for more advanced, consumer-specific balancing or augmentation strategies in future work. Overall, the results confirm that while SMOTE is an effective baseline technique for strengthening global model sensitivity, targeted approaches remain essential for the most challenging consumer types.
- The differing feature subsets produced by LASSO and ReliefF highlight the heterogeneous nature of electricity-theft signatures across consumer types. LASSO favors globally linear, additive relationships and therefore consistently identifies core electricity channels as the most informative predictors. ReliefF, in contrast, captures local non-linear interactions and occasionally elevates additional modalities, reflecting class-dependent variability in consumption behavior. This divergence underscores that no single feature-selection method fully characterizes the range of theft patterns present in the dataset. Accordingly, future work should explore hybrid or stability-aware feature-selection frameworks that integrate both linear and non-linear criteria to improve robustness, particularly for consumer classes with complex or highly variable load profiles.
- Consumer-specific models significantly outperform global models, with restaurant and retail classes achieving near-perfect detection, while residential classes remain harder to classify.
- Extra Trees combined with ReliefF feature selection emerges as the most reliable configuration when SMOTE is applied due to its ability to generalize well on synthetic minority samples.

Overall, this research highlights that consumer-specific modeling and appropriate feature selection are critical for improving detection accuracy. Future work will evaluate hybrid feature selection pipelines, including stability selection and multi-stage selectors that combine LASSO's linear sparsity with ReliefF's non-linear neighborhood sensitivity. These approaches are well-suited to the correlated and heterogeneous nature of multi-utility smart meter data and may provide more robust feature rankings than single-run selectors. More globally, future work should focus on addressing variability in residential consumption patterns, integrating real-world theft data, and exploring advanced techniques such as federated learning and explainable AI for scalable and privacy-preserving deployment.

Author Contributions: Conceptualization, formal analysis, validation, software, writing—original draft, writing—review and editing, and methodology, F.A. and H.G.; conceptualization and writing—review and editing, V.M., Z.B.S., I.B., P.S. and P.T.; conceptualization, writing—review and editing, methodology, and supervision, I.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially supported by the project "OMMU: Optimised Microgrid Management in Ukraine" funded by Innovate UK with project number 10092144.

Data Availability Statement: https://data.mendeley.com/datasets/c3c7329tjj/1, accessed on 15 December 2025.

Conflicts of Interest: Authors Zaid Bin Saeed and Pouya Tarassodi were employed by the Innvotek Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Northeast Group, LLC. $96 Billion Is Lost Every Year to Electricity Theft. Available online: https://www.prnewswire.com/news-releases/96-billion-is-lost-every-year-to-electricity-theft-300453411.html (accessed on 5 January 2025).
2. Sadovskaia, K.; Bogdanov, D.; Honkapuro, S.; Breyer, C. Power Transmission and Distribution Losses—A Model Based on Available Empirical Data and Future Trends for All Countries Globally. *Int. J. Electr. Power Energy Syst.* **2019**, *107*, 98–109. [CrossRef]
3. Carr, D.; Thomson, M. Non-Technical Electricity Losses. *Energies* **2022**, *15*, 2218. [CrossRef]
4. Baffi, E.; Urioste, R.M.L.; Peñalba, M.A. Potential Benefits of Distributed Generation in the Reduction of Non-Technical Losses. *Renew. Energy Power Qual. J.* **2018**, *16*, 39–44. [CrossRef]
5. Kang, L.; Shang, Y.; Zhang, M.; Liao, L. Research on Monitoring Technology of Power Stealing Behavior in Bitcoin Mining Based on Analyzing Electric Energy Data. *Energy Rep.* **2022**, *8*, 1183–1189. [CrossRef]
6. Es'haghi, A.S.; Afjei, E.; Marini, A.; Karimi, M. Detection of Illegal Cryptocurrency Mining Farms in Distribution Systems Using Harmonic State Estimation. In Proceedings of the 2023 13th Smart Grid Conference (SGC), Tehran, Iran, 5–6 December 2023; pp. 1–6. [CrossRef]
7. Wong, H.L.; Tan, C.K.; Tan, W.N.; Gan, M.T.; Yip, S.C.; Bakar, A.H.A. Energy Theft Identification: A State Estimation Model for Partial Meter Bypass. *IEEE Access* **2025**, *13*, 121412–121431. [CrossRef]
8. Plecas, D.; Diplock, J.; Garis, L.; Carlisle, B.; Neal, P.; Landry, S. The Marihuana Indoor Production Calculator: A Tool for Estimating Domestic and Export Production Levels and Values. *J. Crim. Justice Res.* **2010**, *1*, 1–12.
9. Arango, L.; Deccache, E.; Bonatto, B.D.; Arango, H.; Ribeiro, P.; Silveira, P.M. Impact of Electricity Theft on Power Quality. In Proceedings of the 2016 17th International Conference on Harmonics and Quality of Power (ICHQP), Belo Horizonte, Brazil, 16–19 October 2016; pp. 557–562. [CrossRef]
10. Gao, Y.; Foggo, B.; Yu, N. A Physically Inspired Data-Driven Model for Electricity Theft Detection with Smart Meter Data. *IEEE Trans. Ind. Inform.* **2019**, *15*, 5076–5088. [CrossRef]
11. Li, D.; Yang, Q.; Zhang, F.; Wang, Y.; Qian, Y.; An, D. Research on Privacy Issues in Smart Metering System: An Improved TCN-Based NILM Attack Method and Practical DRL-Based Rechargeable Battery Assisted Privacy Preserving Method. *IEEE Trans. Autom. Sci. Eng.* **2024**, *21*, 2882–2899. [CrossRef]
12. Innovate Ukraine. OMM-Ukraine: Optimised Microgrid Management Empowering Off-Grid Communities with Access to Clean Energy—Innovate Ukraine. Available online: https://innovateukraine.io/case_study/case-study-1/ (accessed on 5 January 2025).
13. Gupta, A.K.; Routray, A.; Naikan, V.A. Detection of Power Theft in Low Voltage Distribution Systems: A Review from the Indian Perspective. *IETE J. Res.* **2022**, *68*, 4180–4197. [CrossRef]
14. Badr, M.M.; Mahmoud, M.M.E.A.; Abdulaal, M.; Aljohani, A.J.; Alsolami, F.; Balamsh, A. A Novel Evasion Attack Against Global Electricity Theft Detectors and a Countermeasure. *IEEE Internet Things J.* **2023**, *10*, 11038–11053. [CrossRef]
15. Lamb, M. Advanced Metering Infrastructure: Continued Evolution and Opportunities to Deliver Greater Value. *Clim. Energy* **2025**, *41*, 14–20. [CrossRef]
16. Jokar, P.; Arianpoo, N.; Leung, V.C.M. Electricity Theft Detection in AMI Using Customers' Consumption Patterns. *IEEE Trans. Smart Grid* **2016**, *7*, 216–226. [CrossRef]
17. Zidi, S.; Mihoub, A.; Qaisar, S.M.; Krichen, M.; Al-Haija, Q.A. Theft Detection Dataset for Benchmarking and Machine Learning Based Classification in a Smart Grid Environment. *J. King Saud Univ. Comput. Inf. Sci.* **2022**, *35*, 13–25. [CrossRef]
18. Mohammad, F.; Saleem, K.; Al-Muhtadi, J. Ensemble-Learning-Based Decision Support System for Energy-Theft Detection in Smart-Grid Environment. *Energies* **2023**, *16*, 1907. [CrossRef]
19. Reyes, O.; Morell, C.; Ventura, S. Scalable Extensions of the ReliefF Algorithm for Weighting and Selecting Features on the Multi-Label Learning Context. *Neurocomputing* **2015**, *161*, 168–182. [CrossRef]
20. Liao, W.; Zhu, R.; Yang, Y.; Jia, Y.; Yang, Z.; Rehtanz, C. Electricity Theft Detection with Multi-Source Data: Integrating Electricity, Water, and Gas. *IEEE Trans. Ind. Appl.* **2025**, *62*, 3187–3197. [CrossRef]
21. Hou, Z.; Liu, J. Enhancing Smart Grid Sustainability: Using Advanced Hybrid Machine Learning Techniques While Considering Multiple Influencing Factors for Imputing Missing Electric Load Data. *Sustainability* **2024**, *16*, 8092. [CrossRef]

22. Liu, J.; Hou, Z.; Wang, B.; Yin, T. Optimizing Microgrid Energy Management via DE-HHO Hybrid Metaheuristics. *Comput. Mater. Contin.* **2025**, *84*, 4729–4754. [CrossRef]
23. Hou, Z.; Liu, J.; Shao, Z.; Ma, Q.; Liu, W. Machine Learning Innovations in Renewable Energy Systems with Integrated NRBO-TXAD for Enhanced Wind Speed Forecasting Accuracy. *Electronics* **2025**, *14*, 2329. [CrossRef]
24. Liao, W.; Zhu, R.; Yang, Z.; Liu, K.; Zhang, B.; Zhu, S.; Feng, B. Electricity Theft Detection Using Dynamic Graph Construction and Graph Attention Network. *IEEE Trans. Ind. Inform.* **2024**, *20*, 5074–5086. [CrossRef]
25. Zheng, Z.; Yang, Y.; Niu, X.; Dai, H.N.; Zhou, Y. Wide and Deep Convolutional Neural Networks for Electricity-Theft Detection to Secure Smart Grids. *IEEE Trans. Ind. Inform.* **2018**, *14*, 1606–1615. [CrossRef]
26. Duan, J. Deep Learning Anomaly Detection in AI-powered Intelligent Power Distribution Systems. *Front. Energy Res.* **2024**, *12*, 1364456. [CrossRef]
27. Nevisi, M.M.S.; Shoeibi, M.; Hernando-Gallego, F.; Martín, D.; Khatami, S.S. An Evolutionary Deep Reinforcement Learning-Based Framework for Efficient Anomaly Detection in Smart Power Distribution Grids. *Energies* **2025**, *18*, 2435. [CrossRef]
28. Badr, M.M.; Ibrahem, M.I.; Mahmoud, M.; Fouda, M.M.; Alasmary, W. Detection of False-Reading Attacks in the AMI Net-Metering System. *arXiv* **2020**, arXiv:2012.01983. [CrossRef]
29. Kolter, J.Z.; Johnson, M.J. REDD: A Public Data Set for Energy Disaggregation Research. In Proceedings of the Workshop on Data Mining Applications in Sustainability (SIGKDD), San Diego, CA, USA, 21–24 August 2011.
30. Khalid, A.; Mustafa, G.; Rana, M.R.R.; Alshahrani, S.M.; Alymani, M. RNN-BiLSTM-CRF Based Amalgamated Deep Learning Model for Electricity Theft Detection to Secure Smart Grids. *PeerJ Comput. Sci.* **2024**, *10*, e1872. [CrossRef]
31. Blazakis, K.; Schetakis, N.; Badr, M.M.; Aghamalyan, D.; Stavrakakis, K.; Stavrakakis, G. Power Theft Detection in Smart Grids Using Quantum Machine Learning. *IEEE Access* **2025**, *13*, 61511–61525. [CrossRef]
32. Islam Sajol, M.S.; Ahmed, I.; Mahmud, Q.S. Synthetic Minority Oversampling Technique Enhanced Machine Learning Models for Energy Theft Detection. In Proceedings of the 2024 IEEE Kansas Power and Energy Conference (KPEC), Manhattan, KS, USA, 25–26 April 2024; pp. 1–6. [CrossRef]
33. Mohammad, F.; Al-Ahmadi, S.; Al-Muhtadi, J. RoGRUT: A Hybrid Deep Learning Model for Detecting Power Trapping in Smart Grids. *Comput. Mater. Contin.* **2024**, *79*, 3175–3192. [CrossRef]
34. Abbas, S.; Bouazzi, I.; Ojo, S.; Sampedro, G.A.; Almadhor, A.S.; Hejaili, A.A.; Stolicna, Z. Improving Smart Grids Security: An Active Learning Approach for Smart Grid-Based Energy Theft Detection. *IEEE Access* **2024**, *12*, 1706–1717. [CrossRef]
35. Gunduz, M.Z.; Das, R. Smart Grid Security: An Effective Hybrid CNN-Based Approach for Detecting Energy Theft Using Consumption Patterns. *Sensors* **2024**, *24*, 1148. [CrossRef] [PubMed]
36. Shahid, S.S.; Salman, T.; Baza, M.; Srivastava, G. Efficient Energy Theft Detection Utilizing Hierarchical Federated Learning. In Proceedings of the 2025 International Conference on Smart Applications, Communications and Networking (SmartNets), Istanbul, Turkiye, 22–24 July 2025; pp. 1–6. [CrossRef]
37. Shahid, S.S.; Salman, T.; Baza, M. Mitigating Gradient Inversion Attacks in Energy Theft Using Hierarchical Federated Learning. In Proceedings of the 2025 1st International Conference on Secure IoT, Assured and Trusted Computing (SATC), Dayton, OH, USA, 25–27 February 2025; pp. 1–5. [CrossRef]
38. Satyapal, K.S.; Patil, A. Enhancing the Electricity Theft Detection Using Extreme Gradient Boosting with Optuna Optimization in Smart Grid. In Proceedings of the 2024 IEEE PES Innovative Smart Grid Technologies—Asia (ISGT Asia), Bengaluru, India, 10–13 November 2024; pp. 1–6. [CrossRef]
39. Saqib, S.M.; Mazhar, T.; Iqbal, M.; Shahzad, T.; Almogren, A.; Ouahada, K.; Hamam, H. Deep Learning-Based Electricity Theft Prediction in Non-Smart Grid Environments. *Heliyon* **2024**, *10*, e35167. [CrossRef]
40. Bai, W.; Xiong, L.; Liao, Y.; Tan, Z.; Wang, J.; Zhang, Z. Detection Method for Three-Phase Electricity Theft Based on Multi-Dimensional Feature Extraction. *Sensors* **2024**, *24*, 6057. [CrossRef]
41. Hashim, M.; Khan, L.; Javaid, N.; Ullah, Z.; Javed, A. Stacked Machine Learning Models for Non-Technical Loss Detection in Smart Grid: A Comparative Analysis. *Energy Rep.* **2024**, *12*, 1235–1253. [CrossRef]
42. Yan, Z.; Wen, H. Electricity Theft Detection Base on Extreme Gradient Boosting in AMI. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–9. [CrossRef]
43. Badr, M.M.; Ibrahem, M.I.; Mahmoud, M.; Fouda, M.M.; Alsolami, F.; Alasmary, W. Detection of False-Reading Attacks in Smart Grid Net-Metering System. *IEEE Internet Things J.* **2022**, *9*, 1386–1401. [CrossRef]
44. Patsidis, A.; Dyśko, A.; Booth, C.; Rousis, A.O.; Kalliga, P.; Tzelepis, D.; Patsidis, A.; Dyśko, A.; Booth, C.; Rousis, A.O.; et al. Digital Architecture for Monitoring and Operational Analytics of Multi-Vector Microgrids Utilizing Cloud Computing, Advanced Virtualization Techniques, and Data Analytics Methods. *Energies* **2023**, *16*, 5908. [CrossRef]
45. Liu, Y.; Wang, Y.; Ma, J. Non-Intrusive Load Monitoring in Smart Grids: A Comprehensive Review. *arXiv* **2024**, arXiv:2403.06474. [CrossRef]
46. Bouquet, P.; Jackson, I.; Nick, M.; Kaboli, A. AI-based Forecasting for Optimised Solar Energy Management and Smart Grid Efficiency. *Int. J. Prod. Res.* **2024**, *62*, 4623–4644. [CrossRef]

47. Hernández-Mayoral, E.; Madrigal-Martínez, M.; Mina-Antonio, J.D.; Iracheta-Cortez, R.; Enríquez-Santiago, J.A.; Rodríguez-Rivera, O.; Martínez-Reyes, G.; Mendoza-Santos, E.; Hernández-Mayoral, E.; Madrigal-Martínez, M.; et al. A Comprehensive Review on Power-Quality Issues, Optimization Techniques, and Control Strategies of Microgrid Based on Renewable Energy Sources. *Sustainability* **2023**, *15*, 9847. [CrossRef]
48. Zulu, M.L.T.; Carpanen, R.P.; Tiako, R.; Zulu, M.L.T.; Carpanen, R.P.; Tiako, R. A Comprehensive Review: Study of Artificial Intelligence Optimization Technique Applications in a Hybrid Microgrid at Times of Fault Outbreaks. *Energies* **2023**, *16*, 1786. [CrossRef]
49. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
50. Charbuty, B.; Abdulazeez, A. Classification Based on Decision Tree Algorithm for Machine Learning. *J. Appl. Sci. Technol. Trends* **2021**, *2*, 20–28. [CrossRef]
51. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [CrossRef]
52. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
53. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased Boosting with Categorical Features. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2018; Volume 31.
54. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely Randomized Trees. *Mach. Learn.* **2006**, *63*, 3–42. [CrossRef]
55. Appiah, S.Y.; Akowuah, E.K.; Ikpo, V.C.; Dede, A. Extremely Randomised Trees Machine Learning Model for Electricity Theft Detection. *Mach. Learn. Appl.* **2023**, *12*, 100458. [CrossRef]
56. Zhao, P.; Yu, B. On Model Selection Consistency of Lasso. *J. Mach. Learn. Res.* **2006**, *7*, 2541–2563.
57. Meinshausen, N.; Bühlmann, P. Stability Selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2010**, *72*, 417–473. [CrossRef]
58. Hastie, T.; Tibshirani, R.; Wainwright, M. *Statistical Learning with Sparsity: The Lasso and Generalizations*; Chapman and Hall/CRC: New York, NY, USA, 2015. [CrossRef]
59. Bolón-Canedo, V.; Sánchez-Maroño, N.; Alonso-Betanzos, A. A Review of Feature Selection Methods on Synthetic Data. *Knowl. Inf. Syst.* **2013**, *34*, 483–519. [CrossRef]