IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 7 January 2024, accepted 1 February 2024, date of publication 14 February 2024, date of current version 23 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3366493

RESEARCH ARTICLE

Electricity Theft Detection for Smart Homes: Harnessing the Power of Machine Learning With Real and Synthetic Attacks

OLUFEMI ABIODUN ABRAHAM[1], **(Graduate Student Member, IEEE),**
HIDEYA OCHIAI[2], **(Member, IEEE), MD. DELWAR HOSSAIN**[1], **(Member, IEEE),**
YUZO TAENAKA[1], **(Member, IEEE), AND YOUKI KADOBAYASHI**[1], **(Member, IEEE)**
[1]Division of Information Science, Graduate School of Science and Technology, Nara Institute of Science and Technology, Ikoma, Nara 630-0192, Japan
[2]Graduate School of Information Science and Technology, The University of Tokyo, Tokyo 113-8654, Japan

Corresponding author: Olufemi Abiodun Abraham (abraham.olufemi_abiodun.aj9@naist.ac.jp)

This work was partly funded by the Industrial Cyber Security Centre of Excellence (ICSCoE) Core Human Resources Development Program, Japan.

ABSTRACT Electricity theft is a pervasive issue with economic implications that necessitate innovative approaches for its detection, given the critical challenge of limited labeled data. However, connecting smart home devices introduces numerous vectors for electricity theft. Therefore, this study introduces an innovative approach to detecting electricity theft in smart homes, leveraging knowledge-based, fine-grained, time-series appliance benign and anomalous consumption patterns. We simulated five attack classes and extended our model's detection capabilities to unknown anomalies across residential settings by segmenting the anonymized data into three different home categories. We validated our experiment using simulated and real building attack data. Extreme Gradient Boost (XGB), Random Forest, and Multilayer Perceptron (MLP) outperform the legacy unsupervised model (LUM), which included MLP-Autoencoder (AE), 1D-CONV-AE, and Isolation Forest (RF). XGB had the highest average AUC scores of 98.69% and 98.74% for simulated and real attack detection, respectively, followed by RF at 96.76% and 97.07%, respectively, across all homes, indicating the robustness of our model in detecting benign and anomalous appliance consumption patterns. This study contributes to the academic discourse in the field and offers practical solutions to energy providers and stakeholders in the smart home industry.

INDEX TERMS Electricity theft detection, machine learning, synthetic attack data, smart home, real attack data, unsupervised learning, supervised learning.

I. INTRODUCTION

Electricity theft is a pervasive issue with economic implications that, necessitate innovative detection approaches. Although invisible, it possesses substantial economic value and is an essential resource for modern society. The issue of electricity theft, often referred to as Electricity Theft Attacks (ETAs), represents a pervasive and costly problem worldwide, particularly in developing countries [1], [2]. For instance, in India, it is estimated that more than one-fifth of the total electricity production is lost owing to theft [3]. Moreover, within the United States, these losses are estimated to be approximately $6 billion annually [4], [5]. Traditionally, the detection of ETAs relies on human intervention through routine inspections of power meters and cables, user-reported suspicions, or the identification of meter device failures [6]. Nevertheless, these methods are human-centric and struggle to effectively pinpoint instances of electricity theft, primarily because of the inherent complexity of power systems.

Each electrical appliance exhibits a distinctive consumption signature in the time series data, which is aggregated with the consumption patterns of other appliances at the metering point [16]. Consequently, the output of a smart meter encapsulates the collective patterns of electricity consumed

The associate editor coordinating the review of this manuscript and approving it for publication was Arash Asrari.

TABLE 1. Literature review.

Reference	Method	Contributions	Limitations	Our Model's Resolution
Punmiya et al., 2019	Simple statistical models for synthetic attack loads	Initial frameworks for synthetic load generation and attack simulation.	Limited complexity and realism in attack scenarios.	Novel attack scenarios that simulate realistic electricity theft patterns.
Abraham et al., 2023	Knowledge-based synthetic attack generation	Understanding the intentions behind various attacks.	Lack of empirical validation with real attack data.	Validation with real attack data for binary classification and anomaly detection.
Dayaratne et al., 2021; Singh et al., 2017; Punmiya et al., 2019	Protection of utility companies from electricity theft	Large-scale protection to prevent revenue loss from power stealing.	Not addressing power theft between individual homes.	Detection and protection against electricity theft at the household level.
Jokar et al., 2015	Limited Reduction Techniques	Differentiation between benign and attack data in binary classification models.	A gap in knowledge regarding the utilization of appliance consumption patterns for ETD.	Proposes a classifier that uses reduced datasets for efficient and effective theft detection.
Moradzadeh et al., 2020	Improving residential load disaggregation	Utilized PCA for sustainable energy development.	Did not focus on electricity theft detection.	Utilization of appliance consumption patterns for theft detection.
Singh & Yassine, 2018	Big data mining for energy consumption forecasting	Advanced forecasting using big data mining techniques.	Lack of application to electricity theft detection.	Data mining of usage patterns for electricity theft detection.
Wang & Yang, 2018	Iterative load disaggregation	Addressing complexity in appliance-specific patterns.	Focused on disaggregation, not theft detection.	Incorporation of consumption patterns to distinguish theft behavior.
Wilhelm & Kasbauer, 2021	NILM for human activity recognition	Exploited meter measurements for activity recognition.	Did not apply findings to theft detection.	NILM techniques to identify theft-related irregularities.
Zhang et al., 2014	Simulation of human living activities	Transformation between activities and power patterns.	Focus was not on theft detection.	Relationship between activities and consumption used for theft detection.

by legitimate and potentially malicious loads [17]. This aggregation poses a challenge for the detection of electricity theft through traditional heuristic analysis of meter readings alone [18]. However, with the rapid advancement of machine learning technologies, there is a burgeoning opportunity to harness the power of artificial intelligence to detect electricity theft from aggregated consumption patterns.

This study introduces an innovative Electricity Theft Detection (ETD) framework tailored for smart homes, representing a notable departure from traditional approaches that integrate knowledge-based synthetic attack data. Our approach is particularly well-suited for handling unlabeled datasets, which are commonly addressed using unsupervised learning techniques [19]. Although synthetic data generation is often considered a form of supervised learning owing to the creation of labels through a synthetic approach [20], [21], it applies to unsupervised methods when dealing with unlabeled data. However, our research demonstrates the superior performance of our approach compared with conventional unsupervised learning techniques. This finding underscores the primary message conveyed in the present study.

A. CHALLENGES AND MOTIVATION

One of the key challenges in deploying machine learning classifiers for ETD in smart homes is the pervasive issue of data imbalance [22], [23], where the volumes of normal and abnormal samples exhibit substantial disparities. While benign samples are readily accessible through historical data, attack or theft samples are often scarce, and, in some cases, entirely absent for certain customers.

To address this dilemma, we addressed the issue of imbalanced data and zero-day attacks [24] through the creation of a synthetic attack dataset. Leveraging the inherent predictability of theft patterns [5], this innovative approach significantly elevates the detection rate and empowers the identification of a diverse range of attack types.

The prospect of training an ETD for smart homes machine learning model using synthetic attack data offers a range of advantages. Firstly, it circumvents the need for extensive time and resource investments, which would otherwise be required to acquire authentic attack consumption patterns alongside legitimate patterns in real-life scenarios. This cost-effective approach also allows for the evaluation of synthetic attack scenarios with real data of a similar nature. Secondly, the calculation of attack-contained meter readings becomes straightforward by simply adding synthetic malicious consumption to legitimate usage. Such a practical application was previously unattainable in computer network communications, where node behavior hinged on the presence of attacks. Additionally, the flexibility of synthetic attack data, drawing from a broad spectrum of knowledge, holds the potential to significantly enhance the training of robust ETD models for precise attack classifications.

B. LITERATURE REVIEW

Table 1 shows some available literature on appliance consumption, however, there is a significant gap in knowledge regarding the utilization of appliance consumption patterns to train a classifier for electricity theft detection in smart homes. Although there is extensive research on appliance consumption patterns [14], [25], and [15], there is a lack of specific focus on using these patterns to train a classifier for electricity theft detection. Existing literature primarily discusses load disaggregation, human activity recognition, and energy consumption forecasting based on appliance power consumption patterns. However, the direct application of these patterns to train a classifier for electricity theft detection has not been explicitly addressed.

The literature provides insights into the challenges associated with appliance consumption patterns, such as the complexity of identifying appliance-specific consumption patterns and overlapping operation of appliances, which makes event detection difficult [13].

Additionally, some studies discuss the application of appliance power consumption patterns for simulating human living activities [15] and improving residential load disaggregation [11]. Furthermore, some studies have emphasized the accuracy of identifying appliance usage patterns using the proposed models [14], [25].

However, the specific task of using appliance consumption patterns to train a classifier for electricity theft detection in smart homes remains underexplored. The references did not directly address the development of a classifier for detecting electricity theft based on appliance consumption patterns. Therefore, there is a clear gap in existing knowledge regarding this specific application.

Although the literature provides valuable insights into appliance consumption patterns and their applications, there is a notable gap in knowledge concerning the direct utilization of these patterns to train a classifier for electricity theft detection in smart homes.

1) RESEARCH GAPS

Our analysis reveals several key areas in which existing research falls short, necessitating further investigation and innovation.

a: LIMITED SCOPE OF REDUCTION TECHNIQUES

Previous studies, exemplified by the work of [5], have primarily focused on basic reduction techniques such as random, mean, and reverse-time-based methods for differentiating between normal and anomalous data in binary classification models. Although foundational, these approaches offer limited complexity in capturing the nuanced dynamics of electricity theft.

b: GAP IN REALISM AND VALIDATION

There are notable deficiencies in the simulation of electricity theft patterns that authentically represent the complexities encountered in real-world scenarios.

c: CHALLENGES IN SIMULATING SOPHISTICATED THEFT SCENARIOS

The literature does not adequately address the challenge of creating advanced and realistic simulations that can accurately emulate the multifaceted nature of electricity theft, which hinders the development of robust detection systems.

d: OPPORTUNITIES FOR ENHANCED ANOMALY DETECTION

There is a significant opportunity to improve methods for anomaly detection, particularly by employing real-world attack instances in the validation of models, thus enhancing their predictive accuracy and reliability.

C. CONTRIBUTIONS OF THE PROPOSED MODELS

The development and implementation of our models for electricity theft detection in smart homes have led to several significant contributions to the field.

- **Innovative Hybrid Approach:** Our method merges domain expertise with machine learning to tackle labeled data scarcity in electricity theft detection. Utilizing a knowledge-based synthetic attack classifier, the SYNBDM, and LUM anomaly detectors, we notably boost the detection accuracy of anomalies in smart homes.
- **Advanced Data Analysis with UMAP:** Employing Uniform Manifold Approximation and Projection (UMAP) improves differentiation and adaptation to diverse household power usage patterns, highlighting our models' versatility in smart home environments and increasing accuracy in identifying theft-related anomalies.
- **Robust Anomaly Detection:** Our models, by aggregating power-base consumption patterns, enhance the ability to differentiate legitimate consumption from potential theft, improving theft detection reliability and reducing false positives from non-attack-related power usage changes.
- **Privacy-Sensitive Design:** Our framework's use of anonymized data emphasizes a dedication to user privacy in smart home systems, enabling efficient electricity theft detection while protecting consumer privacy.
- **Real-World Use and Testing:** Validating our models with extensive real-world appliance datasets demonstrates their applicability in detecting electricity theft across diverse real-life scenarios. Specifically, our knowledge-based synthetic attack classifier enhances robust identification, proving effective even when labeled data is scarce.

The remainder of this paper is organized as follows. Section II discusses related work. Section III describes the attack model before proceeding to Section IV to explain our proposal: electricity theft detection with synthetic attack data. Section V describes the dataset for electricity theft detection. Section VI presents the performance evaluation. Section VII provides a discussion and future work, and Section VIII concludes the paper.

II. RELATED WORKS

The authors [26] identified five distinct methods of electricity theft attacks from the perspective of utility companies: (1) meter bypass, (2) meter hacking, (3) direct hooking, (4) tapping, and (5) meter tampering. Additionally, the concept of a False Data Injection Attack (FDIA) for electricity theft [9], [10] has been discussed as a cyber-attack within the realm of cyber-physical systems. Jokar et al. [5] also included FDIA alongside physical attacks that involve bypassing or tampering with meters, all of which manipulate meter readings to reduce electricity consumption [27]. Consequently, the focus of attention has shifted towards electricity theft detection through the analysis of meter readings.

Various machine-learning techniques have been explored for classifying electricity consumption patterns and detecting electricity theft. These methods include Support Vector Machines (SVM) [5], [28], Decision Trees [29], and more recently, Gradient-Boosting [7], [30]. These approaches typically adopt supervised learning principles and, rely on labeled datasets containing both benign and attack data samples. However, obtaining ground truth labels for real-life scenarios is challenging.

An alternative approach to electricity theft detection, particularly when labeled data are unavailable, is anomaly detection. This method encompasses anomaly pattern detection based on hypothesis testing (APD-HT) [31], hierarchical self-organizing maps (SOM) [32], and stacked sparse denoising autoencoders [33]. Anomaly detection, while effective, has the drawback of flagging any abnormal occurrences, including instances such as an unusual appliance usage pattern during the nighttime [34].

Our work operates within the context of real-life scenarios where labeled data are scarce. However, we introduced a novel approach by incorporating knowledge of potential attack scenarios and synthetic attack data to train a supervised model using a non-labeled real-world dataset. Furthermore, our work capitalizes on fine-grained time-series data within a smart home environment, a resource that is currently unavailable in today's smart grid landscape.

In our previous work [8], we introduced nine algorithms for detecting five real-world simulated attack classes in smart homes based on appliance consumption patterns. The present work is an extension of [8]. This paper is focused on electricity theft detection in smart homes and improvements relative to [8] including making the algorithm robust against unclassified attacks, application of synthetic binary discriminator, and legacy unsupervised techniques to enhance classification accuracy, employment of real building appliance consumption dataset for performance evaluations and model comparison with other existing models.

III. ELECTRICITY THEFT ATTACK

The main objective of an electricity theft attack (ETA) is to pay less than the real value for the consumed energy.

In the context of detecting electricity theft and ensuring appliance usage authentication in smart homes, various attack scenarios can threaten the integrity, availability, and confidentiality of the system [35]. These threats are not limited to physical interactions with the distribution board, as in Figure 1A, but can also involve digital intrusions and manipulative tactics. Here are some potential attack scenarios:

A. PHYSICAL ATTACKS

1) **Meter Swapping:** Swapping meters with those from vacant or low-consumption premises.
2) **Power Diversion:** Rerouting the power supply within a community.
3) **Meter Tampering:** This encompasses removing or disconnecting meters, inverting meters, employing magnets to disrupt readings, and unauthorized Smart Meter (SM) access.

B. CYBER ATTACKS

1) **Credential Theft:** Gaining unauthorized meter access via stolen login details.
2) **Firmware Hacking:** Compromising Smart Meter firmware remotely.
3) **Data Tampering:** Modifying stored meter data, including total energy use, audit trails, and cryptographic keys.

C. DATA ATTACKS

1) **Zero/Negative Reporting:** Incorrectly reporting no or negative energy use.
2) **Consumption Report Alteration:** Halting or modifying energy consumption reports.
3) **Measurement Exclusion:** Excluding high-usage appliances from records.

A comprehensive description of the attacks within each category can be found in [36]. Our electricity theft detection (ETD) system is adept at identifying irregularities in appliance consumption patterns. Our synthetic binary discriminator model (SYNBDM) can find all the types of attacks we've talked about so far because the goal of all of them is to change the meter readings of different appliances.

Our previous study, [8] was primarily focused on the multiclass classification of attack scenarios using fine-grained time-series electricity consumption appliances with data points as shown in Table 2 from AMPds2 dataset [37]. The five attack types, as shown in Figure 1B and visualized in Figure 4, simulate various forms of electricity theft generated and their corresponding attack impacts as seen in Table 6 before deployment of data augmentation.

We extended the attack classes to unknown anomalies n and classification by the attack classifier as detailed in section IV.

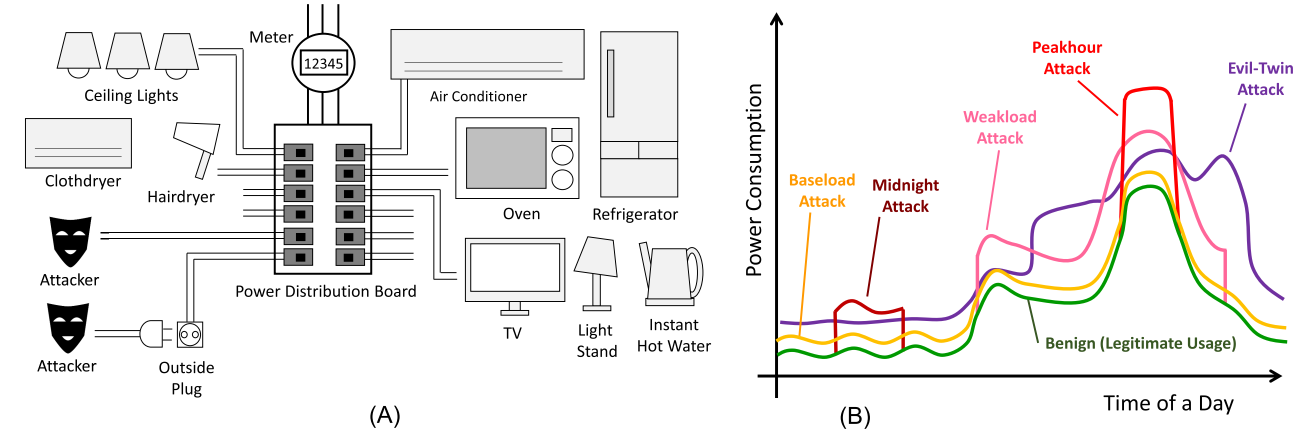


FIGURE 1. A: Power Distribution Board with connected sample appliances. B: Scenarios of electricity theft attacks. Power distribution boards and cables are deployed behind walls and are usually not visible. An attacker uses the pre-deployed cable in a complex building or the outside plug to steal electricity. Depending on the theft pattern, we identify five classes of attacks.

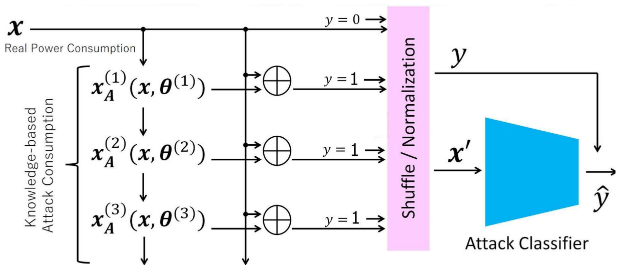

x
Real Power Consumption
$y=0$
Knowledge-based Attack Consumption
$x_A^{(1)}(x,\theta^{(1)})$
$y=1$
$x_A^{(2)}(x,\theta^{(2)})$
$y=1$
$x_A^{(3)}(x,\theta^{(3)})$
$y=1$
Shuffle / Normalization
y
x'
$\hat{y}$
Attack Classifier

FIGURE 2. ETD framework for processing real power consumption data to detect anomalies that indicate cyberattacks by transforming the data, labeling, preparing for analysis, and finally classifying it using a trained attack classifier.

TABLE 2. Appliances and units in AMPds.

Point ID	Description	Point ID	Description
WHE	Whole House Meter	FRE	HVAC/Furnace
B1E	North Bedroom	GRE	Garage
B2E	Master/South BR	HPE	Heat Pump
BME	Basement Plugs/Lights	HTE	Instant Hot Water Unit
CDE	Clothes Dryer	OFE	Home Office
CWE	Clothes Washer	OUE	Outside Plug
DNE	Dinning Room Plugs	TVE	Entertainment TV/PVR/AMP
DWE	Dishwasher	UTE	Utility Room Plug
EBE	Electronics Workbench	WOE	Wall Oven
EQE	Security/Network	UNE	Unmetered Loads
FGE	Kitchen Fridge		.

IV. ELECTRICITY THEFT DETECTION WITH SYNTHETIC ATTACK DATA

In our quest to effectively detect ETAs, we adopt a comprehensive approach that closely monitors power consumption patterns. We emphasize the importance of granularity in both the time and power domains, particularly at the power aggregation point as shown in Tables 4 and 3, and Figure 3.

TABLE 3. The configuration of synthetic attack data threshold for supervised learning algorithm.

Attack Class	Configuration
Baseload	Theft = 100W

Weakload	Threshold=500W
	Theft=500W

Peakhour	Threshold=1500W
	Theft=2000W

Midnight	Starting from 1 AM - 2 AM (at random)
	Duration = 120 minutes
	Theft=1000W

Evil-Twin	One Day was randomly chosen in the dataset.

In this study, we delve into the realm of synthetic and real attack data, leveraging their combined potential to validate machine learning applications. Our primary objective is to accurately identify genuine anomalies originating from legitimate electricity consumption patterns, as shown in the framework of synthetic attack learning for ETD in Figure 2.

Let us consider **x**, a vector of power consumption. This vector contains the power consumption of each timeslot in time order. For example, **x** may represent the power consumption of a certain day, and the i-th element x_i corresponds to the power usage at i-th minute from the beginning of the day. In this case, **x** had 1440 elements, that is, $60 \times 24 = 1440$.

In supervised learning, we assume that each **x** has a corresponding label y to train a classification model for the power consumption patterns. In our case, we consider unsupervised learning; therefore, we can assume that label $y = 0$ as a benign case, and all other attacks such as the Baseload attack, weakload attack, and, other attacks are all labeled $y = 1$ as an attack case. In real practical scenario, we will only obtain a collection of **x** from a house as a result of long-term monitoring, and we will not obtain real attack-enabled cases as anomalies. However, many power-stealing cases can be simulated by arithmetically adding stolen power as power consumption.

Let $\mathbf{x}_A$ be a vector of the stolen power of an attacker. As we assume that the attacker changes the stealing power based on the consumption of the house, $\mathbf{x}_A$ is a function of $\mathbf{x}$ and attacking parameters θ: e.g., $\mathbf{x}_A(\mathbf{x}, \theta)$.

In the case of an attack, that is, $y = 1$, we consider injecting an attack randomly from attack type z. Then, the stolen attack vector is: $\mathbf{x}_A^{(z)}(\mathbf{x}, \theta^{(z)})$.

Finally, we obtain the binary classification (benign and attack) dataset as follows:-

$$(\mathbf{x}', y) = \begin{cases} (\mathbf{x}, 0) & y = 0 \\ (\mathbf{x} + \mathbf{x}_A^{(z)}(\mathbf{x}, \theta^{(z)}), 1) & y = 1. \end{cases} \quad (1)$$

Machine learning models can be applied to the collection of $(\mathbf{x}', y)$ for binary classification and outlier detection.

A. KNOWLEDGE-BASED ATTACK SIMULATION FRAMEWORK

The framework proposed in Figure 2, employs a knowledge-based approach to generate synthetic attack scenarios on power consumption data, encapsulating domain expertise within its operational paradigm. The core of this methodology is the utilization of the actual power consumption profiles, denoted as x, as the foundational dataset from which attack patterns are derived. This framework benefits from recognizing both specific attacks and unclassified anomalies, integrating the strengths of both approaches for a robust security posture.

1) ATTACK DATA GENERATION

Distinct attack scenarios are simulated through a series of transformations applied to real consumption data, parameterized by θ. These transformations—$x_A(x, \theta^{(1)})$, $x_A(x, \theta^{(2)})$, and $x_A(x, \theta^{(3)})$—are crafted based on expert insights into the modus operandi of various attack vectors, with each θ iteration representing a unique attack typology.

2) DATA LABELING AND PREPROCESSING

Further cementing its knowledge-driven architecture, the framework classifies consumption data into normal ($y = 0$) and anomalous ($y = 1$) states, employing pre-established criteria that delineate normalcy from theft-related anomalies. Before classification, data undergo a shuffling and normalization process, for eliminating potential classifier bias attributable to sequential order or feature scale disparities.

3) ATTACK CLASSIFICATION

The culmination of the framework is the attack classifier, a predictive model trained on a rich historical corpus comprising known instances of consumption patterns, both benign and malignant. This classifier is not merely a data-driven algorithm but a knowledge-infused system tuned to recognize and react to the subtle intricacies of electricity theft within smart grid environments.

The framework's reliance on domain-specific knowledge for the generation and processing of data points designates it as knowledge-based. This is exemplified by the methodical application of expert understanding to the identification of theft signatures, which is paramount for effective discrimination between legitimate and fraudulent electricity usage patterns.

Our knowledge-based framework sets a new benchmark for electricity theft detection systems, marrying the depth of domain knowledge with the rigor of machine learning classification. This synergy promises a robust and discerning methodology, poised to advance the state-of-the-art in smart grid security.

Our research focuses on providing home operators with effective tools for cyber threat detection and categorization, offering multiple benefits:

- Our system's binary classification techniques enable immediate threat recognition, facilitating rapid response to cyber incidents.
- We deliver in-depth attack analysis through multiclass classification, crucial for prevention strategies and resource allocation.
- Customized incident response strategies, derived from our attack classification approach, enhance system resilience.
- Compliance with cybersecurity standards boosts customer confidence in our protective measures.
- Maintaining operational integrity is key, with our system's efficiency critical in minimizing the impact of cyber threats.

Figure 3 depicts the architecture of the proposed ETD mode to consolidate appliance consumption data from smart homes, ensuring privacy through anonymization and consistency through normalization. Simulated theft scenarios enhance the dataset, with each instance labeled as normal or fraudulent as described in Section V.

A training set derived from this data trains algorithms to detect consumption patterns, whereas a testing set comprising simulated and real data evaluates accuracy. The model employs both traditional machine learning and neural network classifiers, benchmarked against metrics such as accuracy, area under curve (AUC), and F1-score. In post-validation, the model was deployed, with ongoing retraining to refine its detection capabilities.

V. DATASET FOR ELECTRICITY THEFT DETECTION

A. DATA COLLECTION

For our study, we utilized the AMPds2 dataset (Almanac of Minutely Power Dataset version 2) [37] as a benchmark, representing two years' worth of home power consumption data. AMPds2 includes minute-by-minute power measurements recorded at the outputs of power distribution boards as shown in Figure 1A. The monitoring points within this dataset are detailed in Table 2. Given the varying configurations of homes, it is possible that certain appliances, such as Clothes Dryers, Wall Ovens, or Dishwashers, may not be present in some households. To address this variability, we simulated three distinct home types by excluding the

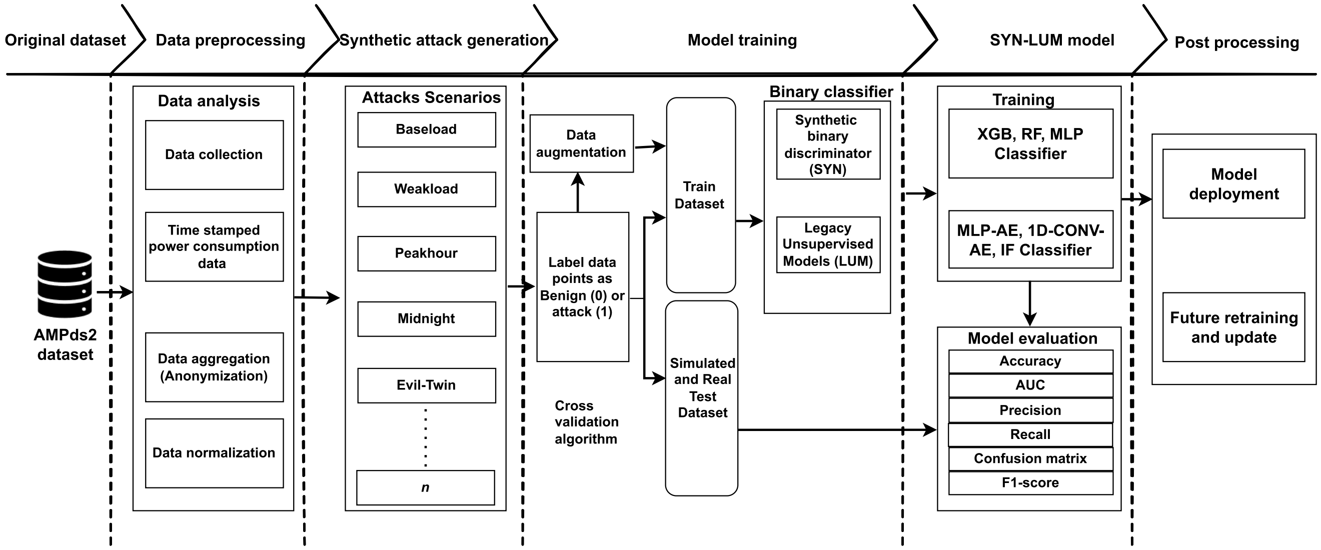


FIGURE 3. Flow of ETD model.

TABLE 4. Home configurations based on appliance data-points.

Home	Aggregated Appliances	Excluded Appliances
A	B1E, B2E, BME, CWE, DNE, HTE, EBE, EQE, FRE, OFE, OUE, TVE, UTE, CDE, HPE, DWE, FGE, WOE	None (full set)
B	B1E, B2E, BME, CWE, DNE, HTE, EBE, EQE, FRE, OFE, OUE, TVE, UTE, HPE, DWE, FGE	CDE, WOE
C	B1E, B2E, BME, CWE, DNE, HTE, EBE, EQE, FRE, OFE, OUE, UTE, CDE, FGE, WOE	DWE, HPE, TVE

power consumption of these optional appliances as listed in Table 4.

In configuring each home, we assumed the following:

- **Home A**, some appliances associated with HPE and WOE have peak power consumption, (Figure 4), which may allow the peak-hour attacker to steal power more efficiently.
- **Home B** that the existence of a cloth dryer and wall oven may influence the accuracy of the attacker detection.
- **Home C** that the existence of a dishwasher, heat pump, and small appliances may influence the accuracy of attacker detection.

Recognizing that electric power consumption data are inherently time-dependent, we adopted a data segmentation strategy. Specifically, we selected the initial 80% of the data, equivalent to 584 days, for our training dataset. During this phase, we applied our synthetic attack data methodology to enrich the dataset. The remaining 20% of the data, spanning 146 days, was reserved for our test dataset in our benchmark experiment as listed in Table 6. Importantly, the test dataset included simulated attacks generated as part of this study.

It is noteworthy that the test dataset did not cover the entire year. This deliberate choice was made to allocate a larger volume of data to the training phase, thereby enhancing the robustness of our models.

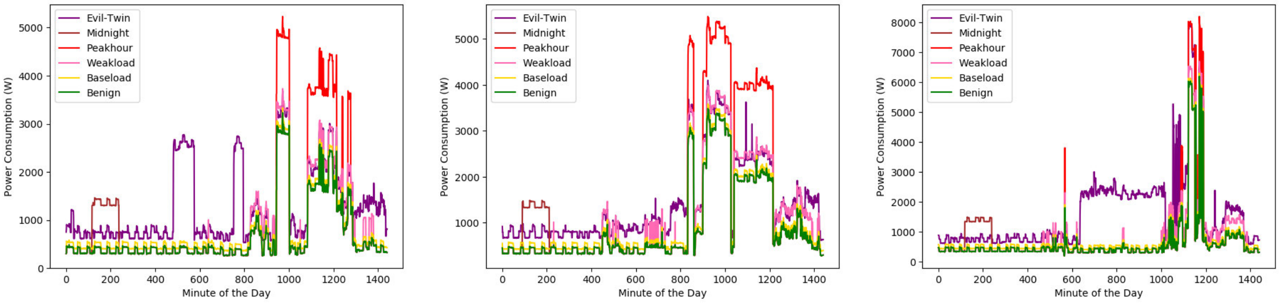
1) DATA PROFILES

Table 6 shows profiles of the data for our electricity theft attack (ETA) detection study. We have 584 benign records for training each home from the original monitoring data. We simulated and added Baseload, Weakload, Peakhour, Midnight, and Evil-Twin attacks as described in our previous work [8] and parameters threshold for each attack as described in Table 3. In some conditions, if the aggregated base power consumption of the home does not reach the threshold, Weakload and Peakhour attacks (smart attacks) are not triggered.

Figure 4 shows examples of Benign and Attack data samples of different three days of Home A. From these figures, we can observe that the electric consumption pattern has a huge change by the day, and we cannot easily recognize the attack class only from a single power consumption data. Figure 5 illustrates the data samples using the Uniform Manifold Approximation and Projection (UMAP) technique for homes A, B, and C [38]. Each plot in this figure corresponds to the data from a single day. Notably, the different homes exhibited distinct data characteristics. For example, in Home A, benign data are dispersed and overlap with various attack samples. Conversely, Home C displayed more discernible clusters which may facilitate classification.

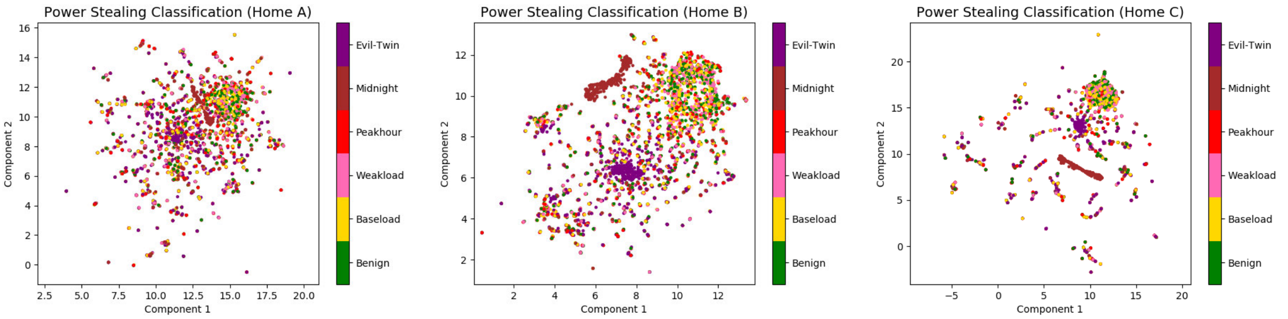
B. OVERVIEW OF THE FEATURE SELECTION

1) **Aggregated Power Consumption Patterns (APCP)** The feature for each minute m on day d is represented as the power consumption at that minute, denoted by $P_{d,m}$, where $d = 1, 2, \ldots, 730$ (for 730 days) and $m = 1, 2, \ldots, 1440$ (for 1440 minutes in a day). The



(a) Day 1 (b) Day 2 (c) Day 3

FIGURE 4. **Electricity usage of Home A with synthetic attack data on different days. The benign consumption patterns drastically change day by day. Especially, The Weakload and Peakhour attacks are sophisticated – having similar consumption patterns with the legitimate usage although the values are different.**



(a) Home A (b) Home B (c) Home C

FIGURE 5. **2D projections of attack contained electricity usage with the benign case by uniform manifold approximation and projection (UMAP) [38]. Each plot corresponds to the consumption pattern of a day on different attack scenarios. Depending on the available appliances of the homes, the shapes of the attack classes are very different. There are more overlaps among classes in Home A, whereas we observe fewer overlaps in Home C.**

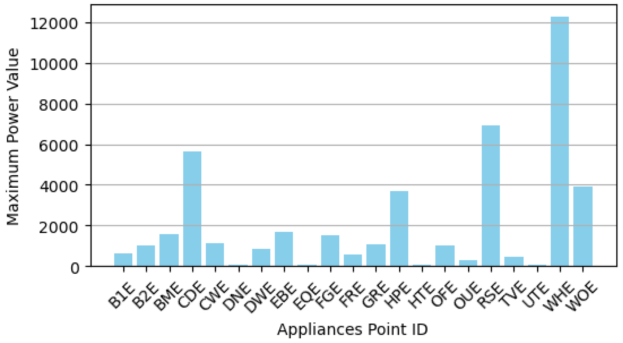


FIGURE 6. **Maximum power values for each appliance.**

aggregated power consumption for a day d is given by:

$$\mathrm{APCP}_d = \sum_{m=1}^{1440} P_{d,m} \tag{2}$$

2) **Deviation from Typical Consumption (DTC)**
This feature measures the deviation of actual consumption from the expected (baseline) consumption. Let $B_{d,m}$ represent the baseline power consumption for minute m on day d. The deviation for that minute is:

$$\mathrm{DTC}_{d,m} = P_{d,m} - B_{d,m} \tag{3}$$

The total deviation for a day d can be aggregated as:

$$\mathrm{DTC}_d = \sum_{m=1}^{1440} |\mathrm{DTC}_{d,m}| \tag{4}$$

3) **Temporal Features (TF)**
Temporal features could include binary indicators for peak and off-peak hours. We define PeakHour(m) as a function that returns 1 if minute m is within peak hours, and 0 otherwise. The temporal feature for a day d is:

$$\mathrm{TF}_d = \sum_{m=1}^{1440} \mathrm{PeakHour}(m) \times P_{d,m} \tag{5}$$

4) **Combined Feature Vector**
For a machine learning model, these features collectively form the input vector for each day d, represented as:

$$\mathrm{FeatureVector}_d = [\mathrm{APCP}_d, \mathrm{DTC}_d, \mathrm{TF}_d] \tag{6}$$

To determine the best feature for measuring appliance consumption patterns, we consider how each feature relates to energy consumption and how well it might differentiate between different consumption patterns [39]. Let us now review each feature:-

1) ***V* (voltage)**: While voltage levels can affect power consumption, in most residential and commercial settings, the voltage is relatively constant. It is not a direct measure of consumption but can be relevant in some analyses.
2) ***I* (Current)**: The current is directly related to the power consumption ($P = VI$, where V is the voltage and I is the current). Fluctuations in current can indicate changes in appliance consumption patterns.
3) ***f* (Frequency)**: The frequency is stable in most power systems. Variations are usually an indication of grid instability rather than appliance consumption patterns.
4) ***DPF* (Displacement Power Factor)**: This measures the efficiency of power usage but does not directly indicate consumption levels. This is more about the quality of consumption than the quantity.
5) ***APF* (Apparent Power Factor)**: Similar to *DPF*, it indicates the efficiency of power usage and is more about power quality.
6) ***P* (Power)**: Power is a direct measure of energy consumption at any given moment. This is one of the most direct measures of appliance consumption.
7) ***Pt* (Total Power)**: If this is cumulative power over time, it is an excellent measure of total consumption but less useful for instantaneous consumption patterns.
8) ***Q* (Reactive Power)**: This is related to the energy stored in the load and returned to the source and is more about the type of load than the quantity of consumption.
9) ***Qt* (Total Reactive Power)**: Similar to Q, but cumulative. It is more relevant to assessing load type over time than consumption patterns.
10) ***S* (Apparent Power)**: This is a combination of reactive power and real power and provides a total power figure but doesn't directly measure consumption efficiency.
11) ***St* (Total Apparent Power)**: Cumulative apparent power over time. Like S, it encompasses active and reactive power but does not directly indicate consumption patterns.

Based on this analysis, when we apply the correlation coefficient for feature selection, for most relevant features, P and Q are relative to S. Figure 7 shows that feature P has a higher correlation to feature S (the orange bar) and a slightly lower, yet still high correlation with feature 2 Q (the blue bar). Also, Figure 8 shows that **(*P*) (Power)** is the best feature for measuring appliance consumption patterns using a mutual information algorithm [40] because it directly reflects the amount of electric power being used at any given moment. In the case of power consumption patterns, the machine automatically learns the key features from the raw sequence of power consumption data without providing any statistically processed data as explicit features.

The power_base dataset has the following features: • Time Components: • Day: The day number relative to the baseline (day_offset). • Minute: The minute of the day. • Aggregated Power Readings: • We extracted the active power P for its processing; the aggregated data in power_base consists of the sum of active power readings from selected meters (base_names) for each minute of each day. Therefore, each entry in power_base represents the total active power consumption (in watts) from a subset of meters, Figure 6, for each minute of each day over 730 days. This aggregated dataset forms a baseline for normal power consumption patterns against which deviations (such as potential electricity theft) can be compared.

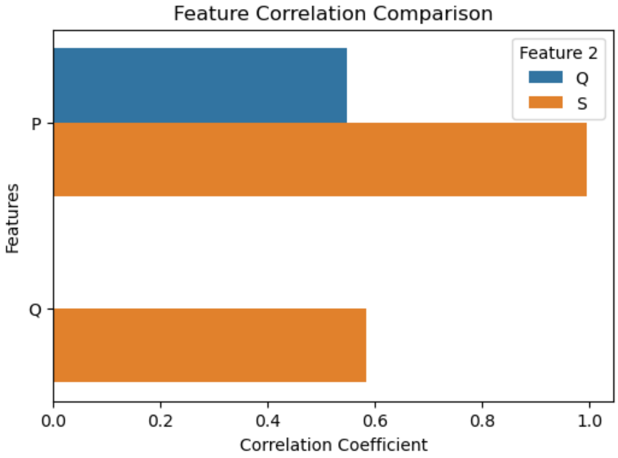


FIGURE 7. Features selection by correlation comparison.

This dataset is pivotal for our analysis and predictive modeling. The structure of `power_base` is organized as a two-dimensional array, where each entry in `power_base[day][min]` denotes the aggregated power consumption for a specific category, as listed in Table 4 for a given day and minute. The data spanned a temporal resolution of one minute, totaling to 1440 min (24 h) per day. This granularity allows for a detailed analysis and forecasting of power consumption patterns.

C. DATASET PREPROCESSING FOR BINARY CLASSIFICATION

The electricity theft attack detection dataset (ETA-DD) consists of:

1) two training sub-datasets, and
2) two testing sub-datasets for homes A, B, and C.

This ETA-DD is assumed for binary classification problems (*Benign* or *Attack*). This is because it is intended for applying unsupervised learning for attack detection and evaluating detection accuracies with real building data. The real building power consumption cannot be easily labeled for the predefined attack cases (as mentioned in subsection A above). These are the reasons why this study focuses on attack detection, rather than classification. Even though it is not intended for attack classification, the evaluation scope has drastically widened.

This paper expands the limitations of the model to detect attacks that are not classified hence we preprocessed our

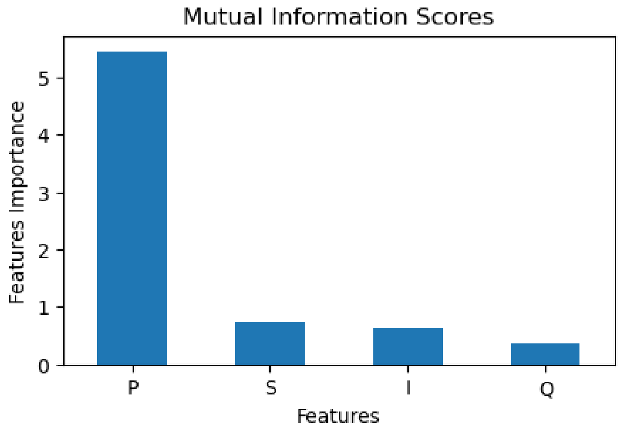


FIGURE 8. Features selection using mutual information scores.

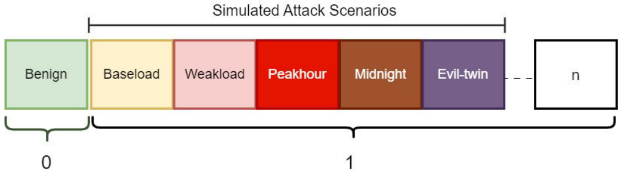


FIGURE 9. Multiclass attack scenarios with unspecified attacks *n*, preprocessed to binary attack class for anomaly detection.

dataset to accommodate the 5 simulated attack scenarios to include unknown or new attacks, n, as *1* and the benign or non-attack data as *0* as shown in Figure 9.

If we denote the original class labels by C where $C \in$ {Benign, Baseload, Weakload, Peakhour, Midnight, Evil-twin, $\ldots, n$}, with their attack parameters as shown in Table 3, the binary classification function $f(C)$ can be defined as

$$f(C) = \begin{cases} 0 & \text{if } C = \text{Benign} \\ 1 & \text{otherwise} \end{cases} \tag{7}$$

Here, label *0* corresponds to the normal (benign) class, and label *1* corresponds to any kind of attack scenario. This binary labeling strategy is a common approach in anomaly or intrusion detection systems [41] where the focus is on differentiating between normal and abnormal behaviors, regardless of the specific type of abnormal activity.

Our real test attack data which have the same features (aggregated power_base consumption patterns), further solidify the evaluation of our model's performance.

We extended our research by transitioning from multiclass [8] to binary classification, incorporating data from three distinct homes (Home A, Home B, and Home C as depicted in Table 4.

D. DATA PREPARATION

1) DATA ANONYMIZATION

Data from various appliances are combined to create a comprehensive view of a home's power usage, with measures to protect user privacy.

The main feature of our dataset is the aggregation of the power consumption of each appliance in each home. The values are only numerical readings without any direct personal identifiers.

The aggregated baseline power consumption for each home (Table 4), $P_{\text{total base}}$ for n appliances is given by the sum of individual baseline power consumptions $P_{\text{base},i}$:

$$P_{\text{total base}} = \sum_{i=1}^{n} P_{\text{base},i} \tag{8}$$

where $P_{\text{base},i}$ is the baseline power consumption of the i-th appliance.

The total energy consumption E_{total}, considering the duration of usage D_i for each appliance, is calculated as:

$$E_{\text{total}} = \sum_{i=1}^{n} (P_{\text{base},i} \times D_i) \tag{9}$$

where D_i is the duration of usage for the i-th appliance in hours, and $P_{\text{base},i}$ is as defined earlier.

2) NORMALIZATION

In our ETD for smart homes, we deployed both StandardScaler and MinMax to normalize the feature vectors before synthetic binary discriminators (SYNBDM) and legacy unsupervised models (LUM) experiments respectively.

- For SYN-supervised models, we used **StandardScaler** to normalize the feature vectors by removing the mean and scaling to unit variance which helped improve the performance and stability of our models.
  The scaler is first fitted on the training data:

$$\mu_j = \frac{1}{n} \sum_{i=1}^{n} X_{train_{ij}} \tag{10}$$

$$\sigma_j = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_{train_{ij}} - \mu_j)^2} \tag{11}$$

  where μ_j and σ_j are the mean and standard deviation for each feature j, and n is the number of training samples. The training data are then transformed using these parameters:

$$X_{train\_scaled_{ij}} = \frac{(X_{train_{ij}} - \mu_j)}{\sigma_j} \tag{12}$$

  The same transformation is applied to the test data:

$$X_{test\_scaled_{ij}} = \frac{(X_{test_{ij}} - \mu_j)}{\sigma_j} \tag{13}$$

  This ensures that both training and test data are on the same scale.
- **MinMaxScaler** was deployed to ensure that the input features contribute equally to the model training, enhancing the learning process of anomaly detection.

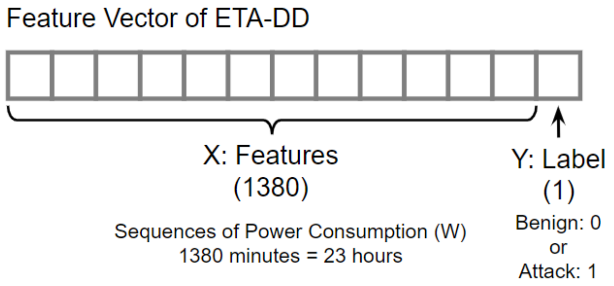


FIGURE 10. Features vector and labels.

The MinMax Scaler linearly transforms each feature to a common scale, typically between 0 and 1. The transformation is defined as:

$$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{14}$$

where $X_{\min}$ and $X_{\max}$ are the minimum and maximum values of the feature in the training dataset, respectively, and X represents the original feature value.

3) FEATURE VECTOR (X) AND LABEL (Y)

Each file starts from a header row, followed by consumption records. Each consumption record is organized as shown in Figure 10.

X: features consisting of 1380 elements, each corresponding to the power consumption of the minute of the day. However, the 1380 min was 23 h. This is because we applied Data Augmentation, as described in the Data Augmentation subsection below.

Y: label that identifies the meaning of the X. *Benign* is defined as 0. *Attack* is defined as 1.

4) DATA AUGMENTATION

Let $\mathbf{X}$ be the original feature, Table 6, vector of length $n = 1380$, representing the power consumption readings every minute for 23 h as depicted in Figure 10:-

$$\mathbf{X} = [x_1, x_2, \ldots, x_n] \tag{15}$$

Let k be the time offset for circular shifting, where $k \in \{0, 1, \ldots, 59\}$.

The augmented feature vector $\mathbf{X}'$ after applying a circular shift of k positions is defined as:

$$\mathbf{X}' = [x_{(i-k) \mod n}, x_{(i-k+1) \mod n}, \ldots, x_{(i-1) \mod n}, x_i, \ldots, x_{(i-k-1) \mod n}] \tag{16}$$

where $i = 1, 2, \ldots, n$ and x_i is the power consumption at the i-th minute. The modulo operation mod ensures that the index wraps around the vector when the shift exceeds the start of the vector.

Label Y remained unchanged for all augmented records. If the original record is labeled as benign ($Y = 0$) or attacked ($Y = 1$), then all augmented versions of the record retain this label:

$$Y' = Y \tag{17}$$

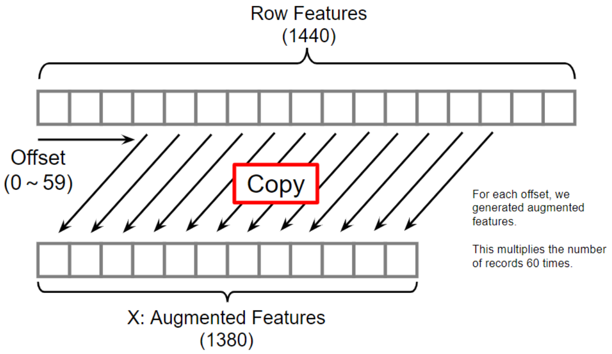

Row Features
(1440)
Offset
(0 ~ 59)
Copy
For each offset, we generated augmented features.
This multiplies the number of records 60 times.
X: Augmented Features
(1380)

FIGURE 11. Framework of data augmentation.

This augmentation process was repeated 60 times for each record in the dataset, as shown in Figure 11, corresponding to each possible offset k, thereby effectively increasing the number of benign records.

The length of the augmented feature vector $\mathbf{X}'$ is 1380 instead of the full 1440 which represents the total number of minutes in 24 hours. The reason for this are twofold:-

1) A circular shift [42] by an offset k ranging from 0 to 59 min implies that we need to have a buffer at the end of the vector to accommodate the maximum possible shift without wrapping the data of the next day.
2) By limiting the feature vector to 1380 min, we ensured that for the largest shift of 59 min, the augmented data still represents a continuous 23-h window from the same day. This maintains the integrity of the daily patterns in power consumption without mixing data from two different days.

The label Y, which represents whether the original 1440-minute vector corresponds to a benign or attack pattern, remains associated with the corresponding 1380-minute augmented vector $\mathbf{X}'$. This ensures that the model learns to detect anomalies based on the most representative and complete daily consumption patterns possible within the constraints of the data augmentation process as listed in Table 5.

To illustrate the effectiveness of this augmentation technique, Figure 12 shows augmented power consumption patterns for a single Home A while Figure 13 displayed multiple homes (Homes A, B, and C), respectively.

In our daily lives, the operation of home appliances may be shifted by approximately one hour. Based on this idea, we have shifted the original data over the time axis up to 60 min and extracted them as benign records.

Table 5 introduces the additional data details and parameters for ETA detection. We expanded the number of benign records to 35,040 benign training records, each home, sourced from the monitoring data. The table reveals that there

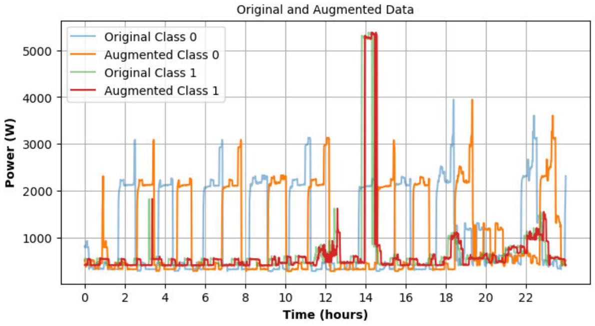


FIGURE 12. Electricity power consumption pattern of Home A for original class and augmented data class for a day.

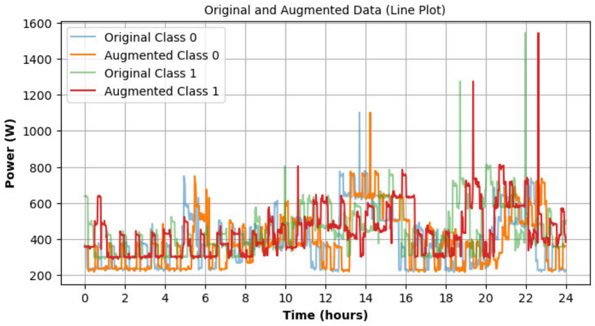


FIGURE 13. Electricity power consumption pattern of all homes for original class and augmented data class for a day.

were "0" instances of unsupervised training data across all homes. The LUM anomaly detectors were trained solely on benign samples and then subjected to testing on datasets containing both benign and simulated attack samples while synthetic supervised detectors were trained with both benign and attack samples. Their performance was further evaluated using real attack samples.

In general, if the AI of energy theft is large, it can be easily detected with higher accuracy. If the AI of the energy theft is small, it can be negligible, that is it does not need to be detected. Therefore, we selected well-balanced power consumption for evaluation from real building facilities.

5) UTOKYO DATA - REAL ATTACK DATA

To test the performance of ETD in both synthetic and unsupervised approaches, we consider including the consumption pattern of real rooms of the University of Tokyo as attack data for the model tests. We measured the power consumption at the power distribution boards of the I-REF building (6th-floor building) from 2012 with a sampling frequency of 1 min. We selected the daily consumption patterns in which the attack impact (AI) corresponded to the attack on the simulation test dataset.

6) ATTACK IMPACT

We define attack impact (AI) as a score to measure electricity theft. This attack affects the amount of energy stolen in the bill.

TABLE 5. Distribution of simulated and real binary dataset.

Dataset	Home	Benign samples	Attack samples
SYN Train	A	35040	33818
	B	35040	33242
	C	35040	31533
UN Train	A	35040	0
	B	35040	0
	C	35040	0
Sim Test	A	146	144
	B	146	142
	C	146	133
Real Test	A	8760	8760
	B	8760	8760
	C	8760	8760

Note: SYN = Synthetic, UN = Unsupervised

Let w_i be the weighted price of the power of x_i on the day at each index i. Let us denote the vector of w_i by $\mathbf{w}$.

The attack impact of electricity theft of the day is:

$$I = \mathbf{w} \cdot \mathbf{x_A} = \sum_i w_i x_{Ai} \tag{18}$$

Here, x_{Ai} represents the power stolen at index i of $\mathbf{x_A}$.

To calculate AI on our dataset, we assumed 0.20 USD/kWh constantly for the unit price. The average daily bill mounted by the attacker was approximately 1 USD on average (see Table 6). It is approximately 30 USD monthly and 360 USD per year. If the base power consumption of the home is larger, such as in the case of an office, shop, restaurant, or factory, the attacker will be able to steal more power.

VI. EVALUATION

For our investigation, we deployed Python Jupyter Notebook 3.9.16 and Keras [43] with TensorFlow as the backend. We conducted our experiment using an Intel Core i9 CPU 2.50GHz with 16GB RAM. Windows 11 (64-bit), and NVIDIA GeForce RTX 3070 Ti Laptop GPU. We present a comprehensive performance evaluation of various machine learning models for ETD, utilizing both real-world and synthetic attack datasets. We primarily focus on the AUC metric from Table 11 as the primary evaluation criterion and complement it with additional metrics from Table 8, including the F1-score, accuracy, precision, and recall.

A. SYNTHETIC BINARY DISCRIMINATOR MODEL (SYNBDM)

Figure 14 shows the evaluation flow of our SYNBDM. We examine the performance of supervised XGB, RF, and MLP classifiers in different homes. Suppose X represents the input features, from the synthetic train dataset, and Y is the output prediction for binary classification,

$$f(X) = Y \tag{19}$$

where f represents the learning function of the Binary Discriminator.

During the testing phase, the trained model, f is evaluated using two different datasets:

TABLE 6. The Profile of the dataset generated with synthetic attack data, and corresponding attack impacts (AI).

Home	Category	Benign	Baseload	Weakload	Peakhour	Midnight	Evil-Twin	Averaged AI	AI Ratio
A	Train	584	584	580	490	584	584	1.15 USD	1.77
	Test	146	146	146	140	146	146	1.17 USD	1.72
B	Train	584	584	580	437	584	584	1.08 USD	1.80
	Test	146	146	146	130	146	146	1.10 USD	1.77
C	Train	584	584	576	306	584	584	0.78 USD	1.66
	Test	146	146	146	84	146	146	0.76 USD	1.68

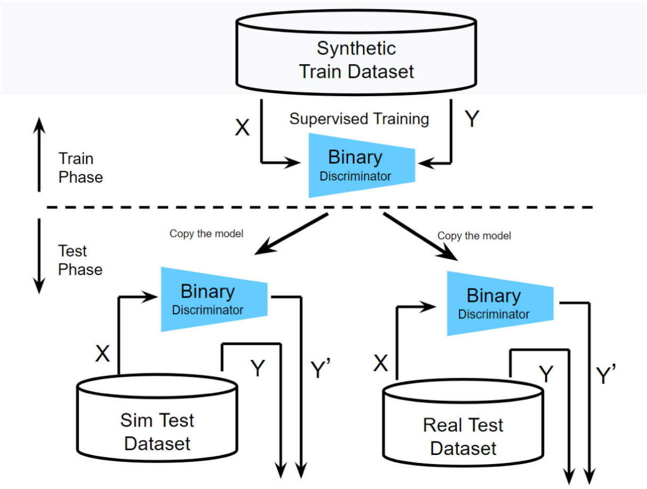


FIGURE 14. Flow of evaluation with Supervised Binary Descriminator.

For the Simulated Test Dataset:

$$f(X_{\text{sim}}) = Y'_{\text{sim}} \tag{20}$$

where X_{sim} are the input features and Y'_{sim} is the output predicted by the model.

For the Real Test Dataset:

$$f(X_{\text{real}}) = Y'_{\text{real}} \tag{21}$$

where X_{real} is the input feature and Y'_{real} is the output predicted by the model.

The performance of the model was assessed based on the accuracy of the predictions Y'_{sim} and Y'_{real} in comparison to the true labels.

1) PROPOSED MODELS CHARACTERISTICS OVERVIEW

XGBoost is a prominent ensemble learning method that, primarily utilizes decision tree structures. It employs gradient boosting, a technique that iteratively refines models by integrating multiple weak learners to formulate a robust predictive framework.

Regularization: A distinctive feature of XGBoost is the incorporation of a regularization term into its objective function [44]. This term is instrumental in mitigating the risk of overfitting, thereby enhancing the model generalization.

$$\text{Objective}_{XGB}(\theta) = \sum_{i=1}^{n} \text{loss}(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{22}$$

where θ denotes the model parameters, n is the number of observations, $\text{loss}(y_i, \hat{y}_i)$ is the loss function, and $\sum_{k=1}^{K} \Omega(f_k)$ is the regularization component.

Random Forest is an ensemble learning technique based on decision tree algorithms. It constructs a multitude of decision trees during training, and their collective output obtained through averaging or majority voting, constitutes the final model prediction.

Overfitting Reduction: The algorithm introduces randomness in tree generation, effectively reducing overfitting compared to individual decision trees [45].

$$\text{Objective}_{RF}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \text{loss}(y_i, \hat{y}_i) \tag{23}$$

where θ represents the model parameters, n is the number of data points, and $\text{loss}(y_i, \hat{y}_i)$ are the loss functions.

Multilayer Perceptron (MLP) is a class of feedforward artificial neural networks, characterized by multiple layers of nodes. Each layer is interconnected through weights and biases, enabling MLPs to capture complex, non-linear relationships in the data [46].

Backpropagation: MLPs rely on backpropagation for training, which is an algorithm that iteratively adjusts weights and biases to minimize the error between the actual and predicted outcomes.

$$\text{Objective}_{MLP}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \text{loss}(y_i, \hat{y}_i) + \alpha \sum_{i=1}^{L-1} ||W_i||^2 \quad (24)$$

where θ denotes the model parameters, n is the number of observations, loss$(y_i, \hat{y}_i)$ the loss function, α is the regularization parameter, and L the number of network layers.

For each model, parameter estimation can often be described using Maximum Likelihood Estimation (MLE), which for classification problems, involves maximizing the log-likelihood function:

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} \log P(y_i|X_i; \theta) \quad (25)$$

where θ represents the parameters, $P(y_i|X_i; \theta)$ is the probability of the target y_i given the input X_i, and $\hat{\theta}$ is the set of parameters that maximizes the likelihood, that is, involves finding the values of hyperparameters, for example, learning rate (eta), max-depth or the number of trees (n_estimators) that maximize the likelihood of observing the actual data.

We performed a grid search with cross-validation techniques, where the objective function, which is a combination of the loss function and regularization was minimized during training, and selection was made based on the hyperparameter tuning of each model.

We normalized our dataset with a StandardScaler to improve performance and, trained our SYNBDM classifier with benign and synthetic attack samples, as shown in Table 5 for Home A. We evaluated the degree to which the classifier is capable of detecting attack instances with real test data. We repeated the same experiment with Homes B and C, on both the simulated and real attack datasets. In the experiment, we used Jupyter Notebook, an open-source web application, written in Python; hence, it was easy to use with TensorFlow. Table 7 lists the parameter values used in the binary classification experiment. We selected the best hyperparameter values by experimenting with grid search. We performed validation through the *fit()* function using validation data, that is, simulated and real data. After training and testing each home, we calculated the accuracy and loss based on the number of correctly classified instances.

Our framework integrates model selection and feature selection to optimize the machine learning pipeline for ETD. It uses synthetic training data for the initial model training and hyperparameter tuning. The model is then validated on synthetic and real test datasets to ensure that it generalizes well to unseen, real-world data. This is achieved by performing a test each time the new appliances are connected. Each new appliance was preprocessed and converted into a proper format consistent with the training set. The proposed XGB is applied to a new sample format to determine whether it belongs to the benign or attack class. The framework incorporates MLE to optimize the objective function, ensuring that the models are well-calibrated, and providing probabilistic outputs that can be interpreted as risk scores for ETD. Table 8 presents the results of our experiments for all the homes.

TABLE 7. Parameters for the binary supervised discriminator.

Model	Parameters	Values
XGB	learning rate	0.1
	n_estimators	100
	reg_alpha	0.01
	reg_lambda	1.0
RF	n_estimators	50
	max_depth	10
	min_samples_leaf	1
	min_samples_split	2
MLP	hidden_layer_sizes	50, 100
	activation	tanh
	batch_size	50
	learning_rate_init	0.001

TABLE 8. Flow of the evaluation with synthetic binary discriminator.

Home	DataSet	Model	Acc	Recall	F1-score	Prec	AUC
A	Sim	XGB	0.9521	0.9169	0.9492	0.9839	0.9876
		RF	0.9197	0.8452	0.9113	0.9886	0.9647
		MLP	0.9733	0.9633	0.9723	0.9816	0.9656
	Real	XGB	0.9555	0.9209	0.9530	0.9875	0.9891
		RF	0.9223	0.8505	0.9146	0.9894	0.9702
		MLP	0.9243	0.9048	0.9212	0.9383	0.9668
B	Sim	XGB	0.9615	0.9281	0.9590	0.9921	0.9878
		RF	0.9359	0.8718	0.9296	0.9957	0.9756
		MLP	0.9514	0.9237	0.9486	0.8718	0.9724
	Real	XGB	0.9624	0.9293	0.9599	0.9927	0.9878
		RF	0.9327	0.8662	0.9259	0.9942	0.9764
		MLP	0.9441	0.9250	0.9413	0.9583	0.9795
C	Sim	XGB	0.9526	0.9116	0.9480	0.9837	0.9853
		RF	0.9269	0.8539	0.9168	0.9906	0.9625
		MLP	0.9168	0.8791	0.9168	0.9406	0.9537
	Real	XGB	0.9570	0.9209	0.9540	0.9894	0.9853
		RF	0.9291	0.8612	0.9216	0.9911	0.9654
		MLP	0.9293	0.9026	0.9251	0.9487	0.9649

B. LEGACY UNSUPERVISED MODEL (LUM)

The process depicted in Figure 15 involves an unsupervised learning model, specifically an autoencoder, which is trained to detect anomalies based on the reconstruction error.

During the training phase, the autoencoder learns parameter θ by minimizing the reconstruction error of the input historical electricity consumption data (benign only) X as follows:

$$\min_{\theta} \|X - \hat{X}(\theta)\|^2 \quad (26)$$

where $\hat{X}(\theta)$ denotes the reconstructed output of the autoencoder, and θ denotes its parameters.

A threshold τ was established based on the reconstruction error distribution during the training phase as shown in Table 10. These thresholds were used to classify the data points as either normal or anomalous.

In the testing phase, the autoencoder reconstructs new data from both Simulated and Real Test Datasets:-

$$\hat{X}' = \text{Autoencoder}(X') \quad (27)$$

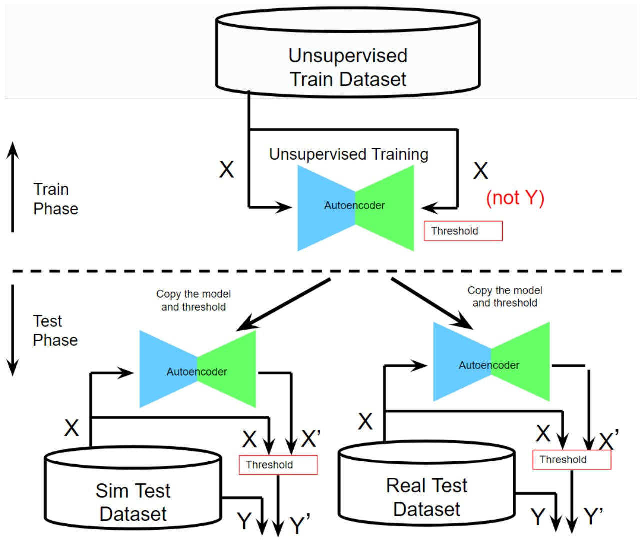


FIGURE 15. Flow of evaluation with Unsupervised Autoencoder.

Subsequently, the reconstruction error E for each data point is computed:

$$E = \|X' - \hat{X}'\| \tag{28}$$

An anomaly is flagged if the reconstruction error E exceeds the predetermined threshold τ:

$$Y' = \begin{cases} \text{Anomaly} & \text{if } E > \tau \\ \text{Normal} & \text{if } E \leq \tau \end{cases} \tag{29}$$

The performance of the autoencoder in anomaly detection is contingent on the accuracy of the threshold τ and its capability to accurately learn the representation of normal data during training.

The binary classification output Y' indicates whether a data point is normal or anomalous, based on the reconstruction error relative to the threshold.

Our experimental parameter settings in Table 9 reference attack detection based on unsupervised binary classification models [47]: Multi-Layer Perceptron Autoencoder (MLP_AE) employed a fixed learning rate of 0.001. The autoencoder has a single hidden layer consisting of 32 neurons. A batch size of 32 was used for the training. The model was trained for 100 epochs, and the Adam optimizer was utilized. A validation split of 10% was employed during the training.

1D Convolutional Autoencoder (1D-CONV_AE): A fixed learning rate of 0.001 was used. The latent space dimension was set to 64. A batch size of 32 was used during the training. The model was trained for 100 iterations. The Adam optimizer was utilized and a validation split of 10% was employed during training. We used 100 trees in the Isolation Forest (IF) algorithm. The random state was set to 42 to ensure reproducibility. The contamination parameter was set to 0.05, which represented the assumed proportion of outliers in the dataset.

TABLE 9. Parameters for legacy unsupervised models.

Model	Parameters	Values
MLP_AE	learning rate	0.001
	hidden_layer_sizes	32
	batch_size	32
	epoch	100
	optimizer	adam
	Validation_split	0.1
1D-CONV_AE	learning rate	0.001
	latent_dim	64
	batch_size	32
	epoch	100
	optimizer	adam
	Validation_split	0.1
IF	n_estimators	100
	random state	42
	contamination	0.05

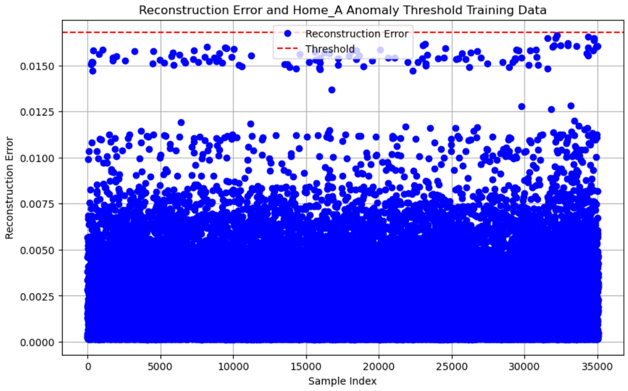


FIGURE 16. Threshold determination for Home A training dataset.

Equations (30), (31), (32), and (33) indicate that TP, TN, FP, and FN are true positive, true negative, false positive, and false negatives respectively. A TP refers to a sample that is malicious and is detected as malicious. TN indicates a benign sample that was detected as benign. FP indicates that the sample is benign but is detected as malicious. An FN represents a malicious sample detected as benign [48].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{30}$$

$$Precision = \frac{TP}{TP + FP} \tag{31}$$

$$Recall = \frac{TP}{TP + FN} \tag{32}$$

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{33}$$

For our electricity consumption data from various homes and datasets, normalization facilitates the scaling of input features. This scaling ensures that the features, such as aggregate power consumption, have uniform scales across different homes and datasets; therefore, the original feature values X are transformed into scaled values X_{scaled} within the *[0, 1]* range. This standardized scaling process is essential for the autoencoder to accurately learn and detect anomalies in electricity consumption patterns.

TABLE 10. Legacy unsupervised AE threshold values for anomaly detection.

Home	Dataset	Model	Threshold
A	Sim & Real	UN-MLP-AE	0.017565
		UN-1D-CONV-AE	0.000020
B		UN-MLP-AE	0.018085
		UN-1D-CONV-AE	0.000055
C		UN-MLP-AE	0.019676
		UN-1D-CONV-AE	0.000055

During training, the autoencoder leveraged these scaled features to reconstruct benign data, and a threshold was determined using the statistical method Median Absolute Deviation (MAD), (Equations 34 and 35), and reconstruction errors to identify anomalies. For example, Figure 16 shows the threshold determination from the training set only for home A while Table 10 depicts the threshold values used in training MLP-AE and 1D-CONV-AE for all homes.

The robust Z-score method uses the Median Absolute Deviation (MAD) [49] instead of the standard deviation, and is not significantly affected by outliers.

The mathematical representation of the Modified Z-score is:

$$\text{MAD} = \text{median}\left(|x_i - \tilde{x}|\right) \tag{34}$$

$$\text{Modified Z-score} = 0.6745 \times \frac{x_i - \tilde{x}}{\text{MAD}} \tag{35}$$

0.6745 is the 0.75th quartile of the standard normal distribution, to which the MAD converges.

- $\tilde{x}$ which is just the median of the sample
- MAD, is calculated by taking the absolute difference between each point and the median, and then calculating the median of those differences.

This feature scaling contributes to the robustness and accuracy of anomaly detection in the context of protecting homeowners from energy theft by identifying unusual electricity consumption patterns, as indicated in the experimental data from different homes in Table 5.

Table 11 shows the results for all the algorithms in different homes. We deployed the receiver operating characteristic curve (AUC-ROC), F1-score, precision, and recall metrics, which are often more informative in such cases. High accuracy can be achieved by simply classifying everything as benign, which does not help in detecting attacks.

In Figure 18 we plotted the ROC curves for the MLP-AE, 1D-CONV-AE, and IF for both simulated and real attack scenarios in Home A. The MLP-AE detector outperformed both 1D-CONV-AE and IF with an AUC score of 0.76 for simulated and 0.59 for real attacks, respectively while 1D-CONV-AE had ROC value 0.67 and 0.61; IF has AUC values 0.64 and 0.54 respectively.

In many real-world scenarios, the attack patterns vary and evolve constantly. Rare attack patterns can be vastly exceeded by benign data. Consequently, the autoencoder may not have sufficient examples of these rare attacks to learn effective representations, making it difficult to detect new attacks.

C. MODEL PERFORMANCE BY ROC AND CONFUSION MATRICES

Figures 17 and 18 further generate ROC curves for the remaining supervised benchmark detectors to facilitate a comparative analysis. The ROC curve provides a general view of the model's performance across all thresholds and, provides a sense of discrimination ability. In contrast, the confusion matrix provides detailed information regarding the performance of our model at a specific threshold level.

In the ROC curve, the AUC for each model (XGBoost, Random Forest, MLP) provides a single measure of performance across all possible classification thresholds, summarizing the trade-off between TPR and FPR.

The confusion matrices in Figure 19 provide a more granular view. For instance:

- The MLP for Home A simulated (Home A_sim) confusion matrix indicates that the model correctly identifies 95% of benign cases (TN), the exact value is 6655, and 89% of attack cases (TP), 6065, at a specific threshold.
- The Random Forest confusion matrix showed a high TN rate of 99%, but a lower TP rate of 84%.
- The XGBoost confusion matrix showed a similarly high TN rate of 99% and a better TP rate of 91%.

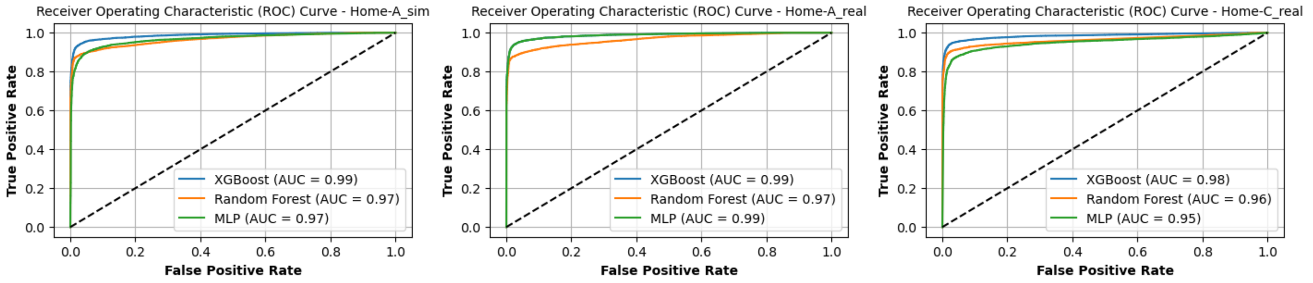
The ROC curve does not show the actual values of TP, FP, TN, and FN; rather, it shows the rate at which these values change with the different thresholds as shown in Figure 17 (a),(b), and (c), samples - taken from home A (sim and real), and home C(real attack) respectively. A high AUC reflects a model with a high TPR and low FPR across different thresholds, which generally corresponds to high values of TP and TN and low values of FP and FN in the confusion matrices at a particular operating threshold.

The same principles were applied to legacy unsupervised models (LUM). For instance, consider the confusion matrices for the simulated and real data from Home B using the 1D-CONV-AE model in Figure 18(b):

- The AUC of 0.78 and 0.72 for simulated and real data respectively on the ROC curve suggests that the model's ability to distinguish between the classes is reasonably good for simulated data and less so for real data.
- For the corresponding confusion matrices in Figure 19, we see high TN rates (0.99 for simulated, 1.00 for real) but varying TP rates (0.78 for simulated, 0.98 for real). This suggests that, while the model is quite good at identifying negative cases (benign), its performance on positive cases (attacks) is inconsistent between the simulated and real data.
- In the confusion matrices, we observed the specific number of instances that are correctly and incorrectly classified, which was reflected in the ROC curve by the closeness of the curve to the top left corner.

From the ROC curve for the isolated forest model in Figure 18(c):

- Home A Simulated has an AUC of 0.64, meaning that the model has a 64% chance of correctly distinguishing

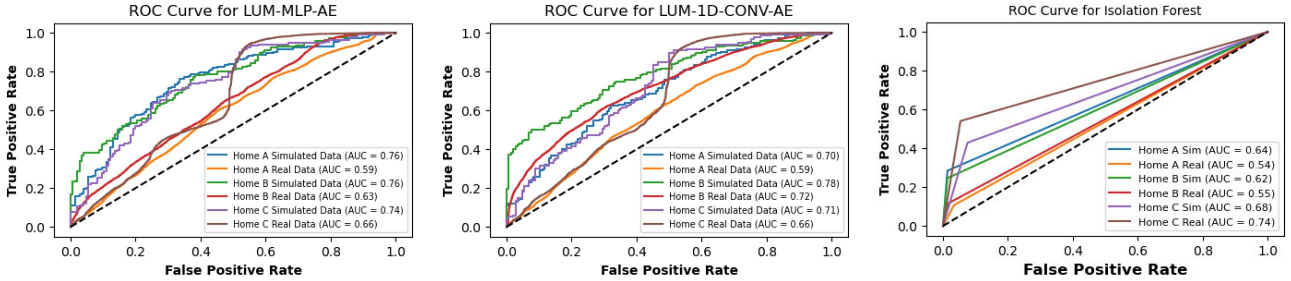


(a) ROC curves for three model-based simulated attacks at Home A

(b) ROC curves for the three models based on real attacks in Home A

(c) ROC curves for the three models based on a real attack at Home C

FIGURE 17. Sampled ROC curves comparison for performance evaluation of our proposed synthetic ETD of some selected homes for simulated and real attack.



(a) MLP-AE: sim vs. real attack

(b) 1D-CONV-AE: sim vs. real attack

(c) IF: sim vs. real attack

FIGURE 18. Performance comparison of ROC curves for legacy unsupervised models across homes.

between a benign and an attack instance for the simulated environment of Home A.

- Home A Real had a lower AUC of 0.54, suggesting that the model was less effective in distinguishing between benign and attack instances in the real-world data of Home A.

Similarly, Home B Simulated and Home B Real have AUCs of 0.62 and 0.55, respectively, and Home C Simulated and Home C Real have AUCs of 0.68 and 0.74. The higher the AUC, the better the model is at distinguishing between positive (attacks) and negative (benign) classes. For example, the model performed best on real data for Home C, with an AUC of 0.74.

From the confusion matrix for Home C real data:

- True Positive (TP): 4737 - The model correctly identified 4737 attack instances.
- True Negatives (TN): 8289 - The model correctly identified 8289 instances as benign.
- False positive (FP): 471 - The model incorrectly identified 471 benign instances as attacks.
- False Negatives (FN): 4023 - The model failed to identify 4023 attacks, mistakenly classified as benign.

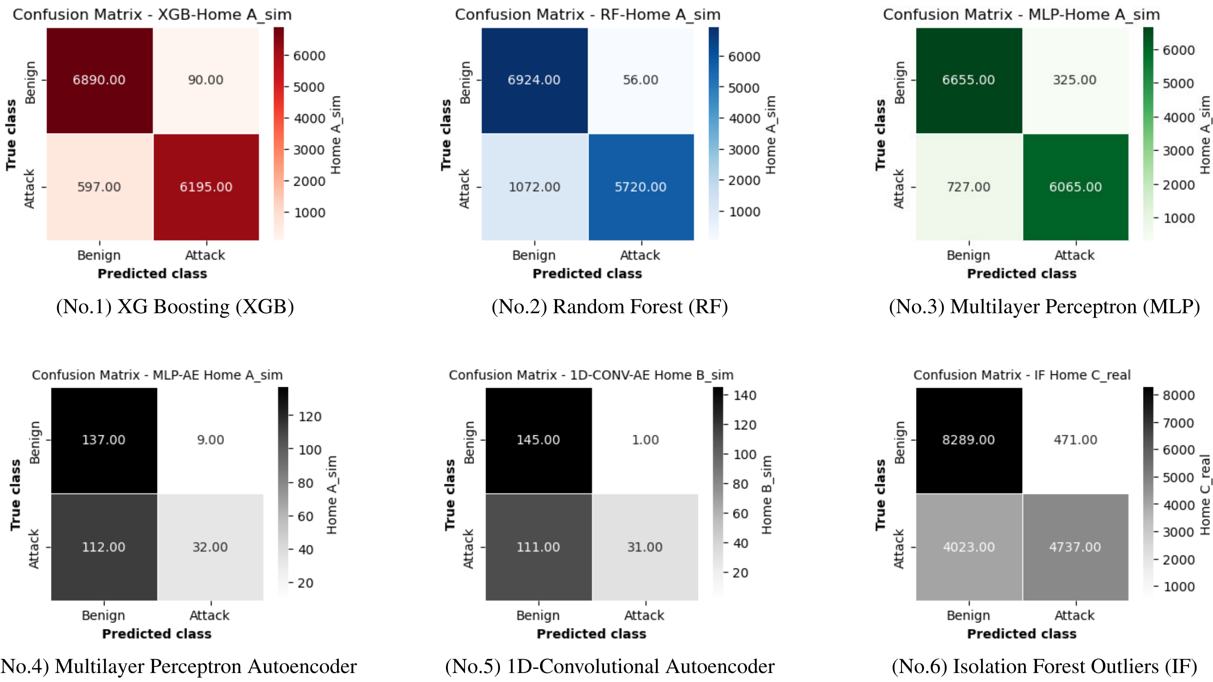
Relating the Confusion Matrix to the ROC Curve:

- The specific values in the confusion matrix correspond to a single point on the ROC curve for Home C's real data. The point is determined by the sensitivity (TPR) and FPR.
- These values would indicate a corresponding point on the ROC curve, but the exact point was not marked in the ROC curve. However, we know that point exists, and if the threshold is adjusted, this point moves along the curve, resulting in different values in the confusion matrix.

Generally, the ROC curve indicates how well the model can separate the two classes, and provides a holistic view of the model's performance across all thresholds. By contrast, the confusion matrix indicates exactly where the model makes mistakes at a specific threshold.

D. MODEL TRAINING AND TEST ERROR

To ensure balanced optimization between the training and test errors, we employed GridSearchCV for meticulous hyperparameter tuning. A five-fold cross-validation was incorporated to enhance the robustness of our model evaluations. For the XGBoost model, the regularization parameters L1 (`reg_alpha`) and L2 (`reg_lambda`) were utilized. Conversely, in the context of the Random Forest model, the parameters `min_samples_split` and `min_samples_leaf` serve a regulative function by constraining the complexity of the decision trees. The MLP model employed an L2 penalty term (`alpha`) and a maximum number of iterations (`max_iter`), combined with an early stopping criterion to prevent overfitting.



(No.1) XG Boosting (XGB)

(No.2) Random Forest (RF)

(No.3) Multilayer Perceptron (MLP)

(No.4) Multilayer Perceptron Autoencoder

(No.5) 1D-Convolutional Autoencoder

(No.6) Isolation Forest Outliers (IF)

FIGURE 19. Confusion matrices of the trained models. The number (No. #) indicates the rank of the overall accuracy (Table 8). The colors indicate the group of accuracy. The attacks class were relatively easily detected by all the synthetic models, especially achieving 99% accuracy with XGB and RF. The false positive rates of XGB, and RF were about 1% which is much better than other models.

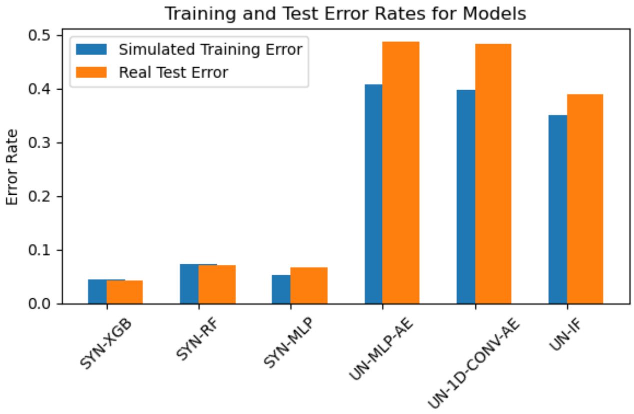


FIGURE 20. Training and test error comparison for our proposed model.

The subsequent results, as shown in Table 12, were obtained through the post-application of the aforementioned hyperparameter tuning and regularization strategies, as detailed for each model in Table 7.

The bar charts in Figure 20 show the training and test error rates for various models, based on simulated and real data aggregate performance outputs in Table 12. The error rates were calculated as one minus the AUC and accuracy (ACC) values for each model. From the charts, we can observe that for the detection of electricity theft in smart homes utilizing aggregated appliance consumption patterns, the comparative performance analysis of various models is pivotal. Our investigation encompassed the following: both supervised and unsupervised learning paradigms, with the supervised models demonstrating superior efficacy.

1) MODEL SELECTION

From our performance analysis with other ETD models for smart homes through aggregated appliance consumption patterns, the XGBoost model (SYN-XGB) emerged as a standout performer, hence it was selected as the best model for our proposed ETD for the following reasons:

It achieved the highest Area Under the Curve (AUC) of 98.69% and accuracy (ACC) of 95.54% among the evaluated models. The XGBoost (SYN-XGB) model also exhibited the lowest error rates across simulated and real datasets, indicating its robustness and high accuracy in discerning normal consumption from theft-related anomalies. The robustness of XGBoost is further enhanced by its ensemble learning framework that employs boosting [50], which has proven effective in managing anomalies and noise prevalent within electricity consumption data. Incorporating regularization within the objective function of the model serves as a bulwark against overfitting, thus facilitating better generalization to unseen data. The model's adeptness in binary classification makes it particularly well-suited for anomaly detection tasks, such as identifying instances of electricity theft in smart home environments. Given its exemplary performance across key metrics, XGBoost is highly recommended for deployment in ETD systems, where the accurate and reliable identification of theft-related irregularities is paramount.

TABLE 11. Evaluation of the ETD model with AUC and accuracy Scores.

Home	DataSet	Model	AUC	ACC
A	Sim	SYN-XGB	**98.76%**	95.21%
		SYN-RF	96.47%	91.97%
		SYN-MLP	96.56%	**97.33%**
		UN-MLP-AE	74.61%	58.28%
		UN-1D-CONV-AE	67.31%	54.83%
		UN-IF	63.55%	63.79%
	Real	SYN-XGB	**98.91%**	**95.55%**
		SYN-RF	97.02%	92.23%
		SYN-MLP	96.68%	92.43%
		UN-MLP-AE	58.96%	51.26%
		UN-1D-CONV-AE	61.12%	50.51%
		UN-IF	53.66%	53.66%
B	Sim	SYN-XGB	**98.78%**	**96.15%**
		SYN-RF	97.56%	93.59%
		SYN-MLP	97.24%	95.14%
		UN-MLP-AE	78.90%	64.58%
		UN-1D-CONV-AE	77.04%	65.63%
		UN-IF	61.64%	62.15%
	Real	SYN-XGB	**98.78%**	**96.24%**
		SYN-RF	97.64%	93.27%
		SYN-MLP	97.95%	94.41%
		UN-MLP-AE	60.61%	52.58%
		UN-1D-CONV-AE	71.77%	53.58%
		UN-IF	55.01%	55.01%
C	Sim	SYN-XGB	**98.53%**	**95.26%**
		SYN-RF	96.25%	92.69%
		SYN-MLP	95.37%	91.68%
		UN-MLP-AE	75.04%	55.20%
		UN-1D-CONV-AE	64.95%	60.22%
		UN-IF	67.66%	68.82%
	Real	SYN-XGB	**98.53%**	**95.70%**
		SYN-RF	96.54%	92.91%
		SYN-MLP	96.49%	92.93%
		UN-MLP-AE	65.64%	50.26%
		UN-1D-CONV-AE	54.95%	51.10%
		UN-IF	74.35%	74.35%

TABLE 12. Aggregate performance evaluation with AUC and ACC metric including training and test reports for all homes.

Home	DataSet	Model	AUC	ACC	Tr_Err	Ts_Err
A+B+C	Sim_av	SYN-XGB	**0.9869**	**0.9554**	**0.0446**	none
		SYN-RF	0.9676	0.9275	0.0725	"
		SYN-MLP	0.9632	0.9472	0.0528	"
		UN-MLP-AE	0.7618	0.5935	0.4065	"
		UN-1D-CONV-AE	0.6977	0.6023	0.3877	"
		UN-IF	0.6428	0.6492	0.3508	"
	Real_av	SYN-XGB	**0.9874**	**0.9583**	none	**0.0417**
		SYN-RF	0.9707	0.9280	"	0.0720
		SYN-MLP	0.9704	0.9326	"	0.0674
		UN-MLP-AE	0.6174	0.5137	"	0.4863
		UN-1D-CONV-AE	0.6261	0.5170	"	0.4830
		UN-IF	0.6428	0.6101	"	0.3899

Note: Tr_Err = Train Error, Ts_Err = Test Error

TABLE 13. Comparison of models based on accuracy and AUC scores.

Model	Multiclass Accuracy	Binary Class Accuracy	Multiclass AUC	Binary Class AUC
XGB	0.9330	0.9569	0.9851	0.9872
RF	0.8652	0.9276	0.8896	0.9692
MLP	0.6610	0.9399	0.6189	0.9668
LR	0.5537	0.8148	0.5743	0.8748
SVM	0.5220	0.9035	0.6478	0.9430

The synthetic random forest (SYN-RF) and synthetic multilayer perceptron (SYN-MLP) also perform well, although they exhibit marginally higher error rates in comparison to SYN-XGB, suggesting room for optimization.

Conversely, unsupervised models, which include the MLP Autoencoder (UN-MLP-AE), the 1D Convolutional Autoencoder (UN-1D-CONV-AE), and the IF (UN-IF), exhibit significantly higher error rates. This is particularly notable in scenarios simulating training conditions, which may point to challenges these models face in capturing complex patterns inherent to electricity theft without labeled training data. Notwithstanding, the UN-IF model demonstrates a lesser increase in error rate transitioning from simulated to real datasets, hinting at a certain level of stability in model performance despite lower overall accuracy.

The findings suggest that, in the context of smart home electricity theft detection, supervised models adeptly leverage the nuanced patterns within aggregated appliance consumption data, thus providing a strong foundation for the development of reliable theft detection systems.

2) TRADE-OFF BETWEEN TRAINING AND TEST ERRORS

In the dedicated exploration of Electricity Theft Detection (ETD) within smart homes, a critical aspect of our experimental design was ensuring equilibrium between the training and test error rates. This balance, a trade-off between training and test errors, is imperative to avert the model's overfitting to training data, which could compromise its generalization capabilities on new, unseen data, a phenomenon that could skew the detection of electricity theft.

Our methodology encompasses the strategic application of GridSearchCV to perform exhaustive hyperparameter tuning, [51], a practice that aids in identifying optimal model parameters setting recorded in Table 7. To further bolster the reliability of our findings, we used a five-fold cross-validation scheme, which provides a more rigorous validation of the model's predictive probability.

In the domain of unsupervised anomaly detection, particularly when employing autoencoders, the goal is to minimize the reconstruction error across both training and test datasets. However, too low a training error (overfitting) may result in poor generalization of the test data. To achieve a good trade-off, we deploy early stopping techniques, and validation data splits, which form the cornerstone of our strategy to fine-tune the model's complexity. Additionally, the nuanced adjustment of hyperparameters, Table 9, including the dimensionality of the encoding and hidden layers, as well as the learning rate, was instrumental in achieving a judicious balance between underfitting and overfitting.

E. MODEL PERFORMANCE COMPARISON

1) COMPARISON WITH BENCHMARK MODELS

The comparison of Table 13 with the accuracy and AUC scores and with other classifiers in Table 14 provides a clearer picture of our model performance with benchmark algorithms such as SVM and LR, each model in multiclass and binary classification tasks.

TABLE 14. Performance metrics of different multiclass classification algorithms.

Models	Accuracy	Precision	Recall	F1-score	AUC
XGB	0.9330	0.9417	0.9291	0.9174	0.9851
RF	0.8652	0.9082	0.8694	0.8701	0.8896
DTC	0.7142	0.7477	0.7068	0.7129	0.8208
MLP-4	0.6610	0.7044	0.6552	0.6812	0.6189
MLP-3	0.6582	0.6877	0.6571	0.6601	0.6097
LR	0.5537	0.5417	0.5543	0.5383	0.5743
RRC	0.5148	0.5280	0.5194	0.5178	0.5060
SVM	0.5220	0.5109	0.5189	0.5085	0.6478

TABLE 15. Proposed model performance metric comparison with binary class. benchmark.

Metric	XGB (Proposed)	RF	MLP	SVM	LR
ACC	0.9554	0.9275	0.9331	0.9035	0.8148
Recall	0.9213	0.8581	0.9472	0.8432	0.7311
Precision	0.9882	0.9916	0.9399	0.9471	0.8565
AUC	0.9869	0.9676	0.9632	0.9430	0.8748

XGB had the highest scores across the board, achieving 93.30% accuracy and 98.51% AUC in multiclass, and 95.69% accuracy and 98.72% AUC in binary classification. RF is a strong contender with 86.52% multiclass accuracy and 88.96% AUC, (Figure 21(a)), along with 92.76% binary accuracy and 96.92% AUC.Figure 21(b). MLP performs moderately with 66.10% multiclass accuracy and 61.89% AUC, improving binary classification with 93.99% accuracy and 96.68% AUC. LR and SVM, while viable, offer lower accuracy and AUC, suggesting that more advanced methods may be preferable for complex classification challenges.

The line plot presented in Figure 21(c) illustrates the error rates for various evaluation metrics across five different machine learning models: XGBoost (proposed), Random Forest (RF), Multilayer Perceptron (MLP), Support Vector Machine (SVM), and Logistic Regression (LR) derived from experimental result Table 15. The error rate for each model is computed as a $1 - metricvalue$, where the metrics include accuracy (ACC), Recall, Precision, and Area Under the Curve (AUC). Figure 21(d) shows the aggregate AUC and ACC binary class performance comparison of the proposed models with the benchmark, existing literature, SVM, and LR models.

- The **XGBoost model (proposed)** shows the most favorable error rates across all metrics, signifying its superior performance relative to the other models.
- The **Random Forest (RF)** and **Multilayer Perceptron (MLP)** models displayed competitive performance, with error rates marginally higher than those of the XGBoost model.
- The **Support Vector Machine (SVM)** and **Logistic Regression (LR)** models exhibit higher error rates, indicating that their performance is not as robust as that of the models above for the tasks evaluated.

The graph underscores the effectiveness of the XGBoost model in minimizing error rates, which correlates with the high predictive accuracy and model reliability for our ETD in smart homes.

VII. DISCUSSION AND FUTURE WORK

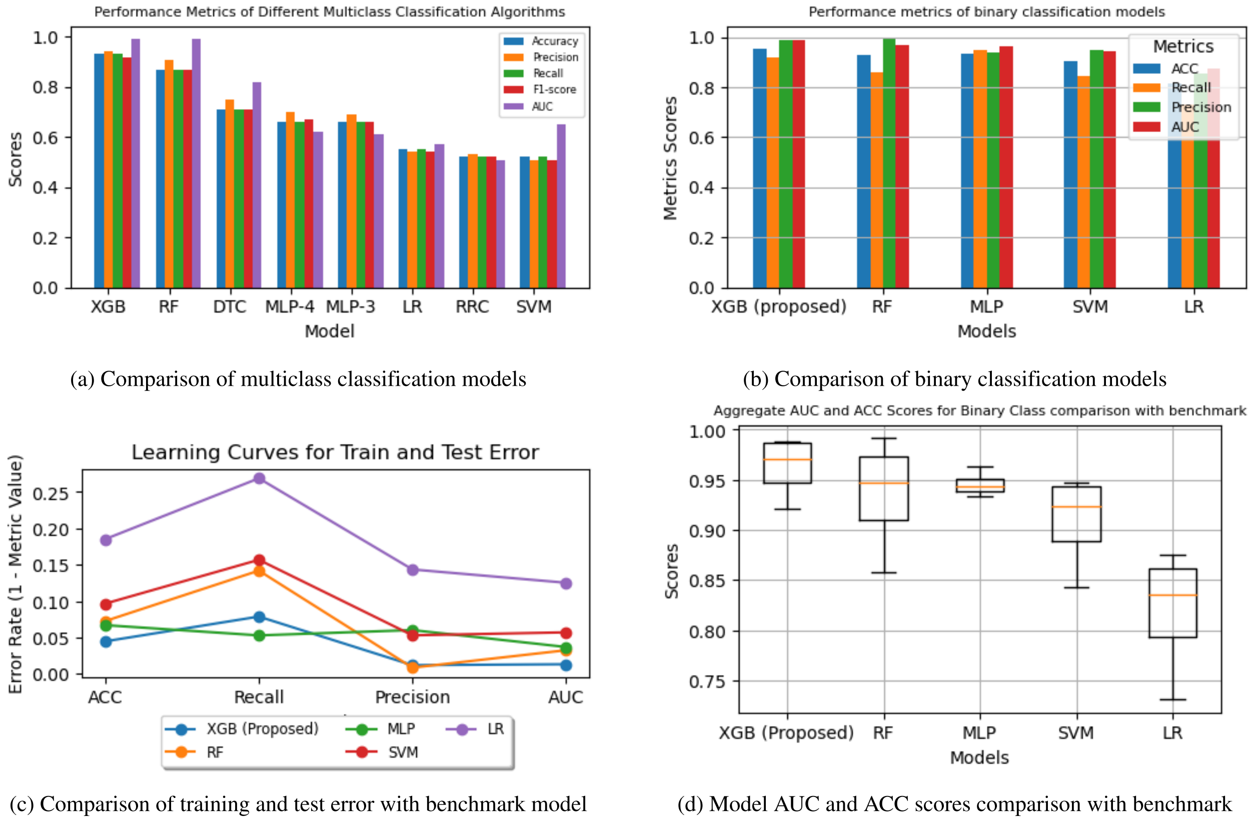
In this study, we explored the complexities inherent in electricity theft detection (ETD) within smart home environments, focusing on the use of aggregated power consumption patterns of appliances. A critical challenge in ETD is the variability of appliance consumption patterns, which can be influenced by a range of non-attack factors, such as temporary electrical spikes, periodic variations, or even permanent changes in usage habits. Such variations pose a risk of false positives in theft detection systems, where benign changes in power usage might be mistakenly identified as malicious activities.

Our synthetic binary discriminator model (SYNBDM) is designed to address these challenges effectively. It incorporates mechanisms to differentiate between short-term unusual behaviors and actual theft incidents. For instance, a transient spike in power usage, which may occur due to typical yet benign activities, is not immediately flagged as an anomaly. This approach significantly reduces the likelihood of false alarms triggered by such short-term changes. The utility of aggregated power base consumption patterns in this context cannot be overstated, as it plays a pivotal role in reducing false positives. By only reporting suspicious behavior when both the smart meters and the XGB model concurrently detect an anomaly, the system ensures a higher degree of accuracy. Consequently, a single appliance's unusual yet non-malicious behavior does not trigger a false alarm unless another appliance is simultaneously compromised, indicating a potential theft scenario.

Furthermore, to refine the system's accuracy, we emphasize the importance of calculating the false positive rate (FPR) and adjusting the sensitivity parameter, denoted as *'m'*. This adjustment is crucial, as it allows the system to avoid overreacting to sporadic or isolated incidents of unusual power usage. By configuring the system to flag theft only upon the recurrence of suspicious behavior, we significantly enhance the reliability of our ETD model.

The application of uniform manifold approximation and projection (UMAP) for clustering adds another layer of sophistication to our approach. This technique enables the algorithm to discern and adapt to various distribution patterns in the dataset, allowing for the training of separate classifiers tailored to specific usage patterns. Such adaptation is particularly beneficial in accounting for the differences in appliance usage between weekdays and weekends or across different seasons. If time-dependent patterns are observed within these clusters, the corresponding classifiers are labeled accordingly, ensuring that new instances are evaluated using the most relevant classifier for that specific time frame.

Lastly, our model is adept at identifying and adjusting to permanent changes in consumption patterns, such as those resulting from new appliances or shifts in weather conditions. This adaptability is key to maintaining the long-term effectiveness of the SYNBDM, ensuring that it remains reliable despite evolving household dynamics.



(a) Comparison of multiclass classification models

(b) Comparison of binary classification models

(c) Comparison of training and test error with benchmark model

(d) Model AUC and ACC scores comparison with benchmark

FIGURE 21. **Performance metrics comparison analysis.**

Our study not only addresses the immediate challenges of detecting electricity theft in smart homes but also lays the foundation for future advancements in this field. By considering a wide array of factors that influence power consumption and employing advanced analytical techniques, our approach demonstrates a comprehensive and robust strategy for ETD.

A. LIMITATIONS OF THE PROPOSED MODEL

Although our models mark significant advancements in detecting electricity theft in smart homes, it is imperative to acknowledge certain limitations that accompany our current methodology.

1) DEPENDENCE ON DATA QUALITY AND GRANULARITY

The effectiveness of our models is closely tied to the quality and granularity of aggregated appliance consumption data. Any inadequacies in data resolution or representativeness could potentially affect the predictive accuracy of the models.

2) ASSUMPTION OF CONSISTENT CONSUMPTION PATTERNS

Our models operate under the assumption that consumption patterns within a household remain relatively stable over time. Significant behavioral changes or the introduction of new appliances could alter these patterns, potentially affecting the model's performance until retraining occurs.

3) OVERFITTING RISKS AND MODEL COMPLEXITY

Despite the implementation of regularization techniques, there remains the risk of overfitting, particularly if the model complexity is not finely calibrated.

4) PRIVACY AND ETHICAL CONSIDERATIONS

Although our dataset was anonymized, the utilization of detailed electricity consumption data raises privacy concerns. This can be resolved using federated learning or secure multiparty computation in smart home energy consumption monitoring [52], [53]. The potential for re-identification or misuse of these data, even in an anonymized form, cannot be entirely ruled out. Ensuring ongoing compliance with privacy regulations and maintaining ethical standards for data usage are paramount.

5) COMPUTATIONAL DEMANDS AND RESOURCE CONSTRAINTS

The computational complexity associated with our models, especially in terms of hyperparameter tuning and processing, presents limitations in terms of resource allocation.

In future work, we plan to address these limitations knowing that embarking on further research in these areas will contribute to the ongoing development of robust and effective ETD systems for smart homes. As the field continues to evolve, these challenges provide exciting avenues for future exploration and innovation in the quest for more secure and reliable smart grids.

VIII. CONCLUSION

In this research, we introduced the SYNBDM and LUM algorithms for electricity theft detection in smart homes, utilizing fine-grained appliance consumption data to distinguish between normal and malicious usage. By employing uniform manifold approximation and projection (UMAP) to identify varied data distributions across different homes, our algorithm effectively leverages aggregated power consumption patterns for robust anomaly detection. Our tests on a real building appliance dataset demonstrated high performance, even with anonymized data, highlighting the algorithm's capability to balance effective theft detection with customer privacy. Though highly effective, we noted that unsupervised learning models require further refinement to better handle the complexities of real-world attack data. Our findings underline the potential of machine learning in enhancing energy security and stress the importance of incorporating appliance consumption patterns in electricity theft detection. This approach offers significant benefits to both consumers and energy providers, aiming for more efficient and secure energy management in smart homes.

CONFLICT OF INTEREST DECLARATION

The authors declare that they have no affiliations with or involvement in any organization or entity with any financial interest in the subject matter or materials discussed in this manuscript.

REFERENCES

[1] B. Lashari. (Oct. 2019). *Electricity Theft, Exploring Social Dimensions*. Accessed: Dec. 6, 2022. [Online]. Available: https://energycentral.com/c/pip/electricity-theft-exploring-social-dimensions

[2] L. Northeast Group. (Dec. 2014). *World Loses $89.3 Billion to Electricity Theft Annually, $58.7 Billion in Emerging Markets*. Accessed: Dec. 6, 2022. [Online]. Available: https://www.prnewswire.com

[3] R. Razavi and M. Fleury, "Socio-economic predictors of electricity theft in developing countries: An Indian case study," *Energy Sustain. Develop.*, vol. 49, pp. 1–10, Apr. 2019.

[4] A. Takiddin, M. Ismail, U. Zafar, and E. Serpedin, "Deep autoencoder-based anomaly detection of electricity theft cyberattacks in smart grids," *IEEE Syst. J.*, vol. 16, no. 3, pp. 4106–4117, Sep. 2022.

[5] P. Jokar, N. Arianpoo, and V. C. M. Leung, "Electricity theft detection in AMI using customers' consumption patterns," *IEEE Trans. Smart Grid*, vol. 7, no. 1, pp. 216–226, Jan. 2016.

[6] N. F. Avila, G. Figueroa, and C. Chu, "NTL detection in electric distribution systems using the maximal overlap discrete wavelet-packet transform and random undersampling boosting," *IEEE Trans. Power Syst.*, vol. 33, no. 6, pp. 7171–7180, Nov. 2018.

[7] R. Punmiya and S. Choe, "Energy theft detection using gradient boosting theft detector with feature engineering-based preprocessing," *IEEE Trans. Smart Grid*, vol. 10, no. 2, pp. 2326–2329, Mar. 2019.

[8] O. A. Abraham, H. Ochiai, M. D. Hossain, Y. Taenaka, and Y. Kadobayashi, "Electricity theft detection for smart homes with knowledge-based synthetic attack data," in *Proc. IEEE 19th Int. Conf. Factory Commun. Syst. (WFCS)*, Apr. 2023, pp. 1–8.

[9] T. Dayaratne, C. Rudolph, A. Liebman, and M. Salehi, "We can pay less: Coordinated false data injection attack against residential demand response in smart grids," in *Proc. 11th ACM Conf. Data Appl. Secur. Privacy*, Apr. 2021, pp. 41–52.

[10] S. K. Singh, K. Khanna, R. Bose, B. K. Panigrahi, and A. Joshi, "Joint-transformation-based detection of false data injection attacks in smart grid," *IEEE Trans. Ind. Informat.*, vol. 14, no. 1, pp. 89–97, Jan. 2018.

[11] A. Moradzadeh, O. Sadeghian, K. Pourhossein, B. Mohammadi-Ivatloo, and A. Anvari-Moghaddam, "Improving residential load disaggregation for sustainable development of energy via principal component analysis," *Sustainability*, vol. 12, no. 8, p. 3158, Apr. 2020.

[12] S. K. Singh, R. Bose, and A. Joshi, "Energy theft detection in advanced metering infrastructure," in *Proc. IEEE 4th World Forum Internet Things (WF-IoT)*, Feb. 2018, pp. 529–534.

[13] H. Wang and W. Yang, "An iterative load disaggregation approach based on appliance consumption pattern," *Appl. Sci.*, vol. 8, no. 4, p. 542, Apr. 2018.

[14] S. Wilhelm and J. Kasbauer, "Exploiting smart meter power consumption measurements for human activity recognition (HAR) with a motif-detection-based non-intrusive load monitoring (NILM) approach," *Sensors*, vol. 21, no. 23, p. 8036, Dec. 2021.

[15] X. Zhang, T. Kato, and T. Matsuyama, "Learning a context-aware personal model of appliance usage patterns in smart home," in *Proc. IEEE Innov. Smart Grid Technol. Asia (ISGT ASIA)*, May 2014, pp. 73–78.

[16] M. A. Devlin and B. P. Hayes, "Non-intrusive load monitoring and classification of activities of daily living using residential smart meter data," *IEEE Trans. Consum. Electron.*, vol. 65, no. 3, pp. 339–348, Aug. 2019.

[17] Z. Li, M. Shahidehpour, and F. Aminifar, "Cybersecurity in distributed power systems," *Proc. IEEE*, vol. 105, no. 7, pp. 1367–1388, Jul. 2017.

[18] M. Nabil, M. Ismail, M. Mahmoud, M. Shahin, K. Qaraqe, and E. Serpedin, "Deep learning-based detection of electricity theft cyber-attacks in smart grid AMI networks," in *Deep Learning Applications for Cyber Security*. Cham, Switzerland: Springer, 2019, pp. 73–102.

[19] S. Li, Y. Han, X. Yao, S. Yingchen, J. Wang, and Q. Zhao, "Electricity theft detection in power grids with deep learning and random forests," *J. Electr. Comput. Eng.*, vol. 2019, pp. 1–12, Oct. 2019.

[20] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, 2020.

[21] Y. Lei, B. Yang, X. Jiang, F. Jia, N. Li, and A. K. Nandi, "Applications of machine learning to machine fault diagnosis: A review and roadmap," *Mech. Syst. Signal Process.*, vol. 138, Apr. 2020, Art. no. 106587.

[22] Pamir, N. Javaid, A. Almogren, M. Adil, M. U. Javed, and M. Zuair, "RFE based feature selection and KNNOR based data balancing for electricity theft detection using BiLSTM-LogitBoost stacking ensemble model," *IEEE Access*, vol. 10, pp. 112948–112963, 2022.

[23] S. Munawar, N. Javaid, Z. A. Khan, N. I. Chaudhary, M. A. Z. Raja, A. H. Milyani, and A. A. Azhari, "Electricity theft detection in smart grids using a hybrid BiGRU–BiLSTM model with feature engineering-based preprocessing," *Sensors*, vol. 22, no. 20, p. 7818, Oct. 2022.

[24] N. Moustafa, J. Hu, and J. Slay, "A holistic review of network anomaly detection systems: A comprehensive survey," *J. Netw. Comput. Appl.*, vol. 128, pp. 33–55, Feb. 2019.

[25] S. Singh and A. Yassine, "Big data mining of energy time series for behavioral analytics and energy consumption forecasting," *Energies*, vol. 11, no. 2, p. 452, Feb. 2018.

[26] A. Aldegheishem, M. Anwar, N. Javaid, N. Alrajeh, M. Shafiq, and H. Ahmed, "Towards sustainable energy efficiency with intelligent electricity theft detection in smart grids emphasising enhanced neural networks," *IEEE Access*, vol. 9, pp. 25036–25061, 2021.

[27] R. K. Ahir and B. Chakraborty, "Pattern-based and context-aware electricity theft detection in smart grid," *Sustain. Energy, Grids Netw.*, vol. 32, Dec. 2022, Art. no. 100833.

[28] S. S. S. R. Depuru, L. Wang, V. Devabhaktuni, and P. Nelapati, "A hybrid neural network model and encoding technique for enhanced classification of energy consumption data," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, Jul. 2011, pp. 1–8.

[29] A. Jindal, A. Dua, K. Kaur, M. Singh, N. Kumar, and S. Mishra, "Decision tree and SVM-based data analytics for theft detection in smart grid," *IEEE Trans. Ind. Informat.*, vol. 12, no. 3, pp. 1005–1016, Jun. 2016.

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

[31] C. H. Park and T. Kim, "Energy theft detection in advanced metering infrastructure based on anomaly pattern detection," *Energies*, vol. 13, no. 15, p. 3832, Jul. 2020.

[32] M. Toshpulatov and N. Zincir-Heywood, "Anomaly detection on smart meters using hierarchical self organizing maps," in *Proc. IEEE Can. Conf. Electr. Comput. Eng. (CCECE)*, Sep. 2021, pp. 1–6.

[33] Y. Huang and Q. Xu, "Electricity theft detection based on stacked sparse denoising autoencoder," *Int. J. Electr. Power Energy Syst.*, vol. 125, Feb. 2021, Art. no. 106448.

[34] S. Alla and S. K. Adari, *Beginning Anomaly Detection Using Python-based Deep Learning*. Cham, Switzerland: Springer, 2019.

[35] Y. S. Chin, "Anomaly detection frameworks for identifying energy theft and meter irregularities in smart grids," Ph.D. dissertation, UM Power Energy Dedicated Adv. Center (UMPEDAC), Faculty Eng., Univ. Malaya, Kuala Lumpur, Malaysia, 2019.

[36] S. McLaughlin, B. Holbert, A. Fawaz, R. Berthier, and S. Zonouz, "A multi-sensor energy theft detection framework for advanced metering infrastructures," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 7, pp. 1319–1330, Jul. 2013.

[37] S. Makonin, B. Ellert, I. V. Bajić, and F. Popowich, "Electricity, water, and natural gas consumption of a residential house in Canada from 2012 to 2014," *Sci. Data*, vol. 3, no. 1, pp. 1–12, Jun. 2016.

[38] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," 2018, *arXiv:1802.03426*.

[39] N. Sadeghianpourhamami, J. Ruyssinck, D. Deschrijver, T. Dhaene, and C. Develder, "Comprehensive feature selection for appliance classification in NILM," *Energy Buildings*, vol. 151, pp. 98–106, Sep. 2017.

[40] Q. Xu, Y. Liu, and K. Luan, "Edge-based NILM system with MDMR filter-based feature selection," in *Proc. IEEE 5th Int. Electr. Energy Conf. (CIEEC)*, May 2022, pp. 5015–5020.

[41] A. Aldweesh, A. Derhab, and A. Z. Emam, "Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues," *Knowl.-Based Syst.*, vol. 189, Feb. 2020, Art. no. 105124.

[42] S. Chen, E. Dobriban, and J. H. Lee, "A group-theoretic framework for data augmentation," *J. Mach. Learn. Res.*, vol. 21, no. 1, pp. 9885–9955, 2020.

[43] Keras and TensorFlow2. *The Python Deep Learning Library*. Accessed: Jul. 4, 2023. [Online]. Available: https://keras.io/

[44] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[45] S. K. Gunturi and D. Sarkar, "Ensemble machine learning models for the detection of energy theft," *Electric Power Syst. Res.*, vol. 192, Mar. 2021, Art. no. 106904.

[46] A. Arif, N. Javaid, A. Aldegheishem, and N. Alrajeh, "Big data analytics for identifying electricity theft using machine learning approaches in microgrids for smart communities," *Concurrency Comput., Pract. Exper.*, vol. 33, no. 17, Sep. 2021, Art. no. e6316.

[47] M. D. Hossain, H. Inoue, H. Ochiai, D. Fall, and Y. Kadobayashi, "LSTM-based intrusion detection system for in-vehicle can bus communications," *IEEE Access*, vol. 8, pp. 185489–185502, 2020.

[48] A. Takiddin, M. Ismail, U. Zafar, and E. Serpedin, "Variational Auto-encoder-based detection of electricity stealth cyber-attacks in AMI networks," in *Proc. 28th Eur. Signal Process. Conf. (EUSIPCO)*, Jan. 2021, pp. 1590–1594.

[49] J. Rodrigues. (May 2018). *Outliers Make Us Go Mad: Univariate Outlier Detection*. Accessed: Jul. 20, 2023. [Online]. Available: https://medium.com/@joaopedroferrazrodrigues/outliers-make-us-go-mad-univariate-outlier-detection-b3a72f1ea8c7

[50] C. Zhao, D. Wu, J. Huang, Y. Yuan, H.-T. Zhang, R. Peng, and Z. Shi, "BoostTree and BoostForest for ensemble learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 8110–8126, Jul. 2023.

[51] H. Tatsat, S. Puri, and B. Lookabaugh, *Machine Learning and Data Science Blueprints for Finance*. Sebastopol, CA, USA: O'Reilly Media, 2020.

[52] P. H. Mirzaee, M. Shojafar, Z. Pooranian, P. Asef, H. Cruickshank, and R. Tafazolli, "FIDS: A federated intrusion detection system for 5G smart metering network," in *Proc. 17th Int. Conf. Mobility, Sens. Netw.*, Dec. 2021, pp. 215–222.

[53] H. M. Khan, A. Khan, F. Jabeen, A. Anjum, and G. Jeon, "Fog-enabled secure multiparty computation based aggregation scheme in smart grid," *Comput. Electr. Eng.*, vol. 94, Sep. 2021, Art. no. 107358.

OLUFEMI ABIODUN ABRAHAM (Graduate Student Member, IEEE) received the B.Tech. degree in computer science from the Federal University of Technology, Akure, Nigeria, and the M.Sc. degree in information science from the Graduate School of Information Technology, Kobe Institute of Computing, Kobe, Japan, in 2020. He is currently pursuing the Ph.D. degree with the Laboratory for Cyber Resilience, Nara Institute of Science and Technology (NAIST). He is a Research Assistant with NAIST. His research interests include cybersecurity, artificial intelligence, smart grid security, and industrial control systems security. He is a Graduate Student Member of the Power and Energy Society (PES).

HIDEYA OCHIAI (Member, IEEE) received the B.E., master's, and Ph.D. degrees in information science and technology from The University of Tokyo, Japan, in 2006, 2008, and 2011, respectively. He was an Assistant Professor with The University of Tokyo, in 2011, where he was an Associate Professor, in 2017. His research interests include the IoT system and protocol designs for peer-to-peer overlay networks, delay-disruption tolerant networks, network security, and decentralized machine learning. He joined the standardization activities of IEEE, in 2008, and ISO/IEC JTC1/SC6, in 2012. He has been the Chair of the Board of the Green University of Tokyo Project, since 2016, the LAN-Security Monitoring Project, since 2018, and the Decentralized AI Project, since 2022.

MD. DELWAR HOSSAIN (Member, IEEE) received the M.Sc. degree in engineering (information systems security) from Bangladesh University of Professionals and the Ph.D. degree in information science and engineering from Nara Institute of Science and Technology (NAIST), Japan. He is currently an Assistant Professor with the Laboratory for Cyber Resilience, NAIST. His research interests include cybersecurity, artificial intelligence, automotive security, smart grid security, and industrial control systems security. He is a member of the IEEE Communication Society.

YUZO TAENAKA (Member, IEEE) received the D.E. degree in information science from Nara Institute of Science and Technology (NAIST), Japan, in 2010. He was an Assistant Professor with The University of Tokyo, Japan. He has been an Associate Professor with the Laboratory for Cyber Resilience, NAIST, since April 2018. His research interests include information networks, cybersecurity, distributed systems, and software-defined technology.

YOUKI KADOBAYASHI (Member, IEEE) received the Ph.D. degree in computer science from Osaka University, Japan, in 1997. He is currently a Professor with the Graduate School of Information Science, Nara Institute of Science and Technology, Japan. Since 2013, he has been the Rapporteur of the ITU-T Q.4/17 for Cybersecurity Standardization. His research interests include cybersecurity, web security, and distributed systems. He is a member of the IEEE Communications Society.