Highlights

How Effective are Synthetic Attack Models to Detect Real-World Energy Theft?

Emran Altamimi, Abdulaziz Al-Ali, Abdulla K. Al-Ali, Hussein Aly, Qutaibah M. Malluhi

- Machine learning models trained on synthetic attack models may fail to classify real-world energy theft cases effectively. Moreover, detection models that excel in identifying synthetic attacks do not necessarily maintain their superiority when applied to real-world scenarios. This raises concerns about the validity of using these models as benchmarks for theft detection.

- Certain synthetic attack models enhance a model's ability to generalize and detect unseen attacks. However, other attack models significantly underperform in this aspect. The study also identifies a sufficient subset of attacks that can effectively emulate the performance of training on the full set of synthetic attacks, potentially reducing the computational complexity of the training process.

- The study found that training models on smaller, select subsets of attacks can lead to better performance compared to training on the entire range of attack scenarios. This suggests that some attack scenarios could be adding unnecessary complexity or confusion to the learned model, possibly because they mimic normal consumption behavior. Further research into identifying the most informative and representative attack scenarios could lead to more efficient and effective theft detection models.

How Effective are Synthetic Attack Models to Detect Real-World Energy Theft?

Emran Altamimi[a,b], Abdulaziz Al-Ali[b], Abdulla K. Al-Ali[a], Hussein Aly[b], Qutaibah M. Malluhi[a,b]

[a] *Department of Computer Science and Engineering, Doha, Qatar*
[b] *KINDI Center for Computing Research, Doha, Qatar*

Abstract

Advanced Metering Infrastructure plays a significant role in smart grid systems by enabling efficient two-way communication between energy suppliers and consumers. Despite its benefits, AMI faces numerous security challenges, including vulnerabilities to cyberattacks and non-technical losses, potentially leading to substantial revenue losses. Data-driven approaches to electricity fraud detection have gained popularity due to smart grids' big data. However, available datasets lack annotated real anomalies, which poses significant challenges in developing effective detection systems. As a response, researchers have developed synthetic attack models to simulate real-world theft. In this paper, we analyze the validity of popular synthetic attack models and their limitations in representing real-world theft scenarios. By conducting an empirical evaluation of synthetic attacks, we provide a practical assessment of their efficacy and their correlation with one another. Our findings suggest that while these models perform well when tested on synthetic attacks, their performance considerably deteriorates on real-world theft, considering the only real-world public dataset. Furthermore, some attack model pairs were found to be more correlated than others. Consequently, this paper suggests a sufficient subset of synthetic attacks for training and testing to effectively capture the full set of synthetic attacks.

Keywords: Energy Theft, Non-Technical Loss, Attack Models, Smart Grid, Machine learning

⋆This publication is supported in part by grant NPRP12C-33905-SP-66 from the Qatar National Research Fund. Corresponding author: Emran Altamimi (email: ea1510662@qu.edu.qa).

1. Introduction

The rapid growth of smart grids has led to increased vulnerabilities to attacks that target Advanced Metering Infrastructure (AMI), resulting in Non-Technical Losses (NTL) and repercussions including substantial revenue losses [1]. Deployment of AMI also introduces several security challenges such as privacy preservation of end users, system resilience against cyber attacks, and power theft prevention. These attacks can compromise the integrity, confidentiality, and availability of the grid, ultimately affecting the consumers' experience and hindering the expansion of AMI [2].

NTL, which are losses that can not be attributed to transmission and distribution losses, pose a significant financial impact globally, with both developing and developed countries losing substantial amounts of total electricity produced annually [3]. This issue is exacerbated by cyber actors exploiting vulnerabilities in power grids controlled by smart grid devices, leading to security risks such as incorrect customer billing, price manipulation, and power outages [4]. This emphasizes the need for robust security measures in smart grid systems.

To address the issue of electricity fraud, various NTL detection techniques have been implemented, ranging from hardware solutions to data-driven methods [5]. These techniques include machine learning-based classifiers to differentiate normal usage from theft [6], game theory approaches for detecting theft [7], and specialized devices like wireless sensors and balance meters [8]. By employing these diverse strategies, power companies can better mitigate electricity theft and secure their smart meter infrastructure [9, 10]. Data-driven methods for detecting NTL garnered much of the recent research attention, fueled by the availability of big data in smart grids, which also enable improvements in system stability, asset utilization, and customer satisfaction [11, 12]. Despite the promising utility of such methods, their development is faced with an important challenge which is the lack of datasets containing real and annotated anomaly examples. This makes it difficult to develop systems that are based on detecting abnormal behavior, such as developing effective NTL or intrusion detection systems in order to address problematic situations and security threats [13, 14, 15, 5].

To tackle the issue of unavailability of the real-world public datasets, researchers have developed attack models that they argue represent real world theft. These attack models would either lower the overall consumption of the consumer or shift the load from peak-hours to non-peak hours with the goal of reducing their overall electricity bill. As reviewed by [16] these attack models are mathematical equations that manipulate the consumption profile with some random factors, ranging from very simple attacks to sophisticated ones. The models are typically applied to real world public datasets to represent theft by consumers that are otherwise assumed to be honest. In this article, we aim to study the validity of these attack models and their limitations in representing real-world theft.

The article is structured as follows; in section II we review the related work and highlight our contribution. Section III defines our methodology and research questions for assessing the attacks. Sections IV and V present the results for synthetic and real-world tests, and the subsequent section VI provides a discussion. Finally the paper is concluded in section VII.

2. Related work

A comprehensive review of NTL detection in smart grids has been conducted in several articles. Jiang et al. [17] discuss the background of AMI and identify major security requirements that AMI should meet, as well as presenting an attack tree-based threat model for energy-theft behaviors. Fragkioudaki et al. [18] review smart metering-based energy-theft detection schemes and categorize them into system state-based and artificial intelligence-based approaches. Viegas et al. [19] provide a detailed review of solutions for NTL detection, as well as a proposed typology for categorizing these solutions. Ahmad et al. [20] focus on investigating NTL in power distribution systems and the use of consumer energy consumption information for NTL analysis.

Glauner et al. [21] explore the challenge of NTL detection using artificial intelligence, while both Wei et al. [10] and Mitra et al. [22] conduct a comprehensive review of cyber-physical attacks and counter defense mechanisms in the AMI, focusing on communication networks and smart meters. Messinis et al. [23] review NTL detection methods and provide a comprehensive list of performance metrics used in the field. The most relevant study to our research is Chuwa et al. [16], which thoroughly discusses NTL attack models, detection methods, and feature engineering methods. The paper also examines the performances of different learning models in detecting various types of attacks.

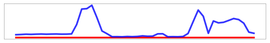
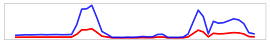
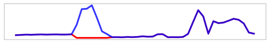
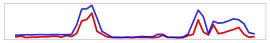
Some studies avoid the use of synthetic attack models altogether and evaluate energy theft detection systems by focusing on the most extreme cases of attacks that can elude the system [24]. These evaluations rely on introducing adversarial attacks, which subtly alter the consumption data in a way that minimizes detection while still being classified as normal by the system. While these evaluation methods are particularly relevant for unsupervised models, they also raise concerns when applied to supervised models.

In supervised learning, models are trained on a dataset with labeled examples, which means the crafted adversarial examples are dependent on the distribution modeled during training. Thee distribution is shaped by synthetic attack models included in the training data. As a result, even when using adversarial examples for evaluation, the issue of reliance on synthetic attack models remains present in supervised models.

Our work is distinct from these articles as it empirically evaluates the synthetic attack models to assess whether they truly represent real-world scenarios. By building upon the findings of these studies, particularly Chuwa et al. [16], our research focuses on a thorough discussion of the attack models, in terms of how they empirically compare to real world theft and their ability to detect unseen attacks.

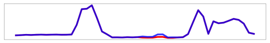
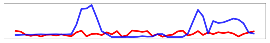
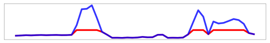
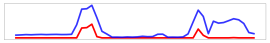
In this study, we bring forth several key contributions. Firstly, we conduct the first evaluation of synthetic attacks within a context of real-world energy theft, providing a much-needed practical assessment of these attack models. Secondly, we analyze the correlations between the synthetic attack models themselves. Lastly, we introduce the adaptation of attack methodologies to datasets that are structured around monthly consumption profiles, rather than daily ones. This approach broadens the scope of current attack models, allowing for evaluation on a real world public dataset [25].

Table 1: Summary of NTL Attack Models. For illustration, blue lines indicate actual consumption over one day, while red lines illustrate the consumption after attack

Attack ID	Brief Description	Equation	Parameters	Illustration
0	Reports zero consumption	$m_0(t) = 0$		
1	Decreases consumption by constant factor	$m_1(t) = \alpha e_t$	α: Constant random value (0,1), e_t:consumption at time t	
2	Zero consumption for random duration	$m_2(t) = \beta_t e_t$	where β_t is zero between t_s: Start time, t_e: End time and 1 otherwise	
3	Multiplies consumption by different random factors	$m_3(t) = \gamma_t e_t$	γ_t: Random factors (0,1)	

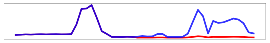
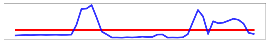
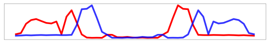
Continued on next page

Table 1 – *Continued from previous page*

Attack ID	Brief Description	Equation	Parameters	Illustration
4	Decreases consumption in random time period	$m_4(t) = \gamma_t e_t$	γ_t: Defined as α when $t_s < t < t_e$ and 1 otherwise	
5	Substitutes consumption with random proportion of the mean	$m_5(t) = \gamma_t * \text{mean}(e_t)$	γ_t: Random variable (0,1)	
6	Replaces values above cut-off point with the cut-off point	$m_6(t) = \begin{cases} e_t, & e_t \leq a \\ a, & e_t > a \end{cases}$	a: Cut-off point	
7	Replaces values below a cut-off point with zeros	$m_7(t) = \max(e_t - a, 0)$	a: Random cut-off point	

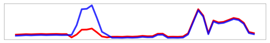
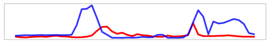
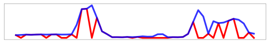
Continued on next page

Table 1 – *Continued from previous page*

Attack ID	Brief Description	Equation	Parameters	Illustration
8	Reduces usage progressively to max intensity	$m_8(t) = (1 - i_t)e_t$	i_t: Attack intensity such that it's 0 until t_s and $s(t - t_s)$ and $i_{\max}$ after $t_{\max}$ where s is the rate of change	
9	Replaces samples with average daily usage	$m_9(t) = \text{mean}(e_t)$		
10	Reverses consumption trend	$m_{10}(t) = e_{p-t}$	p: Total number of samples	

Continued on next page

Table 1 – *Continued from previous page*

Attack ID	Brief Description	Equation	Parameters	Illustration
11	Lowers consumption for certain time	$m_{11}(t) = \begin{cases} e_t - \alpha e_t, & t_s < t < t_e \\ e_t + \epsilon/(N-n), & \text{else} \end{cases}$	t_s, t_e: Start and end time of highest consumption period, n: Duration of period, N: Total samples, ϵ: Sum of reduced energy	
12	Swaps consumption with lower-consuming user	$e_t = e_{t_x}$		
13	Intermittent energy theft or malfunction	$m_{12}(t) = \delta_t e_t$	δ_t: Random binary value (0 or 1)	

3. Methodology

This study aims to evaluate synthetic attack models in NTL detection using a real-world labeled dataset. The attack models refer to mathematical equations that incorporate various factors, including random elements, to manipulate the consumption profiles. These models range from simple ones such as replacing a set of random consumption samples with zero values to more sophisticated models, such replacing the whole consumption profile with another profile that has lower overall consumption. By applying these attack models to publicly available datasets of consumer energy consumption, which are assumed to consist solely of honest consumer behavior, the objective is to represent instances of theft or fraudulent activities. The attack model IDs, description, equation, parameters and visual illustration are provided in Table 1, where e_t refers to consumption at time step t. This section outlines the steps taken to achieve the study's objectives.

3.1. Experimental setup

To evaluate the synthetic attack models, we formulate several pertinent research questions and propose corresponding experimental setups. A summary of the research questions, experiments conducted, and the goal of each experiment are provided in Table 2. The first two research questions aim to establish the best performing model for later subsequent experiments and to validate the re-conceptualization of our problem from classifying the entire consumer as thief to classifying monthly consumption into benign or abnormal. After that, we conduct three experiments to investigate the effectiveness of the synthetic attack models on real world theft, and four other experiments (see Part 2 of Table 2) to investigate the correlations between the synthetic attack models themselves.

To target RQs 1 and 2, we selected three widely-used learning algorithms for the SGCC real-world dataset [25]. These algorithms, namely, Categorical Boosting (CatBoost) [26], Support Vector Machines (SVM), and a hybrid CNN-LSTM model [27], focus on classifying the complete consumption profile. It's worth noting that the observed discrepancies between our results and those of [26, 27] stems from their use of oversampling techniques even on the testing set. We argue that oversampling examples in the test set has a high potential of misrepresenting the real-world scenario and could skew the results of most evaluation measures.

Table 2: Summary of the experiments conducted

RQ No.	Research question	Experiment setup and details	Goal of experiment
Part 1: Effectiveness of the synthetic attack models on the real world theft dataset [25]			
1	What is the expected performance on real world theft when training on the whole consumer?	Train and test on the SGCC [25] real world dataset labelled for theft. The training set is oversampled using ADASYN [28] and the testing set is not over-sampled.	Compare the results with the literature to set a benchmark for the next research question.

Continued on next page

Table 2 – *Continued from previous page*

RQ No.	Research question	Experiment setup and details	Goal of experiment
2	Given that the SGCC dataset labels entire consumer profiles as fraudulent, while our approach labels individual months due to the nature of synthetic attack models, the question arises: is it justifiable to classify on a monthly basis rather than by entire consumer profiles?	Train and test on the monthly consumption from the real world SGCC dataset and classify consumers based on the classified months of each. The preprocessing steps described in [29] were followed, and the dataset is also oversampled using ADASYN [28].	Compare the performance between the voting scheme and training on the whole consumer. If the performances are similar then classifying monthly consumption is an appropriate conceptualization of the problem.
3	Does each synthetic attack model reflect real world theft?	Train the models on each synthetic attack model and test on the real thiefs	A good performance indicates that the attack model reflects characteristics present in real world theft.

Continued on next page

Table 2 – *Continued from previous page*

RQ No.	Research question	Experiment setup and details	Goal of experiment
4	What is the sufficient performing set of attack models on real world theft?	Iteratively remove the attack that degrades the performance of the current set	A better performance of a set of attacks indicates that other attacks actually confuse the model and are not representative of real world theft

Continued on next page

Table 2 – *Continued from previous page*

RQ No.	Research question	Experiment setup and details	Goal of experiment
5	Is the performance on synthetic attacks a good indicative of their performance on real world theft?	Train and test various models on both real world theft and the synthetic attack models. Five popular learning algorithms are picked and compared, which establishes the best learning algorithm used in RQ4 and RQ5	If the models trained and tested on synthetic attack scenarios perform similarly to their performance when applied to real-world theft scenarios, it suggests that these synthetic models provide a realistic benchmark for anticipating model performance in real-world theft. Therefore, the models that outperform others on the synthetic attack models are likely to be superior in addressing real-world theft.
Part 2: Synthetic attacks analysis on the Ausgrid dataset [30]			

Continued on next page

Table 2 – *Continued from previous page*

RQ No.	Research question	Experiment setup and details	Goal of experiment
6	What is the best performing model on the synthetic attack models?	Train and test on the Ausgrid dataset [30] for 3 folds with equal representation of all attacks.	Find the best performing model for subsequent experiments.
7	Which synthetic attack models are hardest to detect?	Train on all synthetic attacks and test on each attack.	Find which attacks are the hardest and easiest to detect.
8	How does each attack perform on the rest?	Train on a single attack and test on each other attack.	Higher performances indicate that the two attacks share similar characteristics that the trained model was able to capture, similarly poor performance indicates no correlation.

Continued on next page

Table 2 – *Continued from previous page*

RQ No.	Research question	Experiment setup and details	Goal of experiment
9	What is the optimal set of attack models to train on?	Attacks were greedily added to find the sufficient set	Determine the sufficient subset of attack models, for each size of the subset, which is the most capable of detecting all other attack models.

For the rest of the experiments where the models classify *monthly* load profiles on the SGCC dataset or the *daily* profiles in the Ausgrid dataset [30], we establish the best performing learning algorithm among traditional benchmark models, Support Vector Machines (SVM) and K-Nearest Neighbors (KNN), as well as three models from highly-cited studies: the CatBoost model from Punmiya et al. (2019) [29], the XGBoost model from Yan et al. (2021) [31], and the Random Forest (RF) model from Gunturi et al. (2021) [32]. Then, we proceed with subsequent experiments using the established algorithm.

3.2. Data Preparation

To the best of our knowledge, the State Grid Corporation of China (SGCC) dataset is the only public dataset labelled for theft [25]. The dataset contains daily electricity consumption (1 reading per day) of 42,372 consumers between 1 January 2014 and 31 October 2016 (1,035 days), where 3,615 of the consumers are labeled as thieves. The labeling of the data is done per consumer rather than per day, which means each individual consumer's total of 1035 readings are labelled together regardless of the day. Since most synthetic attack models used in the literature target higher sampling rates, usually 30-minute readings per data point, we re-conceptualize the problem by grouping each month of a consumer's consumption as one data point, which is labeled as thief or not based on the original label of the consumer. Specifically, instead of classifying whole consumers as thieves we re-conceptualize the problem to detect months were theft occurred. Therefore, to ensure consistency in the analysis, synthetic attack models are modified to be applicable to the granularity of the available monthly consumption data. The adaptation followed in this paper aligns the attack models with the monthly data intervals to allow accurate analysis that can capture and simulate potential attacks on the system. For instance, let us consider attack model 13 that randomly reports either the consumer's honest consumption or a zero reading for each sample at a 30-minute interval throughout the day. To adapt this attack model to monthly consumption data, we adjust the reporting intervals. Instead of using 30-minute intervals, we report the honest consumption or zero reading for each sample at a 1-day interval throughout the month. This adjustment ensures that the attack model remains consistent when applied to the monthly consumption data. A more detailed explanation of how the attack models were adapted to monthly consumption is available in Table 3. In all experiments, we performed 10 folds cross-validation and reported the mean and standard deviation. The dataset was partitioned in a multi-step process, ensuring a balanced representation of various consumer categories. Initially, real-world thieves were segregated from the dataset. In each fold of the process, a distinct group of benign consumers was reserved, creating a 'real set' to balance the classes. The leftover benign consumers in each fold were then divided into synthetic thieves and benign users, forming a 'synthetic set.' Depending on the particular experiment's requirements, both the synthetic and real sets were subsequently subdivided into training, validation, and testing subsets. To ensure consumer data does not overlap in selected partitions, data from any particular consumer was only allowed to exist in one of the partitions only. Thus, each consumer in the validation and testing set was not previously seen by the learning algorithm. In the experiments in Part 2 of table 2, which involves the analysis of correlations between attack models, the Ausgrid dataset is utilized [30]. This dataset comprises consumption data from 300 consumers over a three-year span, from July 1, 2010, to June 30, 2013, with readings taken every 30 minutes. Following the methodologies proposed in previous studies [5, 29, 15], these attack models are applied to daily consumption profiles. Subsequently, these profiles are classified as either benign or indicative of theft.

4. Results: effectiveness of synthetic attack models on real world theft

In this section the results of the first five research questions from Table 2 are discussed in details. These questions aim to analyze the effectiveness of the attack models on real world theft. Section 5 focuses on the last four research questions aiming to analyze the correlations between the attack models using the Ausgrid dataset [30].

Table 3: Attack Models and Their Application to Daily and Monthly Data

Attack Model ID	Daily Data (48 readings)	Monthly Data (1 reading per day)	Implementation Details
1	Multiply each reading by a constant random value between 0 and 1	Multiply each daily reading by a constant random value between 0 and 1	The scale factor is picked to be between (0.1, 0.8) as done in the literature
2	Replace consumption samples with zeros for random durations	Replace daily consumption with zero for random days	Commonly picked to be 6 hours in literature, we used a week (one fourth of the whole profile similar to 6/24)
3	Multiply each reading by a different random factor between 0 and 1	Multiply each daily reading by a different random factor between 0 and 1	Same factor as attack model ID 1
4	Lower energy usage during random time periods with different durations	Lower energy usage during random days	Same time period as attack model ID 2

Continued on next page

Table 3 – *Continued from previous page*

Attack Model ID	Daily Data (48 readings)	Monthly Data (1 reading per day)	Implementation Details
5	Replace each sample with a random factor (0 to 1) multiplied by the mean consumption	Replace each daily reading with a random factor (0 to 1) multiplied by the mean daily consumption	Same factor as in attack model ID 1
6	Cap readings at a random cut-off point	Cap daily readings at a random cut-off point	Commonly picked as 0.6 of the maximum consumption
7	Remove a random cut-off point from each reading, report zero if the result is negative	Remove a random cut-off point from each daily reading, report zero if the result is negative	Same factor as attack model ID 6

Continued on next page

Table 3 – *Continued from previous page*

Attack Model ID	Daily Data (48 readings)	Monthly Data (1 reading per day)	Implementation Details
8	Reduce energy usage progressively until maximum intensity is reached, maintain this level throughout the attack	Reduce energy usage progressively over days until maximum intensity is reached, maintain this level throughout the attack	Attack intensity was a random value between (0.3, 1) and the rate of change to be between (0.05, 0.1) similar to the authors of this attack [33].
9	Replace all measured samples with the average daily energy usage	Replace all daily readings with the average monthly energy usage	
10	Reverse consumption trend of a day	Reverse consumption trend reversing a week	By reversing the week the original goal of shifting peak-hours' load to off-peak is mimicked by shifting the peak-days' of a week to off-peak days

Continued on next page

Table 3 – *Continued from previous page*

Attack Model ID	Daily Data (48 readings)	Monthly Data (1 reading per day)	Implementation Details
11	Lower energy consumption for certain times, distribute the reduced amount throughout the remaining times of the day	Lower energy consumption for certain days, distribute the reduced amount throughout the remaining days of the month	The scaled attack was 0.3 during the maximum consumption and the time duration was one week. The reduced consumption is shifted to the rest.
12	Switch consumption pattern with a customer who has a lower consumption pattern	Switch consumption pattern with a customer who has a lower consumption pattern for a month	Neighborhoods of 100 consumers were picked. The attacker reports the same consumption as a random consumer that has 0.4 or less of his total consumption.
13	Report zero consumption or original consumption sporadically for each reading	Report zero consumption or original consumption sporadically for daily readings	

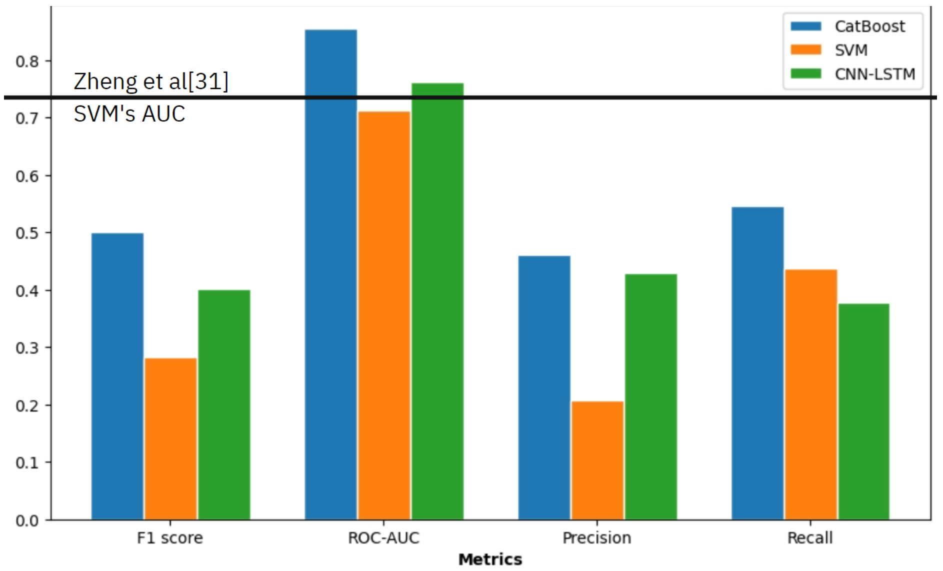


Figure 1: Performance comparison of CatBoost, SVM, and a simple CNN-LSTM model when considering each consumer's entire consumption profile as a single example.

4.1. Comparing theft detection per-month to per-consumer basis (RQ 1-2)

In this set of experiments, we compare the performance of our formulation of classifying theft on a *per-month* basis to classifying consumers' total worth of data (*per-consumer*) to assess its viability. Only real data from the SGCC data was used.

Figure 1 shows the results and also highlights the reported performance of Zheng's et. al. SVM [25] as a reference. Evidently, CatBoost outperformed the others and was thus selected. Next, we validate our conceptualization by testing this algorithm on the two formulations of the problem; *per-month*, and *per-consumer* theft detection. The evaluation of the two approaches is done by calculating the Area Under the Curve (AUC) in two ways. The *per-month* AUC is calculated by taking the predictions of each month separately and the *per-consumer* AUC is calculated by taking the average confidence of all monthly predictions of each consumer and then calculating the AUC per consumer. The results are depicted in Figure 2 which concludes that our formulation of the problem gives comparable results to the literature, allowing us to apply and evaluate the synthetic attack models in the way they were originally intended.

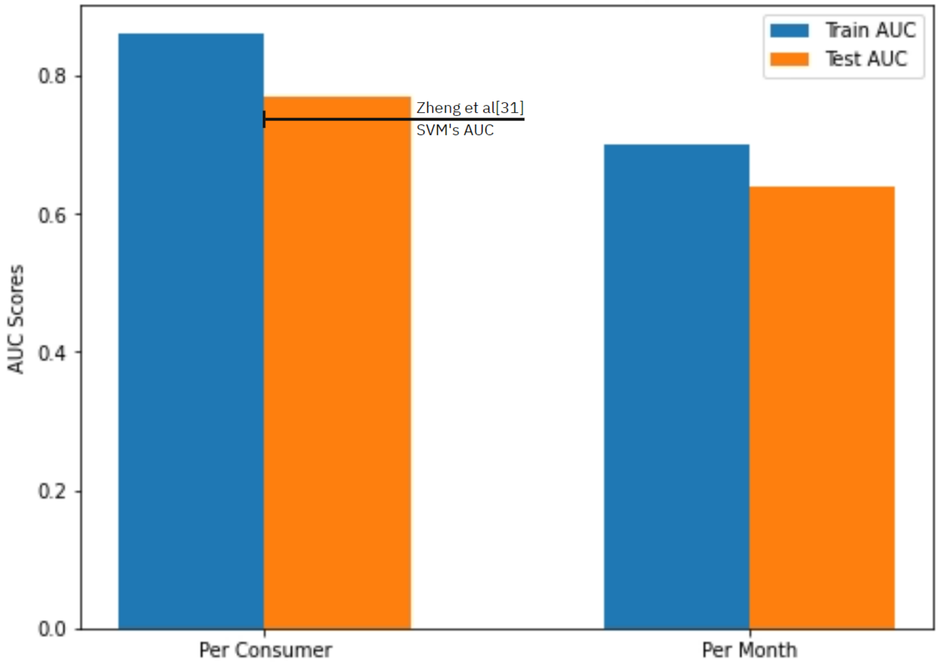


Figure 2: Classification of monthly consumption using CatBoost. AUC calculated both per month and per consumer.

4.2. Effectiveness of individual attack models on detecting real world theft (RQ 3)

In this experiment a single synthetic attack is injected (one at a time) in the training set and the performance is tested on the real theft set, as well as a separate testing set of the same synthetic attack. Figure 3 shows the performance of CatBoost, the best performing learning algorithm from a later experiment (RQ 5), when trained on each synthetic attack and tested on both synthetic and real datasets. It can be concluded that no attack performs better than random guessing on the real set, reflecting that no correlation exists between any of the individual synthetic attack models and real theft behavior. As such, none of the attacks can successfully capture real-world theft behavior in the SGCC dataset when used separately. Whether a subset of the attacks can improve detection performance is tested next.

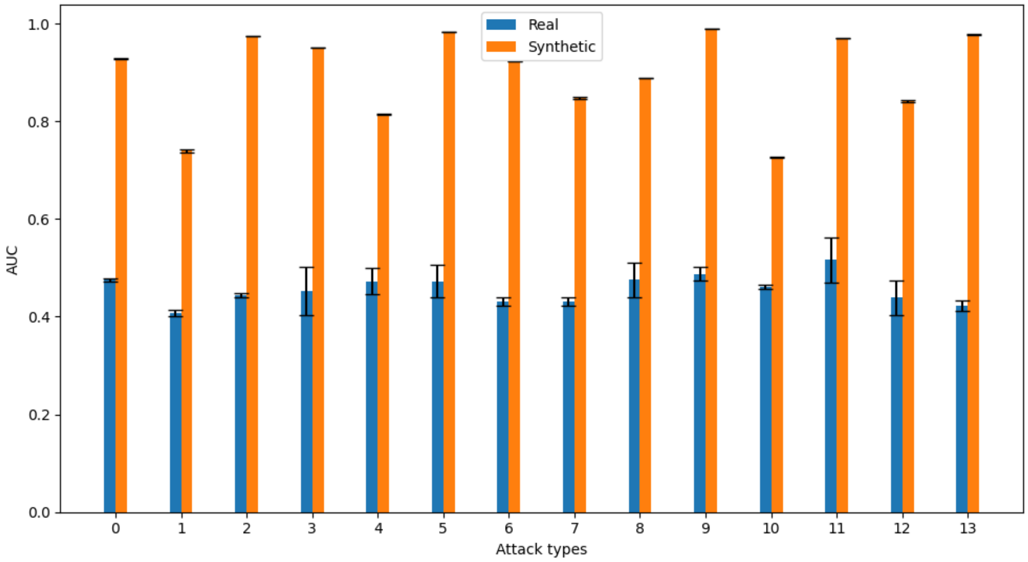


Figure 3: Performance of the CatBoost model when trained on each synthetic attack and tested on both synthetic and real data sets.

4.3. Finding an efficient subset of synthetic attacks for detecting real world theft (RQ 4)

We conducted an experiment to identify whether a sufficient representative subset of the synthetic attacks to capture real-world theft exists. To reduce computation cost, the procedure of our experiment follows a greedy approach as follows:

1. We first train a CatBoost model on all synthetic attacks injected in the training set.
2. Next, we evaluate the model's performance on the real world theft (does not include any synthetic attacks) as the testing set for all 10 folds.
3. We iteratively do the following to eliminate the least effective attack: (a) remove one of the synthetic attacks from the training set at a time, train a new model with the new subset and record it's performance. (b) once all possible removals are evaluated, we choose to eliminate the one with the least performance drop signifying that it has the least information content.
4. Step 3 is repeated to identify the next attack to be eliminated and the process is repeated until we are left with only one attack in the training set.

Through this iterative approach, we can assess the effect of different subsets of synthetic attacks on the model's performance when detecting real-world theft. Figure 4 shows that not only a representative subset of attacks does exist, the removal of some attacks from the training set even led to an increase in detection performance. This can perhaps be attributed to having some of the removed attacks exhibiting behavior similar to benign consumption patterns, thus confusing the classifier. To understand if there were meaningful differences between our baseline group (including all attacks) and the other 14 subsets (each group represents the AUC for 10 folds), we executed a paired t-test for each pair. The null hypothesis for this test was that there is no difference in the means between the baseline group and each of the other groups. Upon performing the t-tests, we obtained a p-value for each test.

Out of the 14 t-tests, 12 of them resulted in p-values smaller than 0.05 suggesting statistically significant differences between the means of those groups and the baseline group. The p-values of the other 2 groups (first (after removing of attack 12) and 6th subset(after removal of attack 9)) were higher than 0.05, implying that the differences between these groups and the baseline could be due to chance.

4.4. Comparing learning algorithms performance ranking between synthetic attack models and real world theft (RQ 5)

The results of the previous experiments indicate that synthetic attack models do not appear to accurately emulate real-world theft scenarios that are present in the SGCC dataset. This raises a critical question: if a learning algorithm demonstrates superior performance compared to others when tested on synthetic attacks, does it maintain this superior performance when confronted with real-world theft? To scrutinize this hypothesis, we have designed a final experiment, in which we evaluate the performance of the five employed classifiers. Initially, each of these models is trained and evaluated using synthetic attack datasets. Following this, we retrain and evaluate the models on real attack datasets. This two-step process allows us to compare the performance of each model on synthetic and real-world datasets, examining whether a model's rank on synthetic attacks is indeed indicative of its performance on real-world attacks. Finally, the performance of the models when trained on synthetic attack models and tested on real world theft is also included in the comparison. This comparison provides valuable insights into

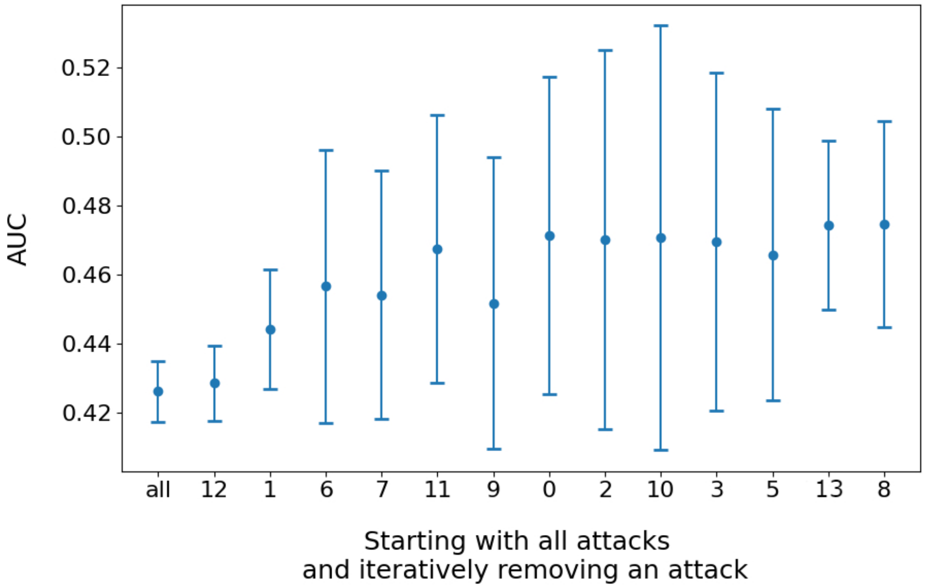


Figure 4: Sufficient subsets of synthetic attacks, where the x-axis represents the attack model removed iteratively from left to right.

whether a high performing learning algorithm tested on synthetic attacks is expected to perform best on real-world theft as well.

Figure 5 visually demonstrates the findings of the study. Firstly, the results indicate that the superiority of the existing proposed models in the literature when tested on synthetic attacks diminishes when confronted with real-world theft scenarios. Secondly, a significant observation is that most models perform similarly when tested on real-world attacks. This suggests that the challenge of accurately capturing real-world attacks may not necessarily be an issue of learning algorithm sophistication, but rather the inherent complexity of the problem itself. Consequently, these analytics suggest that an alternative problem formulation should be explored to simplify the task of the learning algorithms. It should be noted that the comparable performance of the proposed models on real-world attacks does not imply similar performance on different real-world datasets. However, these results serve as valuable evidence that performance gains on synthetic attacks does not always translate to increased performance on real-world attacks.

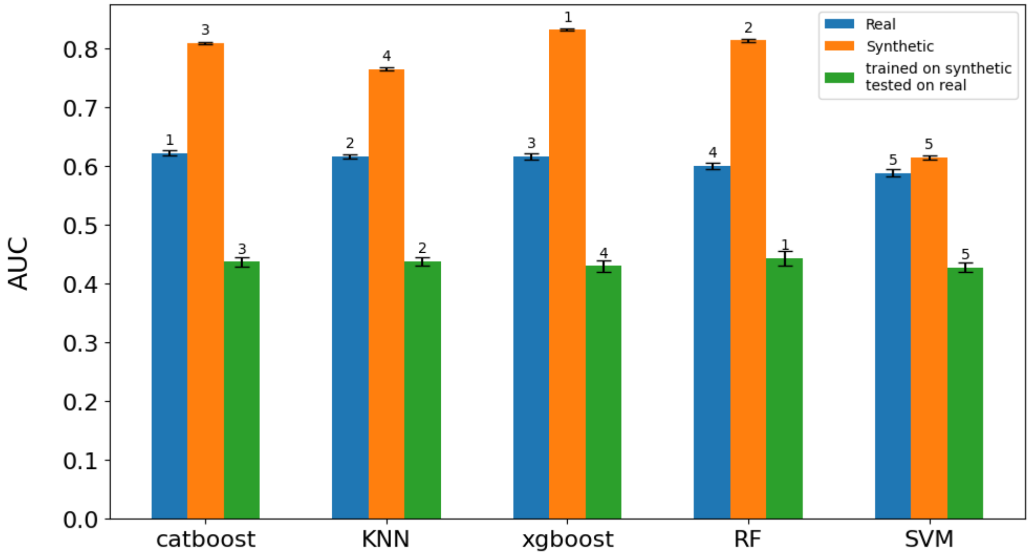


Figure 5: The AUC for 5 models on real and synthetic attacks

5. Results: Synthetic attacks analysis

In this part we aim to understand how synthetic attacks are correlated with each other. The aim of this section is to understand how each set of attacks is able to detect unseen attacks and whether some attacks represent others.

5.1. Comparing models' performance on the synthetic attack models (RQ 6)

To establish the best performing learning algorithm to be used in the next experiments, we first compare the five employed classifiers on all the attack types and found that CatBoost [29] performed the best, albeit with a small margin, as shown in Figure 6. As a result, the rest of the experiments will be done using CatBoost.

5.2. Detection rate of individual attack models (RQ 7)

In the next experiment we evaluated the attacks using a *leave-one-attack-out* fashion; by training the model on all attacks except the one we are testing on. This was done using 3 folds cross validation (using 70 % training, and 30% testing split for each fold). This experiment, as shown in figure 7, highlights which attacks are easier or harder to detect. As expected, simpler

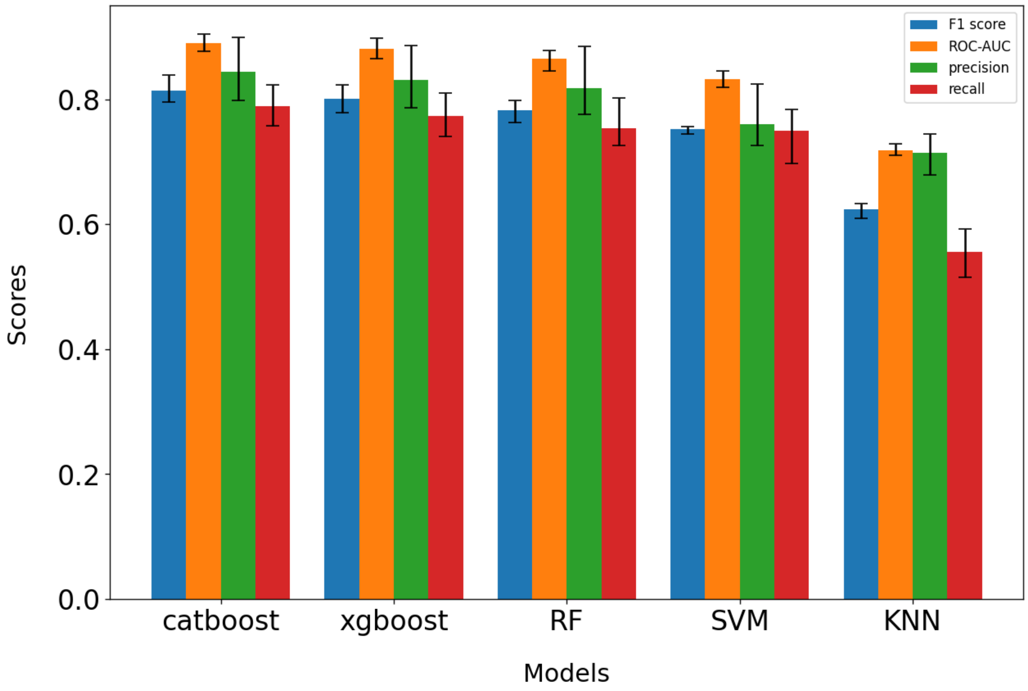


Figure 6: The evaluation metrics of the five models on the Ausgrid dataset.

attacks are easier to detect than more complex ones. For example, attack 9 (which reports the average) and 0 (zero readings) have an AUC of almost 1. On the other hand, attack 12, where attackers swap their consumption with another user, is the hardest to detect.

5.3. Correlation between the attack models (RQ 8)

To better understand the correlation between the attack models, we train the model on a single attack and test its performance on each other attack. Figure 8 highlights the results where the y-axis represent the attack the model was trained on and the x-axis represents the attack the model is tested on. The bottom row represents the average detection AUC of the attack by all other attacks and the right most column represents the average testing AUC performance of that attack on all other attacks. While some attacks are able to represent others very well such as attack 7, the same cannot be said about others. One interesting observation is that some attacks greatly confuse the learned model when trained on them that they perform worse than random

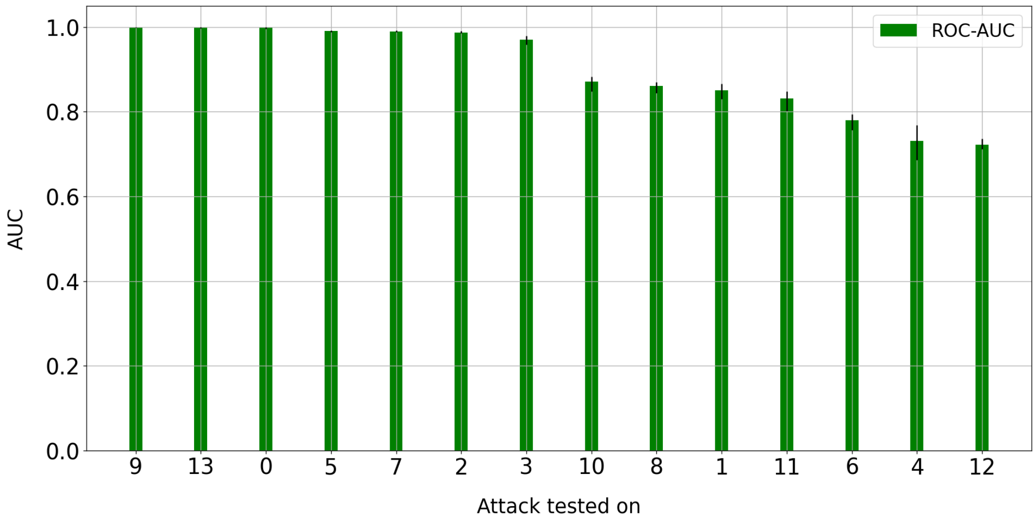


Figure 7: The performance when trained on all attacks on each attack.

guessing on other attacks. For example, when the model is trained on attack 5 and is tested on attack 9 the performance of the model is 0.21 which is worse than random guessing. This result makes sense when taking a closer look at the characteristics of both attacks; attack 5 is erratic whereas attack 9 is constant. Here, since the theft class is represented only by erratic behaviors, the model will fail to identify constant theft behaviors. This further explains why including a combination of attacks is necessary to capture various theft behaviors. We explore this concept in the next experiment.

5.4. Sufficient subset of attack models (RQ 9)

The previous experiment demonstrated that some attacks are able to capture others successfully. This experiment builds upon this by aiming to establish an efficient subset of attacks by adding the most representative attack first, followed by the next most representative attack in a similar greedy fashion to the procedure followed in section 4.3. However, the compared subsets are now tested on synthetic attacks, and the process is now incrementally adds attacks rather than removes them.

The results are shown in figure 9, and it can be noticed that the performance of the attack models converges after including attack model 10. Further additions of attacks beyond this point does not lead to substantial gains. It is crucial to note that in this experiment, attack sets are included

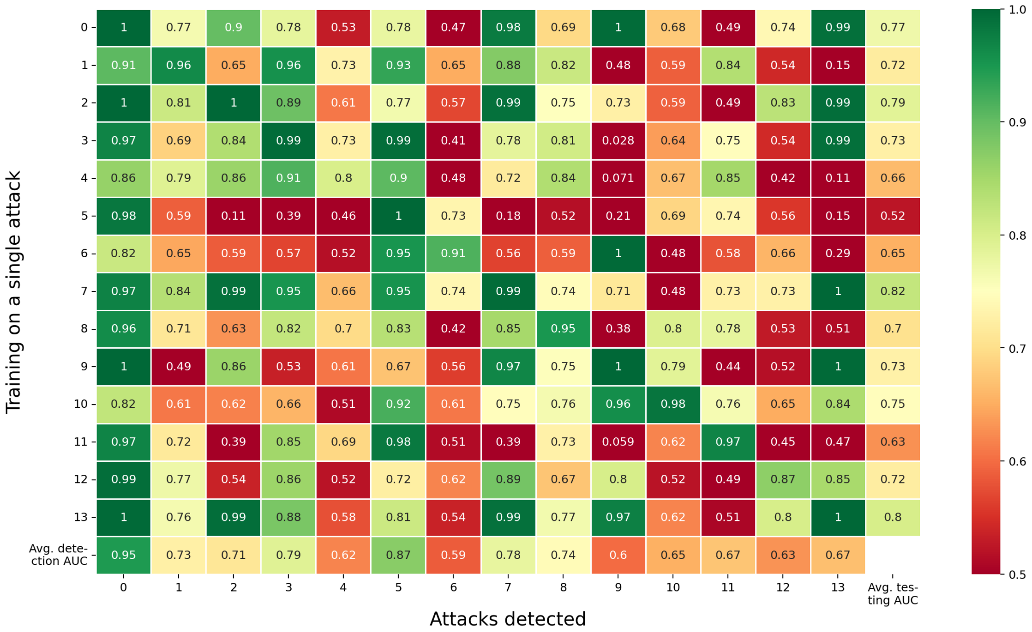


Figure 8: The performance when trained on all attacks except the removed attacks.

in the training set as opposed to results presented in (RQ8).

The results of the paired t-test comparisons indicate a statistically significant difference between the last group (including all attacks together), which we set as a baseline, and each of the two following subsets: the 3rd subset (including attacks 7, 5 and 9) and the 6th subset (including attacks 7, 5, 9, 0, 1 and 8). As a rule of thumb, we suggest using the 8th subset (including attacks 7, 5, 9, 0, 1, 8, 2 and 10) when evaluating machine learning models on synthetic attacks due to the good balance between performance and stability (low standard deviation). Attack 10 can be removed if minimization is desired since the 7th subset (including attacks 7, 5, 9, 0, 1, 8 and 2) is the minimal subset of attacks that are not significantly worse than using all attacks together. It's important to note that our findings are not definitive due to the inherent limitations of the used greedy approach. Other search heuristics and strategies could potentially identify a more optimal set, which underscores the potential for further exploration in this field.

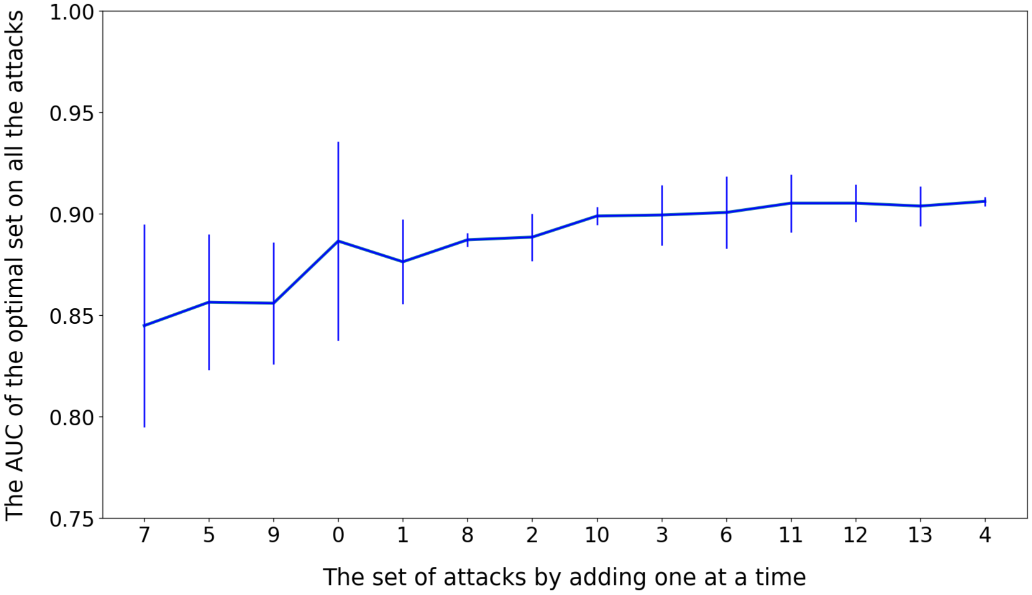


Figure 9: The performance when trained on the set of sufficient attacks and tested on the rest.

6. Discussion

The lack of a real world labelled datasets for theft motivated the need to craft attack models to evaluate the proposed machine learning models. However, while the reasoning of why such attack models are realistic were given, none were justified in a systematic manner, which motivated our work. In the first part of our work we have identified the only public labelled dataset for theft [25] and conceptualized our problem as classifying monthly data. This re-conceptualization of the problem provided similar performance to previous work that classify whole consumption profiles. We propose that this problem formulation could function as a practical theft benchmark, supplementing the customary practice of deploying attack models on high-frequency real-world datasets. By doing so, we can enhance our expectations of the model's performance concerning real-world theft scenarios. This approach ensures a more comprehensive understanding of the model's efficacy and applicability, thereby strengthening the insights derived from the study.

Subsequently, we evaluated the performance of the best performing model, CatBoost, on real world theft when trained on each attack model. It was evident that no attack model is statistically superior to normal guessing, suggesting that real world attacks follow a different distribution than those suggested in the attack models. Nonetheless, we performed an experiment that greedily finds the best optimal subset of attacks that can perform best on real world attacks. The results showed that the models performed better when trained on smaller, select subsets of attacks compared to the entire range of attack scenarios. This implies that some attack scenarios could be adding unnecessary complexity or confusion to the learned model perhaps because they mimic normal consumption behavior. On the last experiment, we ranked the performance of five classification models by comparing their detection performance on real and synthetic attacks. Interestingly, we found that while the performance on synthetic attacks aligns with the results in the literature, the performance on real attacks for all models is similar. This suggests, that at-least for this particular dataset, poor performance of the models is rather attributed to the difficulty of the examples themselves and not the models. In other words, the problem of real world theft is more complex than the synthetic attack models.

In evaluating the performance of our models, we utilized AUC as our primary metric due to its prevalent use in the literature when working on the SGCC dataset. While using AUC values is convenient to measure general performance, it is important to consider other metrics such as F1 score, precision, and recall to highlight the sources of error. Furthermore, as evidenced in RQ 5, the performance difference among various models on the real-world dataset was minor in terms of AUC. Consequently, practitioners must judiciously consider the trade-off between the choice of evaluation measure, the incremental performance enhancement, and the complexity introduced by more sophisticated learning models. For example, a utility provider may prefer to optimize recall to reduce the chance of false negatives (i.e., uncaught theft). This careful evaluation ensures the most efficient and effective deployment of these models in practice.

We then shifted our focus to explore attacks that are harder to detect, the correlation between different synthetic attacks, and the evaluation of attacks' ability to detect other unseen attacks. First, we tested detection performance on each attack when a model is trained on attacks altogether. The results followed intuition where attacks that resemble benign consumption the most were the hardest to detect. The experiment after tried to study the correlation between attacks (RQ 8). We found that some attacks, such as, 7, 13, 2 and 0, allow the model to generalize well on other unseen attacks. On the other hand, training on attacks 5, 11, 6, and 4 had very poor performance.

The issue of poor performance of the learning algorithms when trained on some attack models and tested on others suggests that a slight deviation from the distribution of the trained attack model (which is expected in real world theft) would deteriorate the learning alogrithm's performance. Observations from our study indicate that the relationships between various attack models are not symmetrical. In other words, learning algorithms trained on attack model A and tested on attack model B will not necessarily have the same performance when trained on attack model B and tested on attack model A. This complex relationship among the synthetic attack models points to a compelling area for future study. By carefully crafting new attack models that encapsulate the unique characteristics of the existing models, it is plausible to enhance the performance in detecting real-world theft. This underscores an emerging research direction that emphasizes the need for employing nuanced approaches in the design of theft detection models.

We have also applied a greedy approach to find a sufficient subset of synthetic attack models that demonstrate comparable performance to the set of all synthetic attacks. The results suggest that a subset that includes attacks 7, 5, 9, 0, 1, 8 and 2 can capture the same information as including all the attacks during training. That being said, there are numerous other possible heuristics to find optimal subsets such as iteratively adding the hardest attack model to detect or iteratively removing attack models that are the easiest.

7. Conclusion

This study employs a set of experiments aimed at analysing and studying how popular synthetic attack models, summarized in Table 1, compare to real world theft and whether they are correlated to each other. One finding of the study is that formulating the real world dataset [25] as classifying monthly examples is viable and achieves results similar to the literature. This motivates evaluating energy theft detection models on monthly consumption examples from real world labelled datasets of theft behaviors. When using such datasets, the learning algorithms discern synthetic theft from a lesser sample resolution—31 monthly readings compared to the usual 48 daily readings. One important finding from our experiments is that machine learning models that are trained on synthetic attack models can fail to classify real world attacks. Furthermore, detection models that had superior performance on synthetic attacks did not have the same advantage when tested on real world attacks, which raises the concern of their validity of being used as benchmark detection models. Another observation is that some attacks, namely 7, 13, 2, and 0, are found to enhance the model's generalizability for unseen attacks, while other attacks such as attacks 5, 11, 6, and 4 significantly underperform in this aspect. The issue of generalizability suggests that even slight deviations from the trained attack model distribution could adversely impact the model's performance, which is to be expected in real world scenarios. Lastly, when targeting accurate detection of synthetic attacks, we identified a sufficient subset of attacks that competently emulate the performance of training on the full set of synthetic attacks. This exploration introduces the possibility of reducing the set of synthetic attacks used during training.

8. Acknowledgement

This publication is supported in part by grant NPRP12C-33905-SP-66 from the Qatar National Research Fund. The findings achieved herein are solely the responsibility of the authors.

References

[1] A. O. Otuoze, M. W. Mustafa, R. M. Larik, Smart grids security challenges: Classification by sources of threats, Journal of Electrical Systems and Information Technology 5 (2018) 468–483.

[2] Y. Yan, R. Q. Hu, S. K. Das, H. Sharif, Y. Qian, An efficient security protocol for advanced metering infrastructure in smart grid, IEEE Network 27 (2013) 64–71.

[3] I. Kaspersky, Threat landscape for industrial automation systems, 2021.

[4] A. S. Musleh, G. Chen, Z. Y. Dong, A survey on the detection algorithms for false data injection attacks in smart grids, IEEE Transactions on Smart Grid 11 (2019) 2218–2234.

[5] P. Jokar, N. Arianpoo, V. C. Leung, Electricity theft detection in ami using customers' consumption patterns, IEEE Transactions on Smart Grid 7 (2015) 216–226.

[6] V. Badrinath Krishna, G. A. Weaver, W. H. Sanders, Pca-based method for detecting integrity attacks on advanced metering infrastructure, in: Quantitative Evaluation of Systems: 12th International Conference, QEST 2015, Madrid, Spain, September 1-3, 2015, Proceedings 12, Springer, 2015, pp. 70–85.

[7] S. Amin, G. A. Schwartz, A. A. Cardenas, S. S. Sastry, Game-theoretic models of electricity theft detection in smart utility networks: Providing new capabilities with advanced metering infrastructure, IEEE Control Systems Magazine 35 (2015) 66–81.

[8] K. Xie, Y.-H. Song, J. Stonham, E. Yu, G. Liu, Decomposition model and interior point methods for optimal spot pricing of electricity in deregulation environments, IEEE Transactions on Power Systems 15 (2000) 39–50.

[9] L. Wei, A. H. Moghadasi, A. Sundararajan, A. I. Sarwat, Defending mechanisms for protecting power systems against intelligent attacks, in: 2015 10th System of Systems Engineering Conference (SoSE), IEEE, 2015, pp. 12–17.

[10] L. Wei, L. P. Rondon, A. Moghadasi, A. I. Sarwat, Review of cyber-physical attacks and counter defense mechanisms for advanced metering infrastructure in smart grid, in: 2018 IEEE/PES Transmission and Distribution Conference and Exposition (T&D), IEEE, 2018, pp. 1–9.

[11] C. Tu, X. He, Z. Shuai, F. Jiang, Big data issues in smart grid–a review, Renewable and Sustainable Energy Reviews 79 (2017) 1099–1107.

[12] W. Luan, J. Peng, M. Maras, J. Lo, B. Harapnuk, Smart meter data analytics for distribution network connectivity verification, IEEE Transactions on Smart Grid 6 (2015) 1964–1971.

[13] Y. Li, W. Xue, T. Wu, H. Wang, B. Zhou, S. Aziz, Y. He, Intrusion detection of cyber physical energy system based on multivariate ensemble classification, Energy 218 (2021) 119505.

[14] E. Villar-Rodriguez, J. Del Ser, I. Oregi, M. N. Bilbao, S. Gil-Lopez, Detection of non-technical losses in smart meter data based on load curve profiling and time series analysis, Energy 137 (2017) 118–128.

[15] K. Zheng, Q. Chen, Y. Wang, C. Kang, Q. Xia, A novel combined data-driven approach for electricity theft detection, IEEE Transactions on Industrial Informatics 15 (2018) 1809–1819.

[16] M. G. Chuwa, F. Wang, A review of non-technical loss attack models and detection methods in the smart grid, Electric Power Systems Research 199 (2021) 107415.

[17] R. Jiang, R. Lu, Y. Wang, J. Luo, C. Shen, X. Shen, Energy-theft detection issues for advanced metering infrastructure in smart grid, Tsinghua Science and Technology 19 (2014) 105–120.

[18] A. Fragkioudaki, P. Cruz-Romero, A. Gómez-Expósito, J. Biscarri, M. J. de Tellechea, Á. Arcos, Detection of non-technical losses in smart distribution networks: A review, Trends in Practical Applications of Scalable Multi-Agent Systems, the PAAMS Collection (2016) 43–54.

[19] J. L. Viegas, P. R. Esteves, R. Melicio, V. Mendes, S. M. Vieira, Solutions for detection of non-technical losses in the electricity grid: A review, Renewable and Sustainable Energy Reviews 80 (2017) 1256–1268.

[20] T. Ahmad, Non-technical loss analysis and prevention using smart meters, Renewable and Sustainable Energy Reviews 72 (2017) 573–589.

[21] P. Glauner, J. A. Meira, P. Valtchev, R. State, F. Bettinger, The challenge of non-technical loss detection using artificial intelligence: A survey, arXiv preprint arXiv:1606.00626 (2016).

[22] S. Mitra, B. Chakraborty, P. Mitra, Smart meter data analytics applications for secure, reliable and robust grid system: Survey and future directions, Energy 289 (2024) 129920.

[23] G. M. Messinis, N. D. Hatziargyriou, Review of non-technical loss detection methods, Electric Power Systems Research 158 (2018) 250–266.

[24] D. Mashima, A. A. Cárdenas, Evaluating electricity theft detectors in smart grid networks, in: Research in Attacks, Intrusions, and Defenses: 15th International Symposium, RAID 2012, Amsterdam, The Netherlands, September 12-14, 2012. Proceedings 15, Springer, 2012, pp. 210–229.

[25] Z. Zheng, Y. Yang, X. Niu, H.-N. Dai, Y. Zhou, Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids, IEEE Transactions on Industrial Informatics 14 (2017) 1606–1615.

[26] S. Hussain, M. W. Mustafa, T. A. Jumani, S. K. Baloch, H. Alotaibi, I. Khan, A. Khan, A novel feature engineered-catboost-based supervised machine learning framework for electricity theft detection, Energy Reports 7 (2021) 4425–4436.

[27] M. N. Hasan, R. N. Toma, A.-A. Nahid, M. M. Islam, J.-M. Kim, Electricity theft detection in smart grid systems: A cnn-lstm based approach, Energies 12 (2019) 3310.

[28] H. He, Y. Bai, E. A. Garcia, S. Li, Adasyn: Adaptive synthetic sampling approach for imbalanced learning, in: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence), IEEE, 2008, pp. 1322–1328.

[29] R. Punmiya, S. Choe, Energy theft detection using gradient boosting theft detector with feature engineering-based preprocessing, IEEE Transactions on Smart Grid 10 (2019) 2326–2329.

[30] Ausgrid's solar home electricity data, 2020. URL: `https://www.ausgrid.com.au/Industry/Our-Research/Data-to-share/Solarhome-electricity-data`.

[31] Z. Yan, H. Wen, Electricity theft detection base on extreme gradient boosting in ami, IEEE Transactions on Instrumentation and Measurement 70 (2021) 1–9.

[32] S. K. Gunturi, D. Sarkar, Ensemble machine learning models for the detection of energy theft, Electric Power Systems Research 192 (2021) 106904.

[33] G. M. Messinis, A. E. Rigas, N. D. Hatziargyriou, A hybrid method for non-technical loss detection in smart distribution grids, IEEE Transactions on Smart Grid 10 (2019) 6080–6091.