applied sciences

MDPI

Article

XAI-Based Accurate Anomaly Detector That Is Robust Against Black-Box Evasion Attacks for the Smart Grid

Islam Elgarhy [1,2], Mahmoud M. Badr [3,4], Mohamed Mahmoud [1,*], Maazen Alsabaan [5], Tariq Alshawi [6] and Muteb Alsaqhan [5]

[1] Department of Electrical and Computer Engineering, Tennessee Technological University, Cookeville, TN 38505, USA; iaelgarhy42@tntech.edu

[2] Department of Computer Systems, Faculty of Computer and Information Sciences, Ain Shams University, Cairo 11566, Egypt

[3] Department of Network and Computer Security, College of Engineering, SUNY Polytechnic Institute, Utica, NY 13502, USA; badrm@sunypoly.edu

[4] Department of Electrical Engineering, Faculty of Engineering at Shoubra, Benha University, Cairo 11629, Egypt

[5] Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11451, Saudi Arabia; malsabaan@ksu.edu.sa (M.A.); 438105121@student.ksu.edu.sa (M.A.)

[6] Department of Electrical Engineering, College of Engineering, King Saud University, Riyadh 11421, Saudi Arabia; talshawi@ksu.edu.sa

* Correspondence: mmahmoud@tntech.edu

Abstract: In the realm of smart grids, machine learning (ML) detectors—both binary (or supervised) and anomaly (or unsupervised)—have proven effective in detecting electricity theft (ET). However, binary detectors are designed for specific attacks, making their performance unpredictable against new attacks. Anomaly detectors, conversely, are trained on benign data and identify deviations from benign patterns as anomalies, but their performance is highly sensitive to the selected threshold values. Additionally, ML detectors are vulnerable to evasion attacks, where attackers make minimal changes to malicious samples to evade detection. To address these limitations, we introduce a hybrid anomaly detector that combines a Deep Auto-Encoder (DAE) with a One-Class Support Vector Machine (OCSVM). This detector not only enhances classification performance but also mitigates the threshold sensitivity of the DAE. Furthermore, we evaluate the vulnerability of this detector to benchmark evasion attacks. Lastly, we propose an accurate and robust cluster-based DAE+OCSVM ET anomaly detector, trained using Explainable Artificial Intelligence (XAI) explanations generated by the Shapley Additive Explanations (SHAP) method on consumption readings. Our experimental results demonstrate that the proposed XAI-based detector achieves superior classification performance and exhibits enhanced robustness against various evasion attacks, including gradient-based and optimization-based methods, under a black-box threat model.

Keywords: security; evasion attacks; explainable artificial intelligence; anomaly detector; electricity theft; smart grid

check for updates

Citation: Elgarhy, I.; Badr, M.M.; Mahmoud, M.; Alsabaan, M.; Alshawi, T.; Alsaqhan, M. XAI-Based Accurate Anomaly Detector That Is Robust Against Black-Box Evasion Attacks for the Smart Grid. *Appl. Sci.* **2024**, *14*, 9897. https://doi.org/10.3390/app14219897

Academic Editors: Ryan Gibson and Hadi Larijani

Received: 13 September 2024
Revised: 9 October 2024
Accepted: 26 October 2024
Published: 29 October 2024

1. Introduction

In smart power grids, smart meters (SMs) are used in advanced metering infrastructures (AMIs) to enable two-way communication between consumers and electricity providers for continuous load monitoring and billing purposes [1–4]. However, AMIs are vulnerable to electricity theft (ET) cyber-attacks, where consumers compromise their SMs to report false power consumption readings [5–7]. ET poses a significant challenge for electricity providers, as it can lead to erroneous energy decisions due to reliance on false power consumption readings in load monitoring and energy management. Moreover, ET has negative economic consequences, with annual losses estimated in the billions of dollars.

For example, the annual losses in the United States and Canada can amount to up to six billion dollars, and these figures are even higher in developing countries [8–11].

Several machine learning (ML)-based ET detectors have been proposed in the literature. These detectors adopt either supervised or unsupervised detection. In supervised ET detection, binary detectors are trained using both benign and malicious consumption readings [5,12,13]. While they achieve high performance on learned (seen) attacks, they are limited in their ability to detect new (unseen) attacks. In anomaly ET detection, unsupervised detectors are trained using only benign readings to learn benign consumption patterns [14,15]. Anomaly detectors are effective in detecting new attacks because they focus on deviations from learned benign patterns rather than specific known attack patterns. However, these detectors achieve high performance only under the assumption of ideal threshold selection to separate benign and malicious data. Moreover, both binary and anomaly detectors are vulnerable to adversarial evasion attacks, where malicious consumers can steal electricity by making a minimal change to low-consumption malicious readings to evade detection [16–18].

Most of the existing defense mechanisms against evasion attacks, including adversarial training [19], defensive distillation [20], and diversity ensemble [21], are designed for binary detectors. They also often sacrifice model accuracy to improve robustness against evasion attacks. On the other hand, there are few defense mechanisms proposed to secure anomaly detectors, including approximate projection [22], principal latent space [23], and sequential ensemble [24]. These mechanisms are sensitive to threshold settings because their robustness relies on enhancing decoder output to maximize reconstruction error for adversarial samples in auto-encoders (AEs) [25].

An intriguing connection has been revealed between adversarial attacks and explainable artificial intelligence (XAI), where adversarial evasion samples result in anomalous XAI model explanations [26,27]. This suggests the potential for *using XAI as a defense mechanism against adversarial evasion attacks* [26–29]. XAI is primarily developed to enhance human understanding of decisions made by ML black-box models. Among XAI techniques, Shapley Additive Explanations (SHAP) [30] method stands out as one of the most widely used methods for interpreting ML models. It employs a unified approach to provide explanations for model predictions by utilizing Shapley values, which originate from cooperative game theory. These Shapley values indicate how each feature in the input data contributes to the model's prediction output.

This paper focuses on securing ML-based ET anomaly detectors against adversarial evasion attacks using XAI. As far as we know, this is *the first work that investigates the SHAP explanations (interpretations) of consumption readings for the detection of evasion attacks by training on these explanations*. Unlike other defense mechanisms, we demonstrate the potential of XAI to secure anomaly detectors against evasion attacks while maintaining the detector's accuracy and relaxing the assumption of selecting ideal threshold values. In particular, we use SHAP explanations of consumption readings to train a cluster-based hybrid anomaly detector that combines a Deep Auto-Encoder (DAE) and a one-class support vector machine (OCSVM). The utilization of XAI alongside DAE+OCSVM anomaly detection brings multiple benefits, including improved classification performance, enhanced robustness against evasion attacks, and the ability to detect zero-day attacks without requiring the selection of optimal threshold value. This occurs because the SHAP explanations effectively distinguish between normal and abnormal consumption patterns, facilitating the identification of anomalies caused by evasion samples. Additionally, training the DAE+OCSVM anomaly detector on SHAP explanations of consumption readings, rather than the readings themselves, enhances classification accuracy. Furthermore, while DAE shows the capability to detect zero-day attacks, its performance is sensitive to the selection of good threshold values. The incorporation of OCSVM overcomes this problem by automatically determining the threshold. The main contributions of our work include the following:

- We propose a hybrid DAE+OCSVM anomaly detector for ET detection. The experimental results indicate that the proposed DAE+OCSVM detector overcomes existing

limitations in the literature, including the inability of binary detectors to detect new (unseen) attacks and the sensitivity of DAE anomaly detectors' performance to the selection of the optimal threshold;
- We investigate the vulnerability of the DAE+OCSVM anomaly detector to gradient- and optimization-based evasion attacks. The experimental results indicate its vulnerability to benchmark evasion attacks, including the Fast Gradient Sign Method (FGSM), Basic Iterative Method (BIM), Carlini Wagner (C&W), Zeroth-Order Optimization (ZOO), and DeepFool;
- We propose a robust and accurate cluster-based DAE+OCSVM ET anomaly detector by training it on the SHAP explanations of consumption readings. The design objectives include enhancing accuracy, robustness, and the ability to detect new ET cyber-attacks. The experimental results illustrate the robustness of our detector against all the experimented evasion attacks while maintaining high accuracy.

The remaining sections of this paper are organized as follows. We explore the prior research conducted in securing ET detection and XAI against adversarial attacks and discuss the limitations and research gaps in Section 2. In Section 3, the ET cyber-attacks, evasion attacks, and threat model addressed in this paper are discussed. Moreover, the proposed robust and accurate ET detector is presented in Section 4. Section 5 discusses the dataset, experimental scenarios, and performance evaluation. Finally, the paper is concluded in Section 6.

2. Related Work

This section reviews the state-of-the-art works on securing ML-based ET detectors against adversarial attacks, and the uses of XAI in the context of adversarial attacks. It also discusses existing limitations in the literature and our proposed research motivations.

2.1. *Securing ET Detectors Against Adversarial Attacks*

Few papers in the literature have investigated securing ML-based ET detectors against adversarial attacks. They consider either binary detectors [31–35] or anomaly detectors [24,36].

Li et al. [31] proposed the SearchFromFree evasion attack algorithm and studied its impact on binary detectors. The authors assessed the vulnerability of three different types of neural networks (NNs) and found that malicious consumers could successfully evade detection, even in black-box threat scenarios, while reporting extremely low energy consumption measurements. Moreover, they proposed using a distillation approach to secure ET detectors against evasion attacks. Badr et al. [32] utilized GAN to propose a new evasion attack. This attack generates fabricated low consumption readings that can evade a global ET detector with a 97% attack success rate. In contrast to [32], which targets binary-based global detectors trained on a diverse range of consumption patterns, this work focuses on a different type of evasion attack that employs gradient-based and optimization-based methods specifically within anomaly-based detectors. The attack presented in [32] is inapplicable in our case because we do not use a global detector in this paper. Moreover, in a previous work by Elgarhy et al. [33–35], they evaluated the robustness of a cluster-based ET binary detector against gradient- and optimization-based evasion attacks. They further proposed using parallel ensemble learning to create a robust cluster-based ET detector under different threat models.

In contrast to [31–35], Takiddin et al. [24] utilized an electricity reading and its neighboring readings to generate evasion samples.The authors observed significant degradation in the performance of benchmark detectors in black-, gray-, and white-box threat scenarios, with decreases of up to 22.2%, 26.9%, and 35.8%, respectively. To counter evasion attacks, they used solely benign data to train a sequential ensemble-based anomaly detector that improved the ET detection rate under evasion attacks. Moreover, in [36], Takiddin et al. investigated the impact of poisoning attacks on both general and consumer-based ET detectors. They showed that both types of detectors suffered from a significant detection rate deterioration of up to 17% due to these attacks. They proposed a sequential ensemble detector based on a DAE with attention, gated recurrent units, and feed-forward NNs, to enhance the robustness of the detectors, achieving stable detection performance with a maximum deterioration of only 3% even under strong poisoning attacks.

2.2. XAI and Adversarial Attacks

XAI was originally developed to express the reasoning behind ML model outputs, thereby enhancing their reliability and trustworthiness. However, it can also play a role in detecting adversarial evasion attacks [26–28]. This can involve training new detectors using XAI explanations or providing these explanations to expert analysts for identifying adversarial samples. Additionally, XAI has been employed to enhance evasion attacks and develop more sophisticated attack strategies [37,38].

Fidel et al. [26] presented a novel detection approach for evasion attacks in binary deep neural network (DNN) classifiers. They utilized SHAP values computed for the internal layers of the binary classifier to distinguish between normal and adversarial inputs across the popular CIFAR-10 and MNIST image classification benchmark datasets. The proposed adversarially trained detector demonstrates high detection accuracy and strong generalization ability in detecting various evasion attacks. By varying the hyperparameters and applying different levels of perturbations, their detector can capture diverse patterns of adversarial samples. Similarly, in [28], Al-Essa et al. proposed an XAI adversarial training defense by fine-tuning their technique using the SHAP explanation of FGSM adversarial samples. They used two benchmark cybersecurity datasets, including Android malware and network traffic security. Watson et al. [27] introduced an accurate and model-agnostic explainable detection method for adversarial samples using SHAP values. The proposed detector achieved a detection accuracy of 77% and 88% on Electronic Health Records and Chest X-ray datasets, respectively. They introduced both fully and semi-supervised methods, capable of generalizing to different attack methods without requiring retraining. Moreover, their work led to an improvement of over 10% in the existing adversarial detection for both datasets.

Unlike the use of XAI as a defense against adversarial attacks in [26–28], Amich et al. [37] employed the feature-based explanations of model predictions to guide the crafting of adversarial samples. Their attack involved adding consequential perturbations, which are likely to induce model evasion while avoiding non-consequential perturbations that are unlikely to contribute to evasion. The proposed attack is model-agnostic and applicable across various threat models, model architectures, and distance metrics. It enhances the evasion rate of state-of-the-art attacks while requiring fewer perturbations across both white-box and black-box threat models on the CIFAR-10 and MNIST datasets. Zhang et al. [38] proposed a white-box and non-targeted attack that generates adversarial inputs by misleading target DNNs and deceiving their coupled interpretation models, namely, ADV2. They demonstrated that, with ADV2, the adversary can arbitrarily designate an input's prediction and interpretation in skin cancer diagnosis.

2.3. Limitations and Research Gaps

In the context of ET detection, most existing studies focus on designing binary detectors. However, their effectiveness in detecting new attacks is often limited and unpredictable. Anomaly detection offers an alternative solution, but anomaly detectors typically suffer from low detection accuracy and their performance is highly dependent on finding an optimal threshold. This is a challenging task, especially when malicious data are not known. Additionally, both binary and anomaly detectors are vulnerable to adversarial evasion attacks. While securing binary detectors against evasion attacks has received extensive attention from the research community, the robustness of anomaly detectors against such attacks has been less explored. This paper aims to bridge this research gap by addressing the following limitations:

- Most existing defense mechanisms, such as adversarial training [19] and defensive distillation [20], are tailored to specific evasion attacks and their performance is unpredictable in the case of new attacks. Moreover, these defense mechanisms sacrifice model accuracy to improve robustness against evasion attacks and are primarily designed for binary detectors, rendering them unsuitable for anomaly detectors;

- Few defense mechanisms have been proposed for AE-based anomaly detection [22–25]. These mechanisms are primarily utilized for applications other than smart grids, with only [24] specifically designed for smart grid use. All of these mechanisms primarily aim to improve the AE's decoder output to maximize the reconstruction error for adversarial samples. However, determining the optimal reconstruction error threshold requires some prior knowledge of the malicious data (i.e., the nature of attacks), which may not be possible practically.

These limitations are addressed in this paper by proposing a hybrid anomaly detector that combines DAE and OCSVM to overcome the difficulty in determining the optimal threshold value. Additionally, we propose using SHAP explanations obtained from consumption readings to train a cluster-based anomaly detector, enhancing both its detection accuracy and robustness against evasion attacks.

3. Evasion Attacks and Threat Model

In this section, we first present ET attacks and evasion attacks targeted for undetectable ET and then explain the different types of attackers considered in this paper.

3.1. Evasion and ET Attacks

Jokar et al. [14] have proposed a continuous attack to emulate how a malicious consumer manipulates actual consumption readings to reduce electricity bills. The continuous attack function, denoted by f_1, is modeled by Equation (1), where E_b is the benign (or actual) reading and β is a constant reduction factor ($0 < \beta < 1$).

$$f_1(E_b) = \beta \times E_b \tag{1}$$

The objective of evasion attack algorithms is to manipulate malicious samples E_m, generated by the continuous attack f_1, by introducing slight perturbations to deceive ET detectors and classify them as benign. The following explains various algorithms used to create undetectable evasion samples, including gradient-based and optimization-based algorithms.

Gradient-Based Algorithms. These attack algorithms generate evasion samples using the cost function gradient with respect to the ML model's input in both single-step and multi-step methods. FGSM [19] is a single-step attack. To generate an evasion sample, it modifies the input sample in the direction (sign) of the gradient to maximize the cost function of the correct class, as indicated by Equation (2). On the other hand, BIM [39] is a multi-step attack, modeled by Equation (3). It modifies the input sample like FGSM, then clips it after each step to keep the perturbation within acceptable limits, thereby generating an evasion sample. Here, the hyper-parameters (ϵ, α, and I) control the perturbation amount, y denotes the input sample's label, θ denotes the ML model's parameters, J_θ denotes ML model's cost function, and ∇ denotes ML model's gradient.

$$E_m + \epsilon \; . \; sign\Big(\nabla_{E_m} J_\theta(E_m, y, \theta)\Big) \tag{2}$$

$$clip_{(-\epsilon,+\epsilon)}\Big(E_{m_i} + \alpha \; . \; sign\Big(\nabla_{E_{m_i}} J_\theta(E_{m_i}, y, \theta)\Big)\Big) \tag{3}$$

Optimization-Based Algorithms. These attack algorithms solve an optimization problem to generate evasion samples. C&W [40] minimizes the changes that δ made to a malicious sample to be classified as benign as indicated by Equation (4), where $r > 0$ represents a regularization parameter and $Z[(E_m)]$ represents the predicted probability (i.e., logits layer representation) that E_m belongs to input label y as malicious and target label t as benign. However, ZOO [41] operates under the constraint of lack of direct access to the gradients of the model and can only observe the model's output for a given input. It modifies the loss function $f(E_m, t)$ in Equation (4) such that it only depends on the output of the model F and the desired class label t. Therefore, it computes an approximate gradient using a finite difference method instead of actual backpropagation on the targeted model and solves the optimization problem via zeroth-order optimization, as indicated by Equation (5). DeepFool [42] is an efficient optimization-based iterative method. It aims to identify the smallest perturbation δ that is capable of inducing misclassification by shifting the input across the decision boundary at each iteration i, as indicated by Equation (6). Unlike gradient-based attacks, optimization-based attacks require additional computational resources, making the generation of evasion samples more expensive.

$$\begin{gathered} minimize_\delta\ D(E_m, E_m + \delta) + r \cdot f(E_m, t) \\ f(E_m, t) = max_{y \neq t}[Z(E_m)]_y - [Z(E_m)]_t \end{gathered} \tag{4}$$

$$\begin{gathered} minimize_\delta\ D(E_m, E_m + \delta) + r \cdot f(E_m, t) \\ f(E_m, t) = max_{y \neq t}[log(F(E_m))]_y - [log(F(E_m))]_t \end{gathered} \tag{5}$$

$$\begin{gathered} minimize_\delta\ ||\delta||_2 \\ sign(f(E_{m_i} + \delta))\ \neq sign(f(E_{m_{i-1}})) \end{gathered} \tag{6}$$

3.2. Threat Model

In this paper, we consider a black-box threat model where the attacker exploits transferability [43] using five adversarial evasion attacks, including FGSM, BIM, C&W, ZOO, and DeepFool. In this threat model, the attacker does not know the training dataset, model architecture, and the fact that the defense model uses XAI. Therefore, he/she trains a surrogate model of different architectures, including a convolutional neural network (CNN) and a feed-forward neural network (FFNN) on a different training dataset. After that, he/she attacks the surrogate model utilizing benchmark evasion attacks to generate evasion samples, as illustrated in Figure 1. These samples are then sent to the defense model, hoping that they evade it.

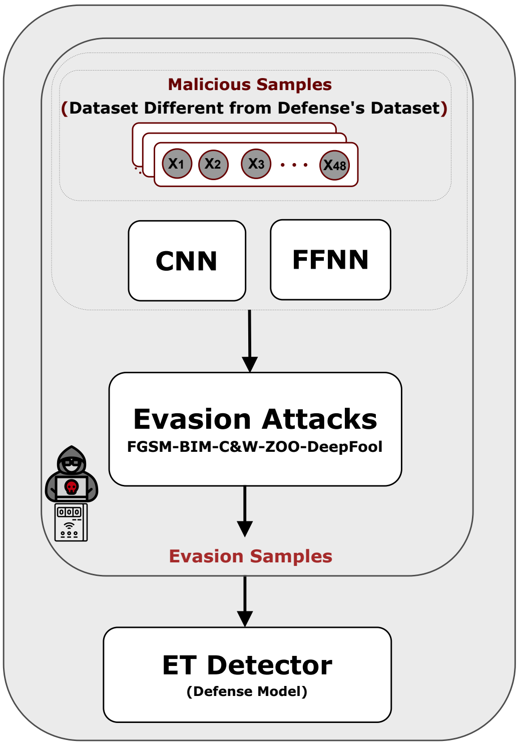


Figure 1. Black-box threat model.

4. Proposed Robust and Accurate ET Anomaly Detector

This section first explains the architecture of the proposed robust and accurate ET anomaly detector and then it discusses the rationale behind its design.

4.1. The Proposed Detector's Architecture

The proposed anomaly detector consists of four components: clustering, XAI, DAE, and OCSVM, as illustrated in Figure 2. The first component is a clustering algorithm that divides electric utility consumers into groups based on their associated metadata, which directly influences their electricity consumption, such as maximum contracted power, house size, occupancy, etc. [32]. Here, we utilize the K-means clustering algorithm to divide consumers into groups based on their consumption levels due to the absence of metadata in the dataset. The second component is the XAI SHAP method that extracts SHAP explanations from the consumption readings. Here, we extract the SHAP explanations for the consumption reading samples. The SHAP algorithm [30] computes Shapley values for each feature (i.e., consumption reading) in each sample across all data samples. These values are determined based on the marginal contribution calculated using a reference subset through iterative calculations. Figure 3 shows an example of the SHAP values for the top-20 features (readings). Practically, we train using all the features (i.e., explanations for all 48 readings). The third component is a DAE that generates latent space representations for the SHAP explanations. The fourth component is an OCSVM that uses the latent space representations to detect ET. Figure 2 depicts both phases of the proposed detector, including training and inference.

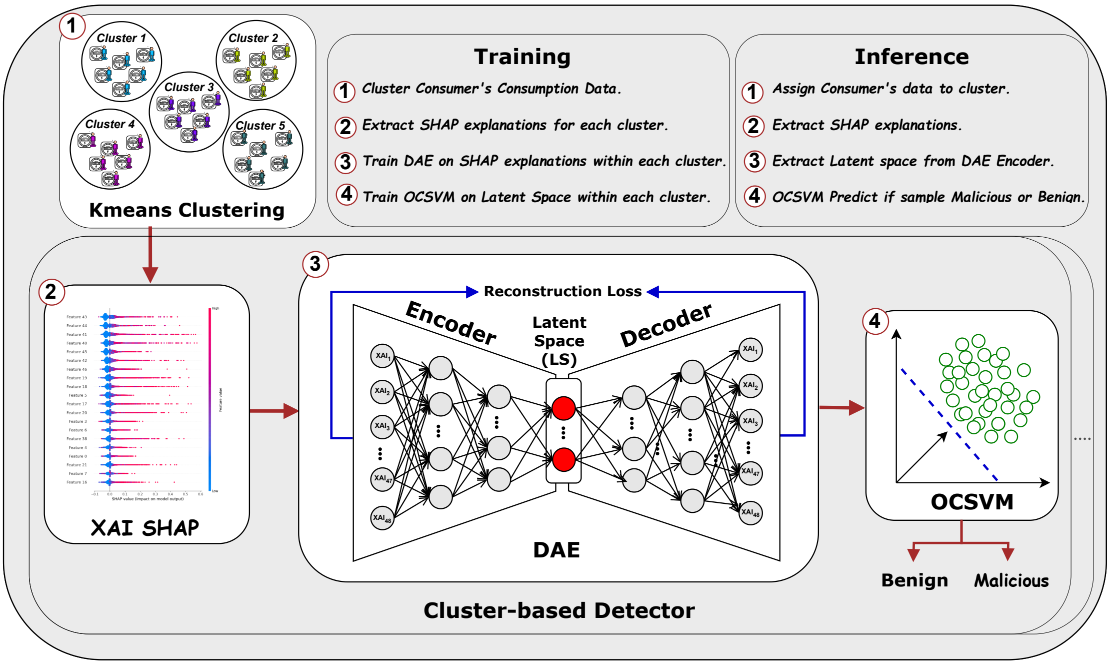
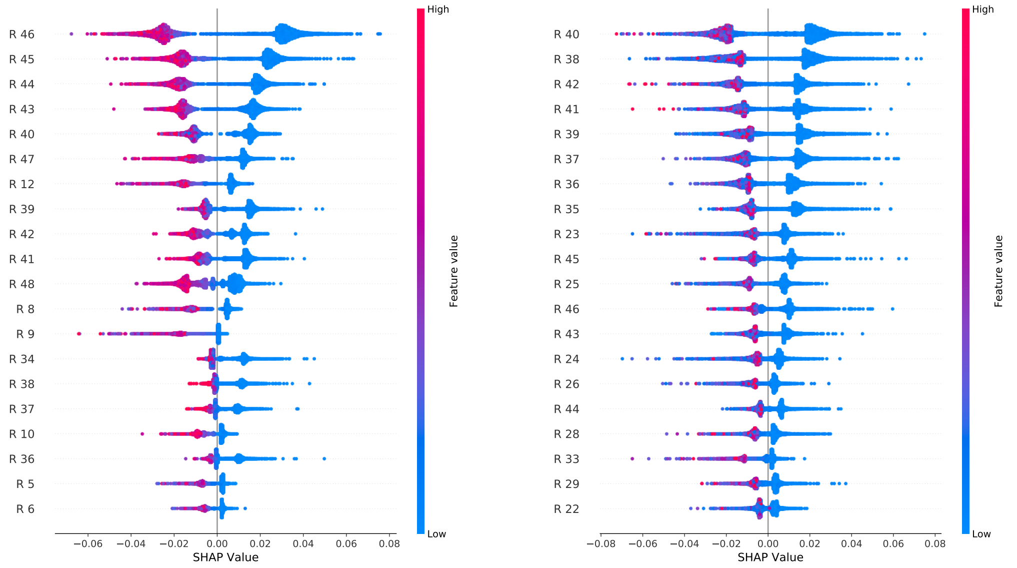

1
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Kmeans Clustering
Training
1 Cluster Consumer's Consumption Data.
2 Extract SHAP explanations for each cluster.
3 Train DAE on SHAP explanations within each cluster.
4 Train OCSVM on Latent Space within each cluster.
Inference
1 Assign Consumer's data to cluster.
2 Extract SHAP explanations.
3 Extract Latent space from DAE Encoder.
4 OCSVM Predict if sample Malicious or Benign.
2
XAI SHAP
3
Reconstruction Loss
Encoder
Latent Space (LS)
Decoder
XAI_1
XAI_2
XAI_3
XAI_{47}
XAI_{48}
DAE
4
OCSVM
Benign
Malicious
Cluster-based Detector

Figure 2. The proposed XAI-based robust and accurate ET anomaly detector.



(**a**) IRISH Dataset. (**b**) PRECON Dataset.

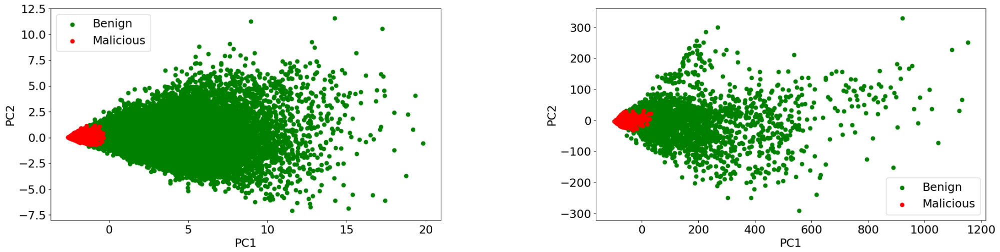
Figure 3. SHAP values: top-20 features (readings).

4.2. The Rationale Behind the Detector's Design

The rationale behind our robust and accurate ET detector is illustrated as follows:

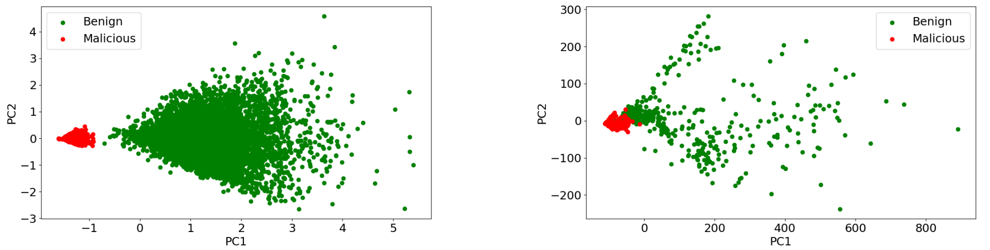
1. There is a trade-off between the level of generalization and robustness against evasion attacks in ML-based detectors [16,44,45]. Therefore, since a cluster-based detector requires less generalization, it leads to increased robustness compared to global detectors. This is because the cluster-based model is trained on data with close consumption patterns, resulting in superior performance and robustness [32,33]. To probe more deeply into the reasons behind this superiority, we applied principal component analysis (PCA) to the consumption readings of the dataset in global and cluster-based settings, as shown in Figure 4 and Figure 5, respectively. PCA allows one to reduce the dimensionality, which facilitates the visualization of complex relationships within the data. Analyzing the plots of the first two PCA components in these figures reveals a notable overlap between benign and malicious consumption patterns in the global setting as opposed to the cluster-based setting. Therefore, we opted for a cluster-based detector rather than a global detector;
2. There is a deep connection between XAI model explanations and adversarial evasion samples. Intuitively, a model's XAI explanation leads to robustness against adversarial evasion samples because evasion samples often result in anomalous XAI explanations [27,29]. To delve deeper into the reasons behind this, we applied PCA to the SHAP explanations of the consumption readings of the dataset in a cluster-based setting, as shown in Figure 6. It is evident from the figure that the SHAP explanations of benign and malicious consumption patterns are significantly distinct. Additionally, upon examining the cumulative variance explained by the principals component, we observe that approximately 90% of the data variance is explained by the first two components of the XAI explanations, compared to only 50% for the consumption readings. Therefore, SHAP explanations are capable of compressing larger amounts of information more efficiently than consumption readings, i.e., with a lower number of PCA components. Consequently, our detector is trained using the SHAP explanations of consumption readings, rather than the readings themselves;
3. Unsupervised anomaly detectors are trained solely on benign data to detect various malicious activities by identifying deviations from learned benign patterns without needing malicious datasets during the training phase. However, they use malicious data to determine the ideal reconstruction threshold for superior detection performance. Comparing DAE and OCSVM anomaly detectors, the DAE achieves superior

detection performance because its deep structure extracts relevant features from the input data, thereby enhancing detection [15,46]. However, the performance of the DAE is susceptible to threshold selection. Conversely, the OCSVM does not require finding optimal threshold values, which may be difficult to find without any knowledge of malicious data [47]. Therefore, we propose a hybrid anomaly detector that combines DAE and OCSVM, achieving superior performance while eliminating the need for determining optimal threshold values.



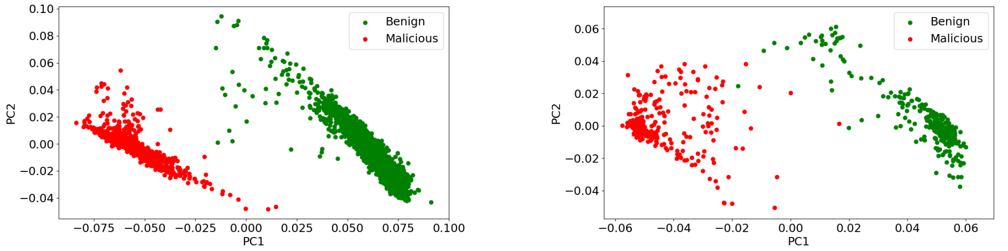
(**a**) IRISH Dataset.
(**b**) PRECON Dataset

Figure 4. PCA applied to global consumption readings.



(**a**) IRISH Dataset.
(**b**) PRECON Dataset

Figure 5. PCA applied to cluster-based consumption readings (e.g., Cluster 1).



(**a**) IRISH Dataset.
(**b**) PRECON Dataset

Figure 6. PCA applied to the cluster-based SHAP explanations of consumption readings (e.g., Cluster 1).

5. Performance Evaluation

This section starts with a description of the dataset, followed by the experimental setup and evaluation metrics. Then, it delves into the experimental results.

5.1. Dataset

5.1.1. IRISH Dataset

The IRISH dataset [48] was created by Electric Ireland and the Sustainable Energy Authority of Ireland. It contains only benign electricity consumption readings from 3639 residential consumers ($\mathcal{C}_{irish}$) recorded at half-hourly intervals. The data span 536 days and were collected between 2009 and 2010.

5.1.2. PRECON Dataset

The PRECON dataset [49] was collected over a period of one year, capturing electricity consumption patterns for 42 residential consumers $(\mathcal{C}_{precon})$ with varied demographics in Lahore, Pakistan, at 1 minute intervals. The dataset spans 321 days and was collected between 2018 and 2019.

5.1.3. Dataset Preparation

The dataset is collected from benign residential consumers. However, practical data collection processes can lead to issues such as erroneous data, missing values, and unintentional anomalies (outliers). These issues can arise from equipment failures, inaccuracies, aging components, transmission errors, and poor connections. Therefore, we first use the preparation methods outlined in [50–52] to ensure that this dataset is clean and free from these issues. Then, we use the cleaned dataset in our experiments to detect intentional cyber-attacks. Due to a lack of publicly available malicious consumption datasets, we use the continuous attack f_1, specified in Equation (1), to generate corresponding malicious samples from the benign samples.

For the IRISH dataset, we consider only $\mathcal{C}'_{irish}$ = 200 consumers from the total residential consumers ($\mathcal{C}_{irish}$). As a result, we have a total of 214,400 samples, divided into 2 categories: 107,200 benign samples and 107,200 malicious samples (generated using the continuous attack function f_1). For the PRECON dataset, we used all consumers $\mathcal{C}_{precon}$, applying the continuous attack function f_1 to generate corresponding malicious samples for these consumers. Thus, the data are balanced, meaning that there is an equal number of benign and malicious samples in both datasets. The IRISH dataset was utilized to evaluate both the classification performance and the robustness of our proposed detector. In contrast, the PRECON dataset was employed specifically to assess the classification performance of the detector. This is because the PRECON dataset contains a limited number of consumers, preventing the effective partitioning of the data into separate attacker and defense sides.

5.2. Experimental Setup

We employed various Python 3.11.6 libraries for different tasks, including sklearn [53] for performance assessment, matplotlib [54] for data visualization, keras [55] for detector training, and pandas and numpy for data preparation. To train binary detectors, we merged the benign and malicious data for each consumer. For anomaly detectors, we only consider benign data. We then split the merged data into testing and training sets with a 1:2 ratio. Global detectors are built using the data from all consumers. Additionally, the selected data are divided into K clusters using the K-means algorithm, where $K = 5$ in the IRISH dataset and $K = 1$ in the PRECON dataset due to the limited number of consumers in the latter dataset. This division is used to build cluster-based detectors utilizing data from consumers within the same cluster.

To ensure a fair comparison among all detectors, we tuned their hyper-parameters and utilized the optimal values. For OCSVM, the optimal values are as follows: kernel = ‘rbf’, gamma = ‘auto’, and nu = ‘0.03’. Table 1 summarize the optimal values for a DAE detector and the attacker’s detectors (FFNN and CNN). The description values in these tables correspond to the layer type, number of neurons, and activation function. In terms of the hyper-parameters for adversarial evasion attacks, we explored a different range for each parameter and selected values that effectively bypass the ET detector while maximizing theft profit. For FGSM and BIM attacks, the optimal values are as follows: α is a portion of optimal ϵ for $I = 100$ and $0.01 < \epsilon < 0.5$. For CW, ZOO, and DeepFool attacks, the optimal values are as follows: maximum iterations = 100, decrease factor = 0.6, learning rate = 0.01, and maximum perturbation = 0.5. Our experiments were conducted using the high-performance cluster of Tennessee Technological University.

Table 1. The optimal hyper-parameters for the DAE defense detector and the FFNN and CNN attacker detectors.

DAE	FFNN	CNN
Input, 48, Linear	Input, 48, Linear	Input, 48, Linear
Dense, 200, Tanh/ReLU	Dense, 96, Linear	Conv1D, 128, ReLU
Dense, 100, Tanh/ReLU	Dense, 192, ReLU	Dense, 256, ReLU
Dense, 50, Tanh/ReLU	Dense, 387, ReLU	Dense, 128, ReLU
Dense, 32, Tanh/ReLU	Dense, 768, ReLU	Dense, 64, Sigmoid
Dense, 50, Tanh/ReLU	Dense, 192, ReLU	Output, 2, Softmax
Dense, 100, Tanh/ReLU	Dense, 200, ReLU	
Dense, 200, Tanh/ReLU	Output, 2, Softmax	
Output, 48, Linear		

5.3. Evaluation Metrics

We have used the following metrics to evaluate our detector and assess the vulnerability to adversarial evasion attacks.

– Accuracy (ACC) represents the percentage of the test samples accurately classified by the detector to the total number of samples in the test dataset. It is calculated as follows:

$$ACC(\%) = \frac{TP + TN}{TP + TN + FP + FN} \times 100. \tag{7}$$

– False Alarm (FA), known also as the false positive rate (FPR), represents the percentage of the false positive samples out of the total number of negative samples. It is calculated as follows:

$$FA(\%) = \frac{FP}{FP + TN} \times 100. \tag{8}$$

– Detection Rate (DR), known also as true positive rate (TPR) and recall, represents the percentage of the true positive samples out of the total number of positive samples. It is calculated as follows:

$$DR(\%) = \frac{TP}{TP + FN} \times 100. \tag{9}$$

– Precision (PR) represents the percentage of true positive samples out of the total number of samples identified by the detector as positive. It is calculated as follows:

$$PR(\%) = \frac{TP}{TP + FP} \times 100. \tag{10}$$

– F1-score ($F1$) represents a statistical measure of both the precision and the detection rate. It is calculated as follows:

$$F1(\%) = \frac{2 * PR * DR}{PR + DR} \times 100. \tag{11}$$

– Highest difference (HD)represents the difference between the detection rate and the false alarm. It is calculated as follows:

$$HD(\%) = DR(\%) - FA(\%). \tag{12}$$

Equations (7)–(12) are derived from the values within the confusion matrix (TP, TN, FN, and FP). TP represents the count of correctly classified malicious samples, known as true positive. TN is the count of correctly classified benign samples, known as true negative. FN is the count of malicious samples incorrectly classified as benign, known as false negative. FP is the count of benign samples incorrectly classified as malicious, known as false positive.

5.4. Experimental Results

We have conducted three experiments. First, we compare the classification performance of a state-of-the-art DAE-based and a DAE+OCSVM ET anomaly detector. Second, we assess the vulnerability of the DAE+OCSVM ET anomaly detector to gradient-based adversarial evasion attacks using transferability. Third, we evaluate the proposed XAI-based DAE+OCSVM ET anomaly detector in terms of classification performance and its robustness against both gradient-based and optimization-based adversarial evasion attacks under a black-box threat model.

5.4.1. Experiment 1

In this experiment, since shallow (or traditional) anomaly detectors (e.g., OCSVM, Isolation Forest) generally achieve lower classification performance compared to deep anomaly detectors, we focused on deep anomaly detection. Therefore, we first trained global deep anomaly detectors using DAE and DAE+OCSVM. Then, we evaluated their performance using ACC, DR, FA, PR, $F1$, and HD. Table 2 compares the performance of three global anomaly detectors: DAE (With-malicious), where the threshold is determined using a malicious electricity consumption dataset; DAE (No-malicious), where the threshold is calculated without using a malicious electricity consumption dataset; and our proposed detector DAE+OCSVM (No-malicious), which does not require a threshold because OCSVM can automatically determine it. From all evaluation metrics in this table, it is evident that the performance of DAE (No-malicious) is limited due to the absence of a malicious electricity consumption dataset, which leads to incorrect threshold selection. Conversely, DAE (With-malicious) achieves good performance under the impractical assumption of ideal threshold selection based on the availability of a real malicious electricity consumption dataset, which is not available. Moreover, DAE+OCSVM (No-malicious) achieves superior performance (e.g., ACC is 93.61% and DR is 96.74%) without the need for the calculation of a threshold or knowledge on malicious electricity consumption data because OCSVM can determine the threshold automatically during the training phase. Therefore, the existing DAE-based methods are only effective when there are known malicious samples, as shown in the DAE (With-malicious). However, in more practical situations where malicious samples are unknown, these methods perform poorly, as shown in the DAE (No-malicious). In contrast, our proposed detector DAE+OCSVM (No-malicious) achieves high performance without needing malicious samples.

Table 2. The performance comparison between the existing ET anomaly detectors and the proposed ET anomaly detector (DAE+OCSVM).

Model Type	ACC	DR	FA	PR	F1	HD
DAE [15,24] (With-malicious)	91.63	92.92	9.67	90.58	91.74	83.25
DAE [15,24] (No-malicious)	51.32	3.33	0.67	83.29	6.41	2.66
DAE+OCSVM (No-malicious)	93.61	96.74	9.51	91.05	93.81	87.23

Note: Malicious samples are used during the threshold selection process.

5.4.2. Experiment 2

In this experiment, we evaluate the robustness of the DAE+OCSVM anomaly detector against benchmark adversarial evasion attacks (FGSM and BIM) using different attacker's model architectures (CNN and FFNN). Table 3 presents the experimental results. The results, including ACC, DR, PR, $F1$, and HD values, demonstrate the vulnerability of the DAE+OCSVM anomaly detector to evasion attacks. Taking the CNN-based attacker model as an example, we observe that the ACC, DR, PR, $F1$, and HD values decrease from

93.22%, 97.48%, 89.82%, 93.49%, and 86.43%, respectively, under no evasion to 70.21%, 51.47%, 82.33%, 63.34%, and 40.38%, respectively, under the FGSM attack and to 67.80%, 46.65%, 80.85%, 59.16%, and 35.56%, respectively, under the BIM attack. Additionally, the *FA* metric remain unchanged and align with the values in Table 2 because the attacker generates adversarial evasion samples only from malicious ones. These results demonstrate the transferability of adversarial evasion attacks across different ML models, even when the attacker's surrogate model has a different architecture from the defense model. *The performance reduction after launching the adversarial evasion attacks proves the severity of these attacks*. This emphasizes the need to secure the DAE+OCSVM ET anomaly detector against adversarial evasion attacks while maintaining high accuracy, as is discussed in Experiment 3.

Table 3. The vulnerability of the DAE+OCSVM anomaly detector to evasion attacks.

		ACC	DR	PR	F1	HD
No Evasion		93.22	97.48	89.82	93.49	86.43
CNN-based Attacker	FGSM	70.21	51.47	82.33	63.34	40.38
	BIM	67.80	46.65	80.85	59.16	35.56
FFNN-based Attacker	FGSM	71.47	53.99	83.01	65.43	42.09
	BIM	68.46	47.96	81.28	60.32	38.87

5.4.3. Experiment 3

In this experiment, we first compare DAE+OCSVM with and without the proposed defense (XAI and clustering) in terms of classification performance, as shown in Table 4. Next, we evaluate the robustness of DAE+OCSVM with and without the proposed defense (XAI and clustering) against gradient-based (FGSM and BIM) and optimization-based evasion attacks (CW, ZOO, and DeepFool). The results of these robustness evaluations are provided in Figures 7 and 8 in terms of *DR*. *It is worth noting that the difference between the DR before (i.e., with no evasion) and after attacks represents the severity of the attacks (i.e., attack success rate)*. The detailed findings from this experiment are as follows:

Table 4. Comparison of DAE+OCSVM with and without the proposed defense (XAI and clustering) in terms of classification performance.

Dataset		ACC	DR	FA	PR	F1	HD
IRISH	No Defense	93.61	96.74	9.51	91.05	93.81	87.23
	Proposed (C1-XAI)	97.75	100.00	4.49	95.70	97.80	95.51
	Proposed (C2-XAI)	97.28	100.00	5.45	94.83	97.35	95.55
	Proposed (C3-XAI)	97.60	100.00	4.80	95.42	97.65	95.20
	Proposed (C4-XAI)	97.79	100.00	4.41	95.78	97.84	95.59
	Proposed (C5-XAI)	97.75	100.00	4.49	95.70	97.80	95.51
PRECON	No Defense	85.31	86.41	15.75	83.99	85.18	70.66
	Proposed (C-XAI)	92.69	96.88	11.34	89.13	92.84	85.54

- Table 4 compares the classification performance (*ACC*, *DR*, *FA*, *PR*, *F*1, and *HD*) of DAE+OCSVM with and without the proposed defense. It is evident from this table that the proposed defense improves all the classification performance metrics, which indicates an ability to accurately distinguish between benign and malicious samples. This improvement occurs because the use of clustering allows detectors to train on data with close consumption patterns, which leads to parameters closer to the optimal detector's parameters for individual consumption patterns (i.e., lower level of generalization). Additionally, the use of XAI distinctly separates benign and malicious consumption patterns, leading to easier ET detection. Additionally, Figure 9 shows the Precision–Recall (PR) and Receiver Operating Characteristic (ROC) *curves*. These figures provide a visual representation of the performance comparison. They indicate that our proposed detectors achieve higher performance, as evidenced by the values of AUC_PR and AUC_ROC, demonstrating that the proposed detectors enhance ET detection in both the IRISH and PRECON datasets;

- Figure 7 compares the robustness of DAE+OCSVM with and without the proposed defense against CNN-based attacker model in terms of DR. The difference between the DR before (i.e., with no evasion) and after attacks represents the severity of the attacks or attack success rate (ASR). In No Defense, the DR shows a minimal decrease from 97.48% to 46.65% under BIM attack. However, with the proposed defense, the DR values remain above 90% for all attacks across clusters C1-XAI to C5-XAI. Figure 8 compares the robustness of DAE+OCSVM with and without the proposed defense against the FFNN-based attacker model in terms of DR. Here, the difference between the DR before and after the attacks also reflects the severity of the attacks ASR. In No Defense, the DR shows a minimal decrease from 97.48% to 47.96% under BIM attack. However, with the proposed defense, the DR values remain above 90% for all attacks across clusters C1-XAI to C5-XAI. It is evident from those DR values that the proposed anomaly detector achieves a promising level of robustness. This is attributed to the use of clustering, which results in a lower level of generalization, and XAI, which facilitates the separation of benign and anomalous consumption patterns, as shown in Figure 5 and Figure 6, respectively. Moreover, the deep structure of the proposed detector extracts relevant features from the SHAP explanations of consumption readings, thereby facilitating the detection of anomalous XAI explanations caused by these adversarial evasion attacks.

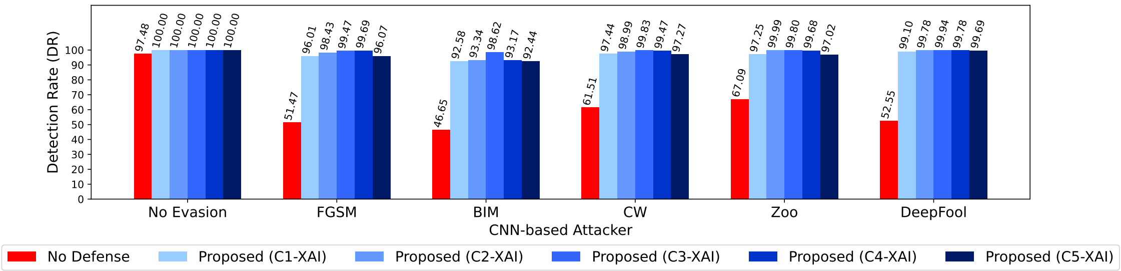


Figure 7. Comparison of the robustness of DAE+OCSVM with and without the proposed defense against CNN-based evasion attacks in terms of DR.

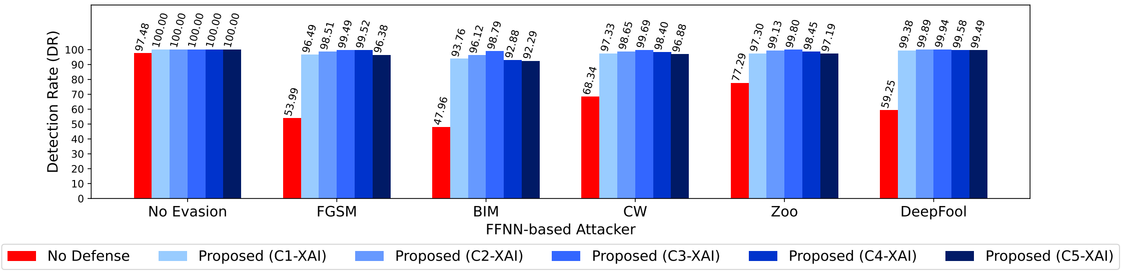
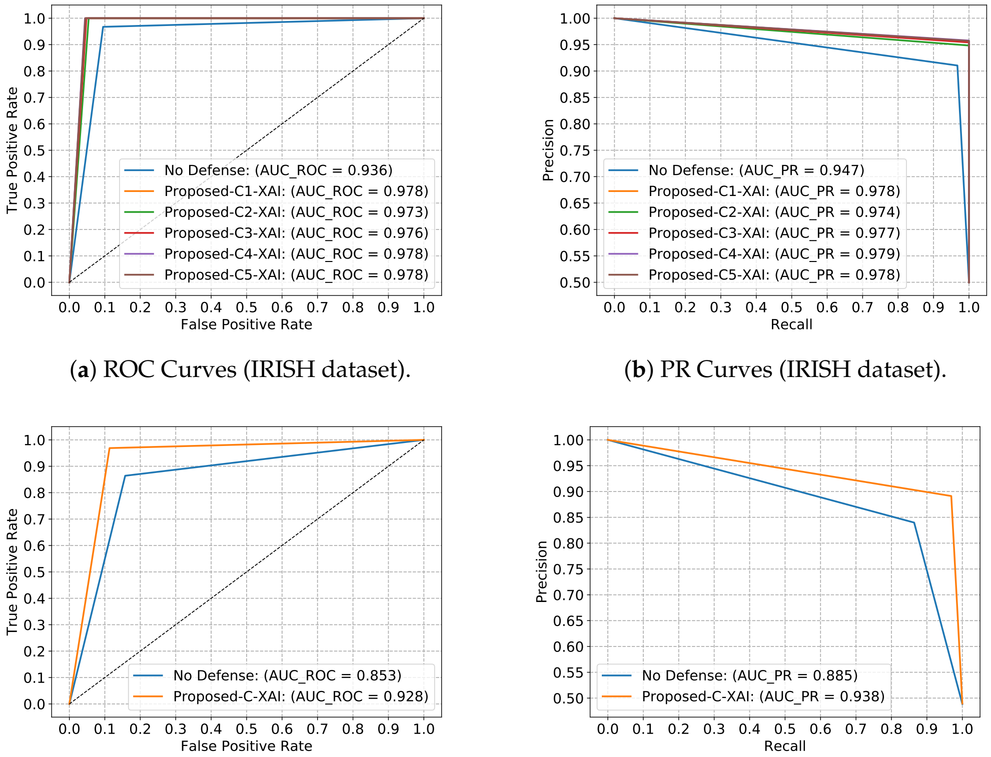


Figure 8. Comparison of the robustness of DAE+OCSVM with and without the proposed defense against FFNN-based evasion attacks in terms of DR.



(**a**) ROC Curves (IRISH dataset).

(**b**) PR Curves (IRISH dataset).

(**c**) ROC Curves (PRECON dataset).

(**d**) PR Curves (PRECON dataset).

Figure 9. PR and ROC curves of DAE+OCSVM with and without the proposed defense (XAI and clustering).

6. Discussion and Conclusions

In this paper, we have investigated the effectiveness of using XAI in ET anomaly detection. First, we proposed a hybrid DAE+OCSVM anomaly detector that not only achieves superior detection performance but also overcomes the threshold sensitivity of DAE anomaly detection. Second, we showed that the DAE+OCSVM anomaly detector is vulnerable to benchmark evasion attacks. Third, we proposed a defense strategy to bolster this ET anomaly detector using XAI and clustering. The utilization of XAI (i.e., the proposed detector trained on the SHAP explanations of consumption readings) facilitates the distinct separation of benign and anomalous consumption patterns, leading to much easier and more robust ET detection. Moreover, the utilization of clustering enhances robustness by reducing the level of generalization. The experimental results illustrate that our detector demonstrates significant robustness against various evasion attacks, including gradient-based and optimization-based. With no defense, the *DR* is reduced from 97.48% to up to 46.65% under evasion attacks. However, with the proposed defense, the minimum *DR* values are 92.58%, 93.34%, 98.62%, 92.88%, and 92.29% for clusters C1-XAI to C5-XAI, respectively. These *DR* values reflects the significant detection capabilities and robustness of the proposed detector. As future work, we aim to explore different explanation (XAI) methods to enhance trust and reliability, as well as to assess the performance of our proposed detector under a white-box threat model. While the black-box threat model is considered the most realistic, it is important to evaluate the detector's effectiveness as the attacker's level of knowledge increases.

Author Contributions: Conceptualization, I.E., M.M.B., M.M., M.A. (Maazen Alsabaan), M.A. (Muteb Alsaqhan) and T.A.; Methodology, I.E., M.M.B., M.M., M.A. (Maazen Alsabaan), M.A. (Muteb Alsaqhan) and T.A.; Software, I.E. and M.A. (Muteb Alsaqhan); Validation, I.E., M.M.B. and M.A. (Maazen Alsabaan); Investigation, I.E., M.M.B., M.M. and M.A. (Muteb Alsaqhan); Writing—original draft, I.E., M.M.B. and M.M.; Writing—review & editing, I.E., M.M.B. and M.M.; Supervision, M.M. and M.A. (Maazen Alsabaan). All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Researchers Supporting Project number (RSPD2024R636), King Saud University, Riyadh, Saudi Arabia and NSF grant #2349104.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Erol-Kantarci, M.; Mouftah, H.T. Smart grid forensic science: Applications, challenges, and open issues. *IEEE Commun. Mag.* **2013**, *51*, 68–74. [CrossRef]
2. Gunduz, M.Z.; Das, R. Smart Grid Security: An Effective Hybrid CNN-Based Approach for Detecting Energy Theft Using Consumption Patterns. *Sensors* **2024**, *24*, 1148. [CrossRef] [PubMed]
3. Hashim, M.; Khan, L.; Javaid, N.; Ullah, Z.; Shaheen, I. Enhancing Smart City Functions through the Mitigation of Electricity Theft in Smart Grids: A Stacked Ensemble Method. *Int. Trans. Electr. Energy Syst.* **2024**, *2024*, 5566402. [CrossRef]
4. Qi, R.; Zheng, J.; Luo, Z.; Li, Q. A novel unsupervised data-driven method for electricity theft detection in AMI using observer meters. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–10. [CrossRef]
5. Buzau, M.M.; Tejedor-Aguilera, J.; Cruz-Romero, P.; Gomez-Exposito, A. Hybrid deep neural networks for detection of non-technical losses in electricity smart meters. *IEEE Trans. Power Syst.* **2019**, *35*, 1254–1263. [CrossRef]
6. Takiddin, A.; Ismail, M.; Zafar, U.; Serpedin, E. Deep Autoencoder-based Detection of Electricity Stealth Cyberattacks in AMI Networks. In Proceedings of the 2021 International Symposium on Signals, Circuits and Systems (ISSCS), Iasi, Romania, 15–16 July 2021; pp. 1–6. [CrossRef]
7. Lepolesa, L.J.; Achari, S.; Cheng, L. Electricity Theft Detection in Smart Grids Based on Deep Neural Network. *IEEE Access* **2022**, *10*, 39638–39655. [CrossRef]
8. McDaniel, P.; McLaughlin, S. Security and privacy challenges in the smart grid. *IEEE Secur. Priv.* **2009**, *7*, 75–77. [CrossRef]
9. Liao, W.; Takiddin, A.; Tariq, M.; Chen, S.; Ge, L.; Yang, Z. Sample adaptive transfer for electricity theft detection with distribution shifts. *IEEE Trans. Power Syst.* **2024**, *39*, 7012–7024. [CrossRef]
10. Emadaleslami, M.; Haghifam, M.R.; Zangiabadi, M. A two stage approach to electricity theft detection in AMI using deep learning. *Int. J. Electr. Power Energy Syst.* **2023**, *150*, 109088. [CrossRef]
11. Yao, R.; Wang, N.; Ke, W.; Chen, P.; Sheng, X. Electricity theft detection in unbalanced sample distribution: A novel approach including a mechanism of sample augmentation. *Appl. Intell.* **2023**, *53*, 11162–11181. [CrossRef]
12. Jindal, A.; Dua, A.; Kaur, K.; Singh, M.; Kumar, N.; Mishra, S. Decision tree and SVM-based data analytics for theft detection in smart grid. *IEEE Trans. Ind. Inform.* **2016**, *12*, 1005–1016. [CrossRef]
13. El-Toukhy, A.T.; Elgarhy, I.; Badr, M.M.; Mahmoud, M.; Fouda, M.M.; Ibrahem, M.I.; Amsaad, F. Securing Smart Grids: Deep Reinforcement Learning Approach for Detecting Cyber-Attacks. In Proceedings of the 2024 International Conference on Smart Applications, Communications and Networking (SmartNets), Harrisonburg, VA, USA, 28–30 May 2024; pp. 1–6. [CrossRef]
14. Jokar, P.; Arianpoo, N.; Leung, V.C. Electricity theft detection in AMI using customers' consumption patterns. *IEEE Trans. Smart Grid* **2015**, *7*, 216–226. [CrossRef]
15. Takiddin, A.; Ismail, M.; Zafar, U.; Serpedin, E. Deep autoencoder-based anomaly detection of electricity theft cyberattacks in smart grids. *IEEE Syst. J.* **2022**, *16*, 4106–4117. [CrossRef]
16. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. In Proceedings of the 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, 14–16 April 2014.
17. Papernot, N.; McDaniel, P.; Jha, S.; Fredrikson, M.; Celik, Z.B.; Swami, A. The limitations of deep learning in adversarial settings. In Proceedings of the IEEE European symposium on security and privacy (EuroS&P), Saarbruecken, Germany, 21–24 March 2016; pp. 372–387.
18. Yuan, X.; He, P.; Zhu, Q.; Li, X. Adversarial examples: Attacks and defenses for deep learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 2805–2824. [CrossRef]
19. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. *arXiv* **2014**, arXiv:1412.6572.
20. Papernot, N.; McDaniel, P.; Wu, X.; Jha, S.; Swami, A. Distillation as a defense to adversarial perturbations against deep neural networks. In Proceedings of the IEEE symposium on security and privacy (SP), San Jose, CA, USA, 22–26 May 2016; pp. 582–597.
21. Wei, W.; Liu, L. Robust deep learning ensemble against deception. *IEEE Trans. Dependable Secur. Comput.* **2020**, *18*, 1513–1527. [CrossRef]

22. Goodge, A.; Hooi, B.; Ng, S.K.; Ng, W.S. Robustness of autoencoders for anomaly detection under adversarial impact. In Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, Yokohama, Japan, 7–15 January 2021; pp. 1244–1250.
23. Lo, S.Y.; Oza, P.; Patel, V.M. Adversarially Robust One-Class Novelty Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 4167–4179. [CrossRef]
24. Takiddin, A.; Ismail, M.; Serpedin, E. Robust Data-Driven Detection of Electricity Theft Adversarial Evasion Attacks in Smart Grids. *IEEE Trans. Smart Grid* **2023**, *14*, 663–676. [CrossRef]
25. Ko, G.; Lim, G. Unsupervised detection of adversarial examples with model explanations. *arXiv* **2021**, arXiv:2107.10480.
26. Fidel, G.; Bitton, R.; Shabtai, A. When Explainability Meets Adversarial Learning: Detecting Adversarial Examples using SHAP Signatures. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–8. [CrossRef]
27. Watson, M.; Al Moubayed, N. Attack-agnostic Adversarial Detection on Medical Data Using Explainable Machine Learning. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 8180–8187. [CrossRef]
28. AL-Essa, M.; Andresini, G.; Appice, A.; Malerba, D. An XAI-based adversarial training approach for cyber-threat detection. In Proceedings of the 2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), Falerna, Italy, 12–15 September 2022; pp. 1–8. [CrossRef]
29. Lin, Y.C.; Yu, F. DeepSHAP Summary for Adversarial Example Detection. In Proceedings of the 2023 IEEE/ACM International Workshop on Deep Learning for Testing and Testing for Deep Learning (DeepTest), Melbourne, Australia, 15 May 2023; pp. 17–24. [CrossRef]
30. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4765–4774.
31. Li, J.; Yang, Y.; Sun, J.S. SearchFromFree: Adversarial measurements for machine learning-based energy theft detection. In Proceedings of the IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), Tempe, AZ, USA, 11–13 November 2020; pp. 1–6.
32. Badr, M.M.; Mahmoud, M.; Abdulaal, M.; Aljohani, A.J.; Alsolami, F.; Balamsh, A. A Novel Evasion Attack Against Global Electricity Theft Detectors and a Countermeasure. *IEEE Internet Things J.* **2023**, *10*, 11038–11053. [CrossRef]
33. Elgarhy, I.; Badr, M.M.; Mahmoud, M.; Fouda, M.M.; Alsabaan, M.; Kholidy, H.A. Clustering and Ensemble Based Approach For Securing Electricity Theft Detectors Against Evasion Attacks. *IEEE Access* **2023**, *11*, 112147–112164. [CrossRef]
34. Elgarhy, I.; El-Toukhy, A.T.; Badr, M.M.; Mahmoud, M.; Fouda, M.M.; Alsabaan, M.; Kholidy, H.A. Secured Cluster-Based Electricity Theft Detectors Against Blackbox Evasion Attacks. In Proceedings of the 2024 IEEE 21st Consumer Communications & Networking Conference (CCNC), Las Vegas, NV, USA, 6–9 January 2024; pp. 333–338. [CrossRef]
35. Elgarhy, I.; Badr, M.M.; Mahmoud, M.; Nabil, M.; Alsabaan, M.; Ibrahem, M.I. Securing Smart Grid False Data Detectors Against White-box Evasion Attacks Without Sacrificing Accuracy. *IEEE Internet Things J.* **2024**, *11*, 33873–33889. [CrossRef]
36. Takiddin, A.; Ismail, M.; Zafar, U.; Serpedin, E. Robust Electricity Theft Detection Against Data Poisoning Attacks in Smart Grids. *IEEE Trans. Smart Grid* **2021**, *12*, 2675–2684. [CrossRef]
37. Amich, A.; Eshete, B. EG-Booster: Explanation-Guided Booster of ML Evasion Attacks. In Proceedings of the Twelfth ACM Conference on Data and Application Security and Privacy, New York, NY, USA, 25–27 April 2022; CODASPY '22, pp. 16–28. [CrossRef]
38. Zhang, X.; Wang, N.; Shen, H.; Ji, S.; Luo, X.; Wang, T. Interpretable deep learning under fire. In Proceedings of the 29th {USENIX} Security Symposium ({USENIX} Security 20), Online, 12–14 August 2020.
39. Kurakin, A.; Goodfellow, I.J.; Bengio, S. Adversarial examples in the physical world. In *Artificial Intelligence Safety and Security*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2018; pp. 99–112.
40. Carlini, N.; Wagner, D. Towards evaluating the robustness of neural networks. In Proceedings of the IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22–26 May 2017; pp. 39–57.
41. Chen, P.Y.; Zhang, H.; Sharma, Y.; Yi, J.; Hsieh, C.J. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, Dallas, TX, USA, 3 November 2017; pp. 15–26.
42. Moosavi-Dezfooli, S.M.; Fawzi, A.; Frossard, P. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, Nevada, USA, 27–30 June 2016.
43. Papernot, N.; McDaniel, P.; Goodfellow, I. Transferability in machine learning: From phenomena to black-box attacks using adversarial samples. *arXiv* **2016**, arXiv:1605.07277.
44. Tanay, T.; Griffin, L. A boundary tilting persepective on the phenomenon of adversarial examples. *arXiv* **2016**, arXiv:1608.07690.
45. Deniz, O.; Pedraza, A.; Vallez, N.; Salido, J.; Bueno, G. Robustness to adversarial examples can be improved with overfitting. *Int. J. Mach. Learn. Cybern.* **2020**, *11*, 935–944. [CrossRef]
46. Principi, E.; Rossetti, D.; Squartini, S.; Piazza, F. Unsupervised electric motor fault detection by using deep autoencoders. *IEEE/CAA J. Autom. Sin.* **2019**, *6*, 441–451. [CrossRef]

47. Kim, C.; Chang, S.Y.; Kim, J.; Lee, D.; Kim, J. Automated, reliable zero-day malware detection based on autoencoding architecture. *IEEE Trans. Netw. Serv. Manag.* **2023**, *20*, 3900–3914. [CrossRef]
48. Commission for Energy Regulation (CER). *CER Smart Metering Project—Electricity Customer Behaviour Trial, 2009-2010 [Dataset]*, 1st ed.; SN: 0012-00; Irish Social Science Data Archive: Dublin, Ireland, 2012. Available online: https://www.ucd.ie/issda/data/commissionforenergyregulationcer/ (accessed on 25 October 2024).
49. Nadeem, A.; Arshad, N. PRECON: Pakistan Residential Electricity Consumption Dataset. In Proceedings of the Tenth ACM International Conference on Future Energy Systems, e-Energy '19, New York, NY, USA, 25–28 June 2019; pp. 52–57. [CrossRef]
50. Breunig, M.M.; Kriegel, H.P.; Ng, R.T.; Sander, J. LOF: Identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas, TX, USA, 16–18 May 2000; pp. 93–104.
51. Yan, Z.; Wen, H. Electricity Theft Detection Base on Extreme Gradient Boosting in AMI. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 2504909. [CrossRef]
52. Zheng, Z.; Yang, Y.; Niu, X.; Dai, H.N.; Zhou, Y. Wide and Deep Convolutional Neural Networks for Electricity-Theft Detection to Secure Smart Grids. *IEEE Trans. Ind. Inform.* **2018**, *14*, 1606–1615. [CrossRef]
53. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
54. Hunter, J.D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95. [CrossRef]
55. The Functional API. Available online: https://keras.io/guides/functional_api/ (accessed on 25 October 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.