Penetrating the Power Grid: Realistic Adversarial Attacks on Smart Grid Intrusion Detection Systems

Nelson Makau Mutua[1(✉)], Simin Nadjm-Tehrani[2], and Petr Matoušek[1]

[1] Brno University of Technology, Bozetechova 1, 612 66 Brno, Czech Republic
{imutua,matousp}@fit.vutbr.cz
[2] Linköping University, 581 83 Linköping, Sweden
simin.nadjm-tehrani@liu.se

Abstract. The widespread adoption and application of Machine Learning (ML) based Intrusion Detection Systems (IDS) has increased the flexibility and efficiency of automated cyber attack detection in smart grid systems. However, the emergence of such IDSes has led to a new attack vector against learning models, known as adversarial attacks. Such attacks could have serious effects in smart grid systems since adversaries can circumvent detection by IDS. This could result in detection of attacks. From the existing literature, a lot of research proposes threat models that are inappropriate for generating realistic adversarial attacks. In this research, we model realistic adversarial attacks with a focus on real attacker capabilities that are feasible to launch adversarial attacks. We discuss how adversarial learning may be used to target ML models using the Jacobian-based Saliency Map Attack (JSMA) and the Fast Gradient Sign Method (FGSM). A power system dataset generated from a smart grid testbed was used for testing the models. The performance of the trained classifiers, Random Forest, XGBoost, and Naive Bayes, dropped when adversarial instances were introduced. The outcomes of this paper are useful for helping researchers model realistic scenarios to avoid dealing with hypothetical problems.

Keywords: Intrusion Detection Systems · Adversarial Attacks · Critical Infrastructure · Machine Learning · Smart Grid Systems

1 Introduction

Smart electrical grids are essential in the digital age of hyper-connected Critical Infrastructures (CIs), providing benefits including improved grid resilience, efficient energy distribution, and smart load management [19]. The adoption of technology enablers such as machine learning (ML), the Internet of Things (IoT), 5G, and Artificial Intelligence (AI) is critical to the life cycle of smart grids. However, this technological breakthrough presents serious cybersecurity concerns, which might have fatal implications, particularly in the energy sector.

G. Oliva et al. (Eds.): CRITIS 2024, LNCS 15549, pp. 249–268, 2025.
https://doi.org/10.1007/978-3-031-84260-3_15

Given the significance of these systems, they have become a desirable target for attackers. By the fact that these systems control physical processes, cyber-attacks may have far-reaching effects on the environment in which they operate and their users [4].

Advanced Persistent Threats (APTs) and multi-step attacks against CIs, like the smart electrical grid, can cause service failures, financial losses, and sometimes tragic accidents. Examples of APT [4] campaigns include Hafnium, Industroyer, the Lazarus Group, and SolarWinds (Sunburst). In 2015, Industroyer lead to a blackout across Ukraine. The NotPetya ransomware caused significant financial damage for various energy-related organizations, making it a notable cybersecurity incident. A more recent CI attack was reported in Denmark in May 2023 where attackers compromised 22 energy organizations in the largest coordinated attack against Denmark's CI [32]. To launch the attacks, hackers exploited multiple vulnerabilities in the firewall for initial access, executing code and gaining complete control over the impacted systems. The attackers successfully compromised 11 energy organizations by executing commands on the vulnerable firewall to obtain device configurations and usernames and thus access to the CI behind it. In this light, security issues about such systems have become a serious concern globally. This prompts the development of not only a safe but also a robust technique that can effectively identify and protect CIs like smart grid networks from cyber attacks.

Although various security methods exist for traditional IT systems, integrating them into smart grid networks is difficult because the monitoring devices have limited resources and the inability to support contemporary security measures. Improvements to security are thus more likely to be brought about by passive security surveillance and other such security approaches. This has resulted in a significant rise in research into more tailor-made IDSes that monitor networks to detect attacks that could disrupt the operation of CIs [20]. IDSes are increasingly being integrated with ML due to their efficiency in attack detection attacks. However, with the introduction of these systems, a new attack vector emerges, such that even the trained models can be attacked. Adversarial Machine Learning (AdvML) refers to deploying attacks against ML systems. Small perturbations can be applied automatically to unseen data points that can result in the model crossing a decision boundary and then classify malicious data as normal.

The existence of such dynamics implies that CI, such as smart grid systems that use ML-IDSes, may be exposed to cyberattacks. AdvML can be used to manipulate data from the Intelligent Electronic Devices (IEDs) that switch circuit breakers or other devices. IEDs introduce perturbations that classify malicious data as benign, hence circumventing the IDS. As a result, there could be delays in detecting attacks, leaks of information, financial losses, or sometimes casualties. As ML-based detection methods grow more prevalent, attackers may have a stronger motive to target them. As a result, they require extensive evaluation against AdvML attacks.

1.1 Motivation and Contribution

Our research is motivated by the recognition that many research papers design, develop, and evaluate IDS in adversarial settings without considering the realism of the proposed attacks or explaining how they can be launched in reality. Many of the proposed research works assume a threat model and report the impact of the attack while providing little or no thought to the viability of the considered perturbation. Moreover, some general techniques are applied to generate adversarial attacks to manipulate the network features in a way that is inconsistent with actual network traffic [3].

Some researchers, for instance, take the view that adversaries are fully aware of the target system [16] while others assume an attacker can attempt as often as possible to breach past the Network Intrusion Detection Systems (NIDS) without getting detected [30]. When generating scenarios in cybersecurity, it is critical to concentrate on the actual problem and adversary. However, it is equally important to characterize how effective all adversarial attacks are on any ML to develop a better detector. Failing to do so could misinform defenders to allocate resources against false cases or hypothetical problems, potentially diverting attention from more critical issues. The abundance of research on adversarial attacks might inadvertently give the impression that any ML-IDS is an unreliable defensive system, contrary to the actual scenario.

Additionally, in a real communication network, manipulating an ML model does not ensure a successful cyberattack. In this study, we propose a more realistic approach to modelling adversarial attacks against ML-IDS for smart grid communication. We identify the necessary conditions and capabilities for the attacker to carry out such attacks. More importantly, this study recreates a realistic dataset gathered from a power system testbed, along with a realistic attack model and assumptions. This research will make the following contributions:

- Detailed review of the feasibility constraints needed to model valid adversarial perturbations on data used as input to an ML-IDS while preserving the network attack's fundamental logic.
- Generating evasion attacks for smart grid network communication capable of evading ML-IDS detection with limited knowledge of the target NIDS.
- Demonstrate the effectiveness of the evasion attack on ML-IDS.

2 Background and Related Work

This section introduces the fundamental ideas of adversarial machine learning. Then, we discuss related work that has utilized adversarial evasion techniques to illustrate how effective they are in evading or reducing the performance of IDS models. Finally, we outline the limitations of the network traffic and detail how the limitations can be upheld for the network to generate an adversarial flow that is valid.

2.1 Adversarial Machine Learning

An adversarial attack refers to the application of small and undetectable alterations to an ML detector [29]. In this case, the altered samples must not have a huge difference from the initial sample to maintain the basic malicious logic and avoid triggering other detection methods. In this research, we focus on evasion attacks. In the event of an evasion attack, the input is manipulated by the attacker in an attempt to trick the model and cause misclassification. There are several approaches to generating adversarial samples, and they vary in terms of their performance, generation speed, and complexity. The easiest way to create adversarial samples is by manually changing the input data points one by one. In such cases, it can take a lot of time to perturb manually, especially with large datasets, and the results could be inaccurate. Therefore, more complicated methods can be used to identify and analyze features automatically by discriminating target values.

Papernot et al. [28] and Goodfellow et al. [13] introduced Jacobian-based Saliency Map Attack (JSMA) and Fast Gradient Sign Method (FGSM) as popular methods for creating perturbed samples automatically. Both methods presume that adding small perturbations (δ) to the original sample *(X)* and can result in adversarial characteristics ($X^* = X + \delta$). This implies that X^* will be misclassified by the target model.

2.2 Fast Gradient Sign Method (FGSM)

The FGSM method for creating adversarial instances is based on the gradient sign method with backpropagation. It is an untargeted attack approach used to obtain max-norm constrained perturbation (η) expressed in Eq. 1. Here (θ) represents the model parameter, x is the input vector to the model, y is the associated label of the input, and J(θ, x, y) is the cost function. FGSM generates perturbation samples with a small noise parameter ϵ [13].

$$adv\_x = x + \epsilon * \text{sign}\left(\nabla_x J(\theta, x, y)\right) \tag{1}$$

2.3 Jacobian-Based Saliency Map Attack (JSMA)

The Jacobian matrix, on the other hand, serves as the foundation for the JSMA technique. The JSMA method is utilized to compute the forward derivative of the cost function *f(x)*. The following formula computes the Jacobian of the neural network function F using input X:

$$JF = \frac{\partial F(X)}{\partial X} \tag{2}$$

Unlike the FGSM, JSMA operates differently from other adversarial attacks by leveraging saliency maps. These maps visually represent the prediction process of a classification model for each pixel, illustrating how each pixel influences the model prediction of a specific class. JSMA has various advantages as well as

disadvantages. One advantage of using JSMA is its ability to make small perturbations while maintaining high success rates. These minimal changes make it easy to control attacks based on their intensity within a specific ML-IDS. However, JSMA is more computationally intensive than FGSM [28].

2.4 Major Adversarial Attacks Against the NIDS

This section reviews existing research that used adversarial evasion methods to reduce the performance of ML-IDS models. The existing literature identifies detection methods as vulnerable to generic evasive adversarial attacks, which are considered significant threats. However, the previous research failed to evaluate the effectiveness of generated adversarial traffic for real-world attacks.

In their research, Warzyński and Kołaczek [37] demonstrated that a Deep Neural Networks (DNN) binary classifier on the NSL-KDD[1] dataset [35] was degraded by a FGSM attack. Further, they confirmed that network traffic can also be degraded by the FGSM attack, although this attack was first developed for image recognition. Using the Mirai dataset[2], Clements et al. [9] assessed Kitsune's resistance to FGSM attacks as a lightweight intrusion detection system for IoT networks. Wang [36] discovered that FGSM attacks achieve various degrees of success and use different feature patterns. The author suggested that perturbing specific features may increase the vulnerability of IDS to adversarial traffic. Nevertheless, the research does not show the features had been altered to confirm if the perturbations produced consistent traffic instances.

Asimopoulos et al. [6] introduced an AI-powered IDS for the IEC 60870-5-104 protocol. To test the model, the authors use four ML methods: (a) Decision Tree, (b) RF, (c) eXtreme Gradient Boosting (XGBoost), and (d) Multilayer Perceptron (MLP). The authors used a Conditional Tabular Generative Adversarial Network (CTGAN) adversarial attack generator and the FGSM to assess the possible impact of adversarial attacks on IDS detection performance. In comparison to the CTGAN datasets, the examined models DT, XGBoost, RF, and MLP fared better on the FGSM adversarial datasets. However, the authors did not discuss the realistic implementation of adversarial attacks in their case studies. Additionally, they did not explain how to set the optimum level of perturbations that could trigger an attack.

Huang et al. [17] evaluated the effectiveness of three port-scan attack detection models for Software Defined Networking (SDN) environments: MLP, Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM) against the FGSM attacks. Martins et al. [23] found that FGSM attacks reduced the average performance of RF, SVM, Decision Trees (DT), Naïve Bayes (NB), and Neural Network (NN) classifiers. Sriram et al. [34] evaluated the performance of DNN, RF, Support Vector Machine (SVM), NB, and DT classifiers against FGSM attacks on the NSL-KDD dataset (See footnote [1]). Debicha et al. [10]

[1] See https://www.unb.ca/cic/datasets/nsl.html [May 2024].
[2] See https://ieee-dataport.org/documents/nss-mirai-dataset [May 2024].

concluded that DNN detection model performance was considerably degraded by FGSM attacks.

As indicated by the existing literature, detection models fail to withstand novel adversarial attacks that are considered serious threats. The literature only demonstrates how attacks have high rates of evasion; it fails to show how real and effective the generated adversarial traffic is. In real-world attacks, it is essential to illustrate how generic evasion adversarial attacks can be done realistically. Moreover, the majority of studies have focused on the consequences of adversarial attacks within conventional IP networks [10,11,18,24,31]. Conversely, it is imperative to evaluate security threats in other networking landscapes like smart grids given their critical role of hyper-connected CIs in this digital age.

Based on the previous research, we did not find any research that has verified the realism of adversarial attacks in smart grid networks. Therefore, this paper proposes a realistic approach to modelling adversarial attacks against ML-IDS for smart grid communication by identifying the necessary conditions are required for the attacker to launch such attacks. More importantly, this research develops a reasonable attack model and assumptions and uses real power system datasets.

2.5 Limitations of Previous Research Studies

There were three specific shortcomings of the studies that had been published earlier. First, previous studies failed to observe traffic domain constraints regarding how they created adversarial attacks to maintain the validity and functionality of attack traces. Second, previous studies assumed that the adversary can manipulate or control the number of features without restraint. This would potentially lead to disruption of the semantic connections between interdependent features. Realistically, this assumption may not always hold true in some scenarios because the adversary may be an outsider or may not understand the inner operations of an IDS. Finally, previous studies worked under the conditionality that the threat model is white-box, meaning the adversary has full knowledge of all the parameters of the targeted model, which can be unrealistic in many real-world cases.

3 Case Study

For our case study, we use publicly available power system datasets developed by Mississippi State University and Oak Ridge National Laboratory[3]. Figure 1 depicts the power system framework configuration and components utilized to generate datasets for this research. The power system includes the following components:

– There are two main power generators (G1 and G2).

[3] See https://sites.google.com/a/uah.edu/tommy-morris-uah/ics-data-sets [05/24].

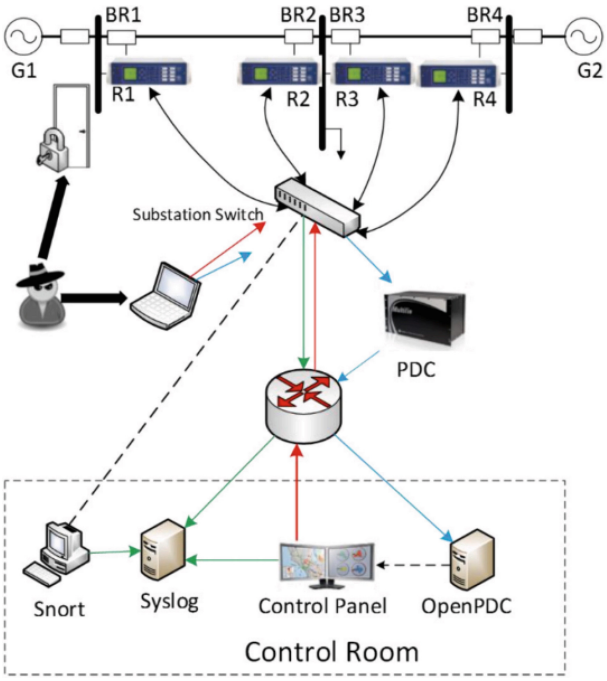

Fig. 1. Power System Testbed [2].

- The Intelligent Electronic Devices (IEDs) R1, R2, R3, and R4 activate the breakers (BR1, BR2, BR3, BR4), which automatically protect electrical circuits from short circuits.
- Each IED controls a single breaker (for example, R1 controls BR1, R2 controls BR2, and so on).
- IEDs lack internal validation; instead, they use a directional relay to trip the breaker in case a fault is detected, regardless of validity.
- System operators can manually trip the breakers by sending commands to the IEDs. To maintain lines as well as system components, operators use manual override.
- The testbed includes additional network monitoring and detection tools, like SNORT and Syslog servers.

3.1 Dataset Description

This dataset contains 128 features from two categories: 1) Phasor Measurement Units (PMU) and 2) logs from the Control Room. With a total of four PMUs, 29 measurements are obtained from each PMU, resulting in 116 features in total. The logs from the control room are categorized into SNORT, control panel, and relay logs, each category with four features. In total, the control role logs have

12 features. More details about the dataset and the description of the features can be found in the original dataset description document [2].

3.2 Simulated Attacks

From the testbed, a data set of malicious and benign data was created. Three main categories are used to classify the data: instances with 'no events', instances with 'natural events', and instances with 'attack events'. 'No event' and 'natural event' are combined to signify benign activity. During the simulation, five different attack scenarios were launched at the power system. The attacks include:

- ***Short-circuit fault.*** This is a power line short circuit that may occur at any distance along the length of the line. The percentage range serves as the location indicator.
- ***Line maintenance.*** To do maintenance on a particular line, one or more relays are disabled.
- ***Remote tripping command injection.*** This attack works by sending a command to a relay that will open a breaker. It is accomplished only after an attacker has penetrated through the outside defence layers.
- ***Alter relay settings.*** In this form of attack, relays have distance protection implemented. In order not to cause the relay to trip at the receipt of a valid command or a fault, the attacker proceeds to change the relay configuration to remove the relay functionality.
- ***Data injection attack.*** Data injection attacks are considered cyberattacks directed at CIs such as power grids. Attackers manipulate system sensors or use other control communication pathways to insert inaccurate data into system control. It can result in erroneous actions that precipitate problems such as instability or blackouts. For instance, an attacker can deliberately fake a power surge or a fault to compel the system to take unnecessary protective actions that interfere with the power supply or cause a blackout.

3.3 Attacker Capabilities

In this research, we model realistic adversarial attacks against Machine Learning NIDS (ML-NIDS) by adopting the taxonomies of Apruzzese et al. [5]. To model them, we evaluate the realistic capabilities of an attacker. This demonstrates how much control the attacker has over the target model. This shows how much control the attacker has over the target detection system. The attacker can have access to the following five elements, as highlighted in Fig. 2.

- ***Training Data*** represents the ability to obtain the dataset required to train the ML-NIDS. There are three types of access: read-only, write-only, and no access.
- ***Feature Set*** indicates the understanding of the features that the IDS employs to continue the detection. There are three types of it: full knowledge, partial, and none.

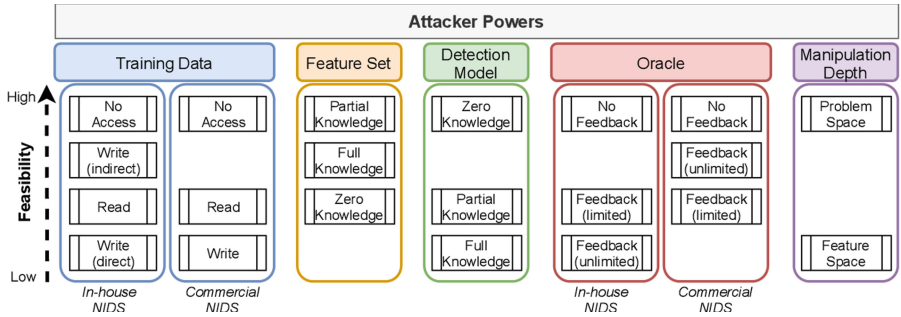

Fig. 2. Power capabilities accessible by the attacker. [5].

- ***Detection Model*** describes a situation whereby the trained machine learning model is employed in the NIDS and used for the detection. This information can be obtained in part, in full, or none at all.
- ***Oracle*** component illustrates how to obtain a response from the IDS output to an attacker manipulation input. The attacker may get no feedback, limited, or unlimited feedback.
- ***Manipulation Depth*** pertains to the degree or level of an adversary manipulation that can either analyse the problem space or feature space.

3.4 Threat Model

In this paper, we examine the risk posed by an insider threat actor with administrative access privileges to the network systems of the smart grid network. Insider threats represent a significant yet often overlooked danger to CI [12].

In other words, identifying and mitigating insider threats is a challenging and multifaceted endeavor due to the fact that insiders may have access to the network and often reside within enterprise-level security [22]. The German Federal Office for Information Security defines insider threats as individuals who can potentially misuse their access to information technology systems, sensitive data, or infrastructure. The following groups are specifically regarded as insider threats [1]:

- An individual who has direct access and physically interacts with control systems, such as operators or engineers.
- An individual with privileged access rights, such as system administrators.
- Anyone who has indirect access to the office network..
- Outsourced personnel or external service providers like those who develop and maintain software, suppliers, etc.

These adversaries can deploy several attacks, including:

- Social engineering can be employed to plan follow-up attacks. This involves identifying vulnerable employees, learning about internal processes, and mapping out the IT infrastructure to find potential weaknesses.
- Unauthorized acquisition or alteration of confidential data may occur when an individual accesses file servers, data storage media or historians. Primarily, the motives of such an attack are industrial espionage and whistle-blowing.
- Deliberate acts of sabotage against the company. This may include an attacker modifying control components or installing malware in the system for political or economical reasons. These actions, which may involve an attacker changing control components or installing malware in the system, may be driven by political or financial interests.

As shown in Fig. 3, our research is based on a realistic scenario where an insider already possesses legitimate access to the smart grid network through privileges such as administration and physical access. The network has an ML-NIDS model that should be able to detect any network attack. Given the difficulties associated with analyzing individual packets, this study considers the NIDS to be a flow-based system rather than a packet-based system in high-speed network environments. In the power system scenario presented in Sect. 3, given the capabilities of the attacker in Sub-Sect. 3.3, it is presumed that the adversary is interested in launching an evasion attack. Given the adversary's position, it is assumed that he or she is familiar with the features used by the IDS for classification; nonetheless, he or she is unfamiliar with the detector's specific algorithm configuration. The attacker's primary objective is to identify how to circumvent the NIDS. This will allow him/her to either launch more damaging attacks in the future or exploit the organization for personal gain by selling this information to competitors, ultimately leaving the organization exposed and susceptible to harm. Due to the knowledge acquired by the adversary, this type of attack can be classified as a grey box attack. This threat scenario presented in Fig. 3 was used to generate adversarial data for testing on trained ML model as presented in Sect. 5.

4 Attack Generation

This research examines the use of JSMA and FGSM attack techniques in a grey box setting, where the attacker is aware of the complete datasets and the features but is unaware of the target model. Even if the attacker has no information about the target model, generating samples that will force the target model to declassify the given model, particularly using other ML models, will be possible. This is because adversarial samples are transferable across different ML models.

When creating adversarial traffic, there are four key steps involved, as depicted in Fig. 4. In the first step of this attack model, the attacker generates adversarial traffic to deceive the surrogate model, which was trained by sniffing the actual traffic flows. In step 2, adversarial traffic that goes undetected by using the surrogate model is received and analyzed by the attacker.

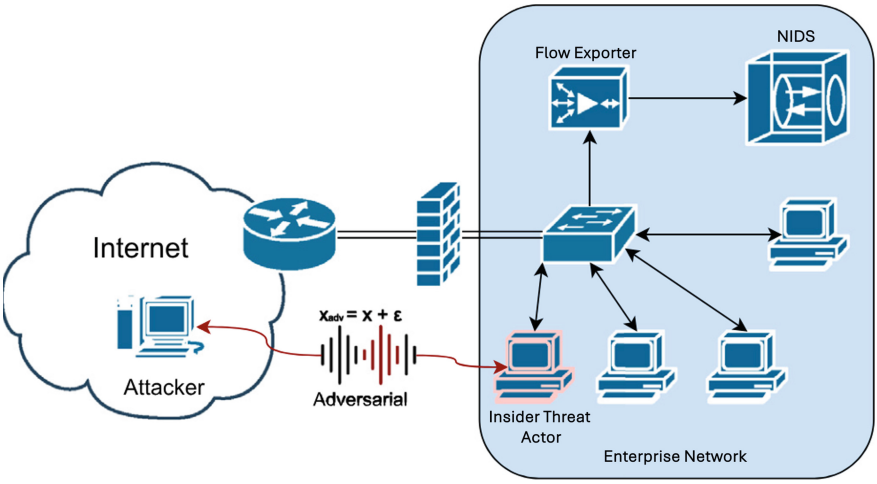

Fig. 3. Illustration of the considered threat scenario

The transferability feature is employed by the attacker in step 3 to send the adversarial traffic to defender NIDS. The adversarial traffic that managed to get past the defender NIDS will reach the insider threat actor's computer in step 4. In this research, the attacks were implemented through the Adversarial Robustness Toolbox (ART)[4]. ART is a Python tool that can generate various adversarial attacks.

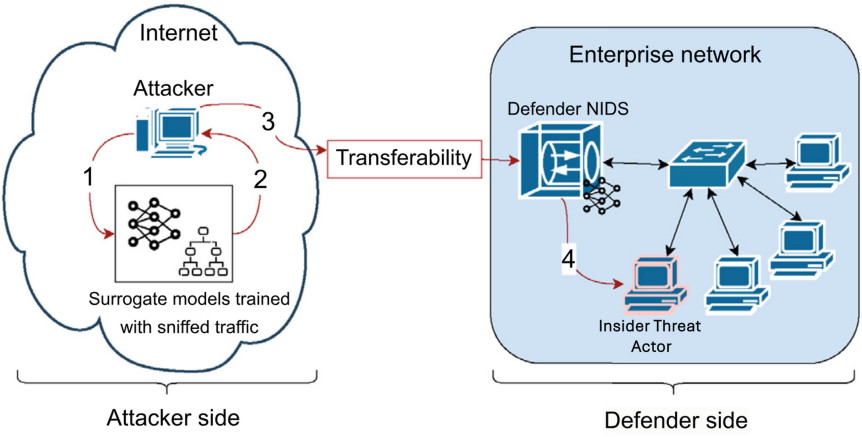

Fig. 4. Illustration of the adversarial traffic generation.

[4] See https://github.com/Trusted-AI/adversarial-robustness-toolbox [05/24].

4.1 Machine Learning-Based NIDS

This research uses some classical ML algorithms to assess how effectively supervised classification techniques can learn to detect cyberattacks in a smart grid environment. On the defender side, the defender uses Random Forest (RF), Naive Bayes (NB) and XGBoost (XGB) algorithms as a model for the NIDS. We selected these three ML models for our work because of their wide usage by the research community [21,25]. Additionally, the selected methods are easy to implement, less computational cost is needed, and they work well with annotated data, making them a suitable choice for our NIDS. Figure 5 illustrates how these algorithms use the same training and testing procedures. The experimental setup offers a standard platform for performance comparison, which aids in determining which model performs the best.

4.2 Hyper Parameters Optimization

XGboost and Naive Bayes models were trained using the default parameters provided by the scikit-learn framework[5]. To ensure optimal performance of RF as recommended by Zhu et al. [38] key tunable hyperparameters were applied, including the number of trees (100), the split method (Gini), and the minimum number of samples required to split (2). Hyperparameter tuning fixes the best value from the search space for the algorithm's parameters. Even though hyperparameter tuning was not conducted in this study, it opens an opportunity for future research.

4.3 Model Training and Testing

To ensure the usability of the research results, the original dataset was divided into subsets and stratified based on their labels. Both the data subsets are the same in terms of size as well as distribution of data. The first subset is assigned for training and evaluation on the defender side. As shown in Fig. 5, the second dataset is used by the attacker to train a surrogate model. Section 3 details how an insider threat scenario can make an attacker obtain this data through network sniffing. To validate the data, the datasets of the defender and the attacker are divided based on 70% of the training data and 30% of the testing data. In both training and testing datasets, malicious traffic is equal to benign traffic. The datasets are divided in this way to be as balanced in representation as possible, thereby minimizing the issue of data imbalance.

While this is a requirement to get the most of our envisaged IDSes, it is not essential to the actual claim of the paper. We are aware of the fact that real traffic data may be rather unbalanced and tuning the IDS to work in those contexts may overcome that problem. However, our evasion attack methodology is not dependent on this aspect. As shown in Fig. 5, the threat model follows the same training and testing process. Besides, the standard ML evaluation metrics were used to measure the performance of the threat models.

[5] See https://scikit-learn.org/stable/ [05/24].

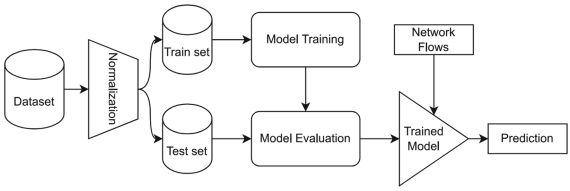

Fig. 5. The training and testing pipeline for an attacker and a defender.

4.4 Performance Metrics

Various evaluation metrics can be utilized when assessing the performance of different IDS models [15]. As shown in Table 1, all the metrics are based on the confusion matrix.

Table 1. IDS Confusion Matrix

Actual Class	Predicted Class	
	Anomaly	Normal
Anomaly	True Positive (TP)	False Negative (FN)
Normal	False Positive (FP)	True Negative (TN)

– Recall, which is sometimes called the "detection rate," helps quantify the proportion of actual positive instances that are correctly identified by the model.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

– Precision evaluates the accuracy of the positive predictions made by a model. Specifically, precision measures the proportion of predicted positive instances that are correct.

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

– The F1 score provides the performance of the combined metrics and is calculated as the harmonic mean of recall and precision. In this respect, the use of the F1 score allows the system not only to offer relevant results but also to refuse the others.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{5}$$

5 Evaluation

This section details the experimental findings. In Sub-Sect. 5.1, the performance of the attacker and defender models on the given metrics, such as F1 score, precision, and recall, is discussed. Moreover, a discussion of the model's ability to perform in an adversarial environment is discussed in Sect. 5.2. Lastly, a detailed analysis of the perturbation difference between the initial malicious instance and the adversarial instance is also provided.

5.1 ML-IDS Model Performance in Clean Settings

Different metrics are used to evaluate the model on clean settings, from both the attacker and defender perspectives. These include recall (Eq. 3), precision (Eq. 4), and F1-score (Eq. 5). In clean settings, the ML-IDS models performed binary classification to distinguish between malicious and benign traffic. As shown in Table 2, the performance of the trained ML models. XGBoost performed better compared to Naive Bayes and the Random Forest models. In general, these initial results demonstrate a good performance of the Random Forest and XGBoost. The F1 scores achieved by the classifiers were 0.845, 0.567 and 0.925 respectively.

Table 2. ML-IDS model performance in clean settings

Classifier	Accuracy	Precision	Recall	F1 Score	Time (s)
Random Forests	0.8473	0.8655	0.8473	0.8454	30
XGBoost	0.9464	0.9075	0.9464	0.9252	45
Naive Bayes	0.5742	0.5831	0.5742	0.5673	24

5.2 ML-IDS Models Performance in Adversarial Settings

In this sub-section, we demonstrate that applying the perturbations using JSMA and FGSM reduces the performance of our trained model. Further, we check whether the missed attacks (in the bigger false negative set) are still attacks and whether they will inpact the power system. To investigate how different parameter combinations can affect the model performance, multiple adversarial samples were created from the testing data using epsilon (ϵ) values ranging from 0 to 0.45 Although the current literature does not recommend a standard value for ϵ, in our research, we adopted a range between 0 to 0.45 to test attack success rates as suggested by Goodfellow et al. [14].

The adversarial dataset was then generated using different (ϵ) values. To determine how the performance of the models could be affected, the adversarial samples were then combined with the benign testing data and tested on the trained model. Figure 6 shows the overall performance for different adversarial

combinations. As the (ϵ) values increased, the model accuracy decreased further. For instance, XGBoost performance decreased from 94.64% at (ϵ) = 0 to 72.03% at (ϵ)= 0.45. On the other hand, Random Forest performance decreased from 84.73% at (ϵ) = 0 to 68.02% at (ϵ) = 0.45. Lastly, the performance for Naive Bayes decreased from 57.42% at (ϵ) = 0 to 32.05% at (ϵ) = 0.45. For FGSM adversarial attack, the attack success rate increased with higher (ϵ) values hence the accuracy declined because to the ML model was deceived by the attack.

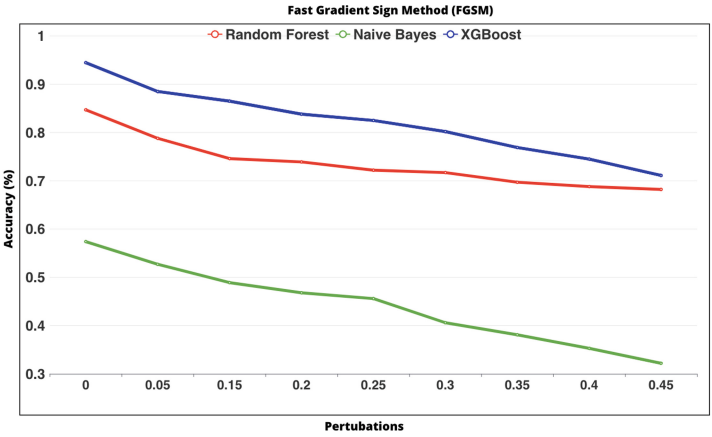

Fig. 6. Performance of trained models in adversarial settings when increasing perturbations (ϵ).

Selecting an appropriate (ϵ) value that will control the perturbation size is very crucial, as a higher (ϵ) value may increase attack success rates but may also increase the detectability of the adversarial samples. Small perturbations are ideal for launching a realistic attack and remain undetected by the IDS. For instance, when $\epsilon = 0.05$, the XGBoost accuracy dropped from 94.64% to 88.04% while Random Forest accuracy dropped from 84.73% to 78.43% and the Naive Bayes accuracy dropped from 57.42% to 52.89%. To consider another instance, when $\epsilon = 0.001$, XGBoost accuracy dropped from 94.64% to 92.36%, the Random Forest accuracy dropped from 84.73% to 82.67% and Naive Bayes accuracy dropped from 57.42% to 55.32% as detailed in Table 3. As per the adversarial performance, all the metrics declined in comparison to the performance of the original datasets in a clean setting.

Naive Bayes exhibited a greater performance drop compared to Random Forest and XGBoost. This could indicate that malicious data is misclassified by Naive Bayes due to its increased sensitivity. On the other hand, XGBoost's classification performance was better. This could suggest that XGBoost is a more reliable classifier when it comes to accurately discriminating between malicious and benign data. The first experiment focuses on JSMA, as indicated in Table 4, to investigate the effects of adversarial instances generated in our evasion attack

Table 3. Degradation of trained models in FGSM adversarial settings ($\epsilon = 0.001$).

Classifier	Accuracy	Precision	Recall	F1 Score
Random Forests	0.8267	0.6667	0.8267	0.6656
XGBoost	0.9236	0.7690	0.9236	0.7493
Naive Bayes	0.5532	0.2808	0.5532	0.2976

and the success of transferring adversarial instances generated by the attacker to the defender's trained model.

To analyze the impact of adversarial instances, recall is used as the detection rate parameter. It quantifies the number of adversarial occurrences classified by the IDS as malicious. The attacker initially creates adversarial instances for each model learned on his or her side (RF, XGBoost, and NB). The adversarial examples generated for one model are then transferred to the other models to determine the transferability of the attacker's trained models.

Table 4. Degradation of trained models in JSMA adversarial settings.

Classifier	Accuracy	Precision	Recall	F1 Score
Random Forests	0.6416	0.6605	0.6434	0.5980
XGBoost	0.7090	0.7289	0.7355	0.7234
Naive Bayes	0.2934	0.2789	0.2978	0.2784

Further, we investigated the incorrectly classified inputs to see if they could cause harm. For example, we use JSMA to alter the features of a malicious data point using various variants of θ and γ. These examples show that the greater the value of θ, the greater is the perturbation of the features. The R1-PA1:VH feature shows an increase from 0.7545 to 0.7550 for $\theta = 0.1$ and 0.5, and to 1 when $\theta = 0.9$. Similarly, the greater the value of γ, the more features are manipulated. Similarly, the higher the value of γ, the more features are perturbed. Table 5 illustrates the frequency of transmitted power features R1:F, R2:F, R3:F, and R4:F.

The frequency of transmitted power for R1:F, R2:F, R3:F, and R4:F increased significantly from an average of 60 Hz to an average of 62 Hz after perturbation. In power systems, the frequency of transmitted power from the grid is kept within a certain range, usually around 60 Hz in North America and 50 Hz in other regions of the world. Maintaining the frequency within the allowed range is critical for the power grid's reliable operation and preventing damage to connected equipment. If the frequency were to rise above 60.3 Hz, several issues could arise [7]. Therefore, from our results, we are confident that this particular attack that was undetected could damage the system, disrupt industrial processes, or cause harm.

Table 5. An example of how the frequency of transmitted power features are perturbed using JSMA.

Dataset	R1-PA1:VH	R1:F	R2:F	R3:F	R4:F
Original Test Data	0.7545	59.986	59.986	59.984	59.988
$\theta = 0.1, \gamma = 0.1$	0.7550	60.059	60.196	60.294	60.468
$\theta = 0.5, \gamma = 0.5$	0.7650	61.574	61.754	61.537	61.643
$\theta = 0.9, \gamma = 0.9$	1.0000	62.566	62.676	62.464	62.376

6 Conclusion

The increasing adoption of NIDS based on ML algorithms presents interesting security challenges. Despite their exceptional performance, these ML models are vulnerable to a wide range of adversarial techniques, including evasion attacks. This paper demonstrated the importance of realistic threat modelling in the context of adversarial attacks on smart grid systems. By highlighting real attacker capabilities and feasible attack scenarios, this research provides a more practical and applicable perspective compared to the existing literature, which often deals with hypothetical or idealized models. Moreover, this research performs an empirical evaluation using a power system dataset generated from a smart grid testbed, which adds significant value, grounding the theoretical insights in real-world data.

To our best knowledge, this is the first realistic approach that aims to evade the NIDS by leveraging on the transferability property without relying on any query methods and with a minimal understanding of the target NIDS. This approach operates within the traffic space and adheres to domain constraints. This paper demonstrates a realistic adversarial approach designed to generate valid and realistic adversarial network traffic by introducing minor perturbations. This allows for bypassing the NIDS protection with a high probability while preserving the core logic of the underlying model. The experiments detailed in this research have shown that evasion attacks can be successfully generated using JSMA and FGSM methods, impacting the classification performance of Random Forest, Naive Bayes and the XGBoost ML models.

Furthermore, our results show that the same set of adversarial examples that managed to deceive one classifier also succeeded in deceiving the other classifiers. For instance, the adversarial samples generated by FGSM managed to decrease the performance of XGBoost from 94.64% to 72.03%, Random Forest from 84.73% to 68.02% and Naive Bayes from 57.42% to 32.05%. This observation can be considered additional evidence for the transferability phenomenon first alluded to by Papernot et al. [27] within the image recognition domain and by Sheatsley et al. [33] within the network intrusion detection domain. Our work in the smart grid domain makes it clear that all three classifiers are vulnerable to adversarial perturbations.

6.1 Future Work

In this research, we have shown how adversarial attacks can be generated successfully by employing both the JSMA and FGSM methods and adversely affect how state-of-the-art supervised models are classified. However, future studies can consider how other methods, such as Carlini Wagner (CW) and Generative Adversarial Networks (GAN) can be used to generate adversarial attacks. In this regard, the current study can be extended and explored more in future works by comparing how different methods of generating adversarial attacks are different or the same. Moreover, adversarial attacks should be investigated against other ML models.

The adversarial attacks against ML models are not limited to the domain of IDS systems but to all systems where ML techniques are implemented. Future research in areas like federated learning can consider the direction of awareness, defence, and mitigation of adversarial attacks against ML [8,26]. Therefore, it would be interesting for further studies to evaluate the applicability of the proposed model in a distributed setting. As mentioned before, there is a great need for research on suitable mitigation techniques against adversarial threats.

Acknowledgments. The first and third author were supported by the Brno University of Technology project "Smart information technology for a resilient society", 2023–2025, code FIT-S-23-8209. The second author was supported by the RICS centre financed by the Swedish Civil contingencies agency (MSB) and project AIR^2 supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP), funded by the Knut and Alice Wallenberg Foundation.

References

1. Industrial Control System Security. Insider threat. https://www.allianz-fuer-cybersicherheit.de/. Accessed 22 Jan 2024
2. Power system attack datasets. http://www.ece.uah.edu/thm0009/icsdatasets. Accessed 30 Jan 2024
3. Alatwi, H.A., Morisset, C.: Realism versus performance for adversarial examples against DL-based NIDS. In: Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing, pp. 1549–1557 (2023)
4. Alshamrani, A., Myneni, S., Chowdhary, A., Huang, D.: A survey on advanced persistent threats: techniques, solutions, challenges, and research opportunities. IEEE Commun. Surv. Tutor. **21**(2), 1851–1877 (2019)
5. Apruzzese, G., Andreolini, M., Ferretti, L., Marchetti, M., Colajanni, M.: Modeling realistic adversarial attacks against network intrusion detection systems. Digit. Threats Res. Pract. (DTRAP) **3**(3), 1–19 (2022)
6. Asimopoulos, D.C., et al.: Breaching the defense: investigating FGSM and CTGAN adversarial attacks on IEC 60870-5-104 AI-enabled intrusion detection systems. In: Proceedings of the 18th International Conference on Availability, Reliability and Security, pp. 1–8 (2023)
7. Barrera-Cárdenas, R., Molinas, M.: Modelling of power electronic components for evaluation of efficiency, power density and power-to-mass ratio of offshore wind power converters. In: Offshore Wind Farms, pp. 193–261. Elsevier (2016)

8. Bouacida, N., Mohapatra, P.: Vulnerabilities in federated learning. IEEE Access **9**, 63229–63249 (2021). https://doi.org/10.1109/ACCESS.2021.3075203
9. Clements, J., Yang, Y., Sharma, A.A., Hu, H., Lao, Y.: Rallying adversarial techniques against deep learning for network security. In: Symposium Series on Computational Intelligence (SSCI), pp. 01–08. IEEE (2021)
10. Debicha, I., Debate, T., Dricot, J.M., Mees, W.: Adversarial training for deep learning-based intrusion detection systems. In: The Sixteenth International Conference on Systems (ICONS) (2021)
11. Fu, X., Zhou, N., Jiao, L., Li, H., Zhang, J.: The robust deep learning-based schemes for intrusion detection in internet of things environments. Ann. Telecommun. **76**(5–6), 273–285 (2021)
12. Gollmann, D.: From insider threats to business processes that are secure-by-design. In: Third International Conference on Intelligent Networking and Collaborative Systems, pp. 627–627 (2011).https://doi.org/10.1109/INCoS.2011.175
13. Goodfellow, I., et al.: Generative adversarial nets. Adv. Neural Inf. Process. Syst. **27** (2014)
14. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples (2015). https://arxiv.org/abs/1412.6572
15. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques, 3rd edn. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2011)
16. Hashemi, M.J., Cusack, G., Keller, E.: Towards evaluation of NIDSs in adversarial setting. In: Proceedings of the 3rd ACM CONEXT Workshop on Big Data, Machine Learning and Artificial Intelligence for Data Communication Networks, pp. 14–21 (2019)
17. Huang, C.H., Lee, T.H., Chang, L.H., Lin, J.R., Horng, G.: Adversarial attacks on SDN-based deep learning IDS system. In: International Conference on Mobile and Wireless Technology (ICMWT), pp. 181–191. Springer (2019)
18. Jeong, J., Kwon, S., Hong, M.P., Kwak, J., Shon, T.: Adversarial attack-based security vulnerability verification using deep learning library for multimedia video surveillance. Multimed. Tools Appl. **79**, 16077–16091 (2020)
19. Kravchik, M., Shabtai, A.: Detecting cyber attacks in industrial control systems using convolutional neural networks. In: Proceedings of the Workshop on Cyber-Physical Systems Security and Privacy, pp. 72–83. CPS-SPC '18, Association for Computing Machinery (2018)
20. Linda, O., Vollmer, T., Manic, M.: Neural network based intrusion detection system for critical infrastructures. In: International Joint Conference on Neural Networks, pp. 1827–1834. IEEE (2009)
21. Liu, C., Gu, Z., Wang, J.: A hybrid intrusion detection system based on scalable k-means+ random forest and deep learning. IEEE Access **9**, 75729–75740 (2021)
22. Liu, L., De Vel, O., Han, Q.L., Zhang, J., Xiang, Y.: Detecting and preventing cyber insider threats: a survey. IEEE Commun. Surv. Tutor. **20**(2), 1397–1417 (2018)
23. Martins, N., Cruz, J.M., Cruz, T., Abreu, P.H.: Analyzing the footprint of classifiers in adversarial denial of service contexts. In: 19th EPIA Conference on Artificial Intelligence, Proceedings, Part II, pp. 256–267. Springer (2019)
24. Merzouk, M.A., Cuppens, F., Boulahia-Cuppens, N., Yaich, R.: A deeper analysis of adversarial examples in intrusion detection. In: Risks and Security of Internet and Systems: 15th International Conference, CRISIS 2020, Revised Selected Papers 15, pp. 67–84. Springer (2021)

25. Min, E., Long, J., Liu, Q., Cui, J., Chen, W.: Anomaly-based intrusion detection through text-convolutional neural network and random forest. Secur. Commun. Netw. (2018)
26. Nguyen, T.D., Rieger, P., Miettinen, M., Sadeghi, A.R.: Poisoning attacks on federated learning-based IoT intrusion detection system. In: Proceedings of the Workshop Decentralized IoT System Security (DISS), vol. 79 (2020)
27. Papernot, N., McDaniel, P., Goodfellow, I.: Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. arXiv preprint arXiv:1605.07277 (2016)
28. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A.: The limitations of deep learning in adversarial settings. In: European Symposium on Security and Privacy (EuroS&P), pp. 372–387. IEEE (2016)
29. Papernot, N., McDaniel, P., Sinha, A., Wellman, M.P.: SOK: security and privacy in machine learning. In: European Symposium on Security and Privacy (EuroS&P), pp. 399–414. IEEE (2018)
30. Peng, X., Huang, W., Shi, Z.: Adversarial attack against DOS intrusion detection: an improved boundary-based method. In: 31st International Conference on Tools with Artificial Intelligence (ICTAI), pp. 1288–1295. IEEE (2019)
31. Peng, Y., Su, J., Shi, X., Zhao, B.: Evaluating deep learning based network intrusion detection system in adversarial environment. In: 9th International Conference on Electronics Information and Emergency Communication (ICEIEC), pp. 61–66. IEEE (2019)
32. SectorCERT: The attack against Danish, critical infrastructure. https://sektorcert.dk/wp-content/uploads/2023/11/SektorCERT-The-attack-against-Danish-critical-infrastructure-TLP-CLEAR.pdf. Accessed 01 Jan 2024
33. Sheatsley, R., Papernot, N., Weisman, M., Verma, G., McDaniel, P.: Adversarial examples in constrained domains (2022). https://arxiv.org/abs/2011.01183
34. Sriram, S., Simran, K., Vinayakumar, R., Akarsh, S., Soman, K.: Towards evaluating the robustness of deep intrusion detection models in adversarial environment. In: International Symposium on Security in Computing and Communication, pp. 111–120. Springer (2019)
35. Tavallaee, M., Bagheri, E., Lu, W., Ghorbani, A.A.: A detailed analysis of the KDD CUP 99 data set. In: Symposium on Computational Intelligence for Security and Defense Applications, pp. 1–6. IEEE (2009).https://doi.org/10.1109/CISDA.2009.5356528
36. Wang, Z.: Deep learning-based intrusion detection with adversaries. IEEE Access **6**, 38367–38384 (2018)
37. Warzyński, A., Kołaczek, G.: Intrusion detection systems vulnerability on adversarial examples. In: Innovations in Intelligent Systems and Applications (INISTA), pp. 1–4. IEEE (2018)
38. Zhu, N., Zhu, C., Zhou, L., Zhu, Y., Zhang, X.: Optimization of the random forest hyperparameters for power industrial control systems intrusion detection using an improved grid search algorithm. Appl. Sci. **12**(20), 10456 (2022)