Rigorous Evaluation of Machine Learning-based Intrusion Detection Against Adversarial Attacks

Onat Gungor, Elvin Li, Zhengli Shang, Yutong Guo, Jing Chen, Johnathan Davis, and Tajana Rosing

Department of Computer Science and Engineering, University of California, San Diego

Abstract—**The rapid growth of the Internet of Things (IoT) has engendered profound security challenges. Intrusion detection system (IDS) is a security measure to mitigate these challenges by continuously monitoring system data and alerting to any suspicious activity. While machine learning (ML) has emerged as a promising IDS solution, its vulnerability to adversarial attacks raises concerns about the reliability of these systems. In this paper, we present a rigorous evaluation framework to assess the performance of ML-based IDS against various adversarial attacks in IoT environments. Our framework employs a wide range of adversarial attack techniques, including white-box, gray-box, and black-box adversarial attacks, across four realistic and recent IoT intrusion datasets. Our results showed that the intrusion detection performance of state-of-the-art ML and DL models deteriorates by up to 49.5× under adversarial attacks. This observation indicates an urgent need for more resilient ML-IDS solutions against adversarial attacks in IoT systems.**

I. Introduction

The Internet of Things (IoT) is the convergence of the Internet and smart objects that can communicate and interact with each other [1]. By 2030, its value could be worth $12.6 trillion globally [2]. Although various application domains such as home automation, industrial processes, and environmental monitoring, benefit from IoT systems, these systems also present several security risks. These risks stem from diverse standards and communication stacks, limited computing power, and a large number of interconnected devices [1]. Most importantly, cyber-attacks against IoT systems can cause enormous financial and reputational losses by impacting critical infrastructures such as power plants and transportation systems [3]. Hence, securing IoT systems plays a critical role in minimizing the impact of these cyber-attacks.

An Intrusion Detection System (IDS) is an IoT security measure that monitors the operations of a host or a network in a continuous manner to detect malicious activities [4]. Machine learning (ML) has become a popular IDS solution due to its strong generalizability in detecting attack variants and novel attacks [5]. An adversarial attack is a type of cyber-attack that imperceptibly perturbs input data and forces ML models to produce incorrect outputs leading to incorrect predictions [6]. These attacks raise significant challenges to the reliability of machine learning-based intrusion detection systems (ML-IDS) in IoT environments. There are a few studies aiming to analyze ML-based IDS against adversarial attacks [7]–[10], yet they fail to accurately capture the magnitude of the risks that adversarial attacks pose to current ML-IDS solutions.

ML-based IDS solutions, which earned their reputation by leveraging ML algorithms such as Random Forests and Deep Neural Networks, are still highly susceptible to adversarial attacks, deeply undermining their effectiveness in securing IoT systems [8]. To address this issue, we propose a comprehensive evaluation framework that conducts a quantitative, multi-model analysis of ML-IDS against adversarial attacks across a broad range of realistic IoT intrusion datasets. Our approach distinguishes itself from previous attempts at tackling this issue in several key aspects. We employ a wide range of adversarial techniques, using white-box, gray-box, and black-box attacks to thoroughly assess the vulnerabilities of ML-IDS. To ensure our findings can be generalized and applicable to the real-world IoT settings, we utilize four diverse and up-to-date IoT intrusion datasets, namely MQTTset [11], WUSTL-IIoT [12], X-IIoTID [13], and Edge-IIoTset [14]. Our results showed that the macro F_1 score of the XGBoost model for the WUSTL-IIoT dataset drops from 99% baseline to 2% under the Fast Gradient Sign Method attack (49.5× performance loss), demonstrating an urgent need for the development of more resilient ML-IDS solutions against adversarial attacks.

II. Related Work

IoT security is challenging due to its complex environment and resource-constrained devices [15]. Failure to address security concerns makes IoT susceptible to security attacks, leading to enormous financial and reputational losses [3]. Globally, the number of IoT attacks reached over 10.5 million in December 2022 [16]. For instance, Mirai botnet attack has infected over half a million IoT devices in late 2016 [17]. Distributed denial-of-service, man-in-the-middle, jamming, eavesdropping, side channel, and spoofing are some IoT cyberattacks [18]. Despite the availability of traditional security with encryption, authentication, and access control, IoT systems are still vulnerable to multiple attacks to disrupt the network, necessitating a second line of defense [19]. Intrusion detection system (IDS) is a pervasive IoT security measure that monitors network traffic or system events to identify and alert any suspicious activity [4]. IDS can be categorized into two groups: Signature-based Intrusion Detection System (SIDS) and Anomaly-based Intrusion Detection System (AIDS) [20]. SIDS uses a rule-based engine to match known patterns. The main drawback of SIDS is the inability to identify zero-day (unknown) attacks since there is no matching signature in the database. As a remedy to this approach, AIDS identifies patterns that deviate

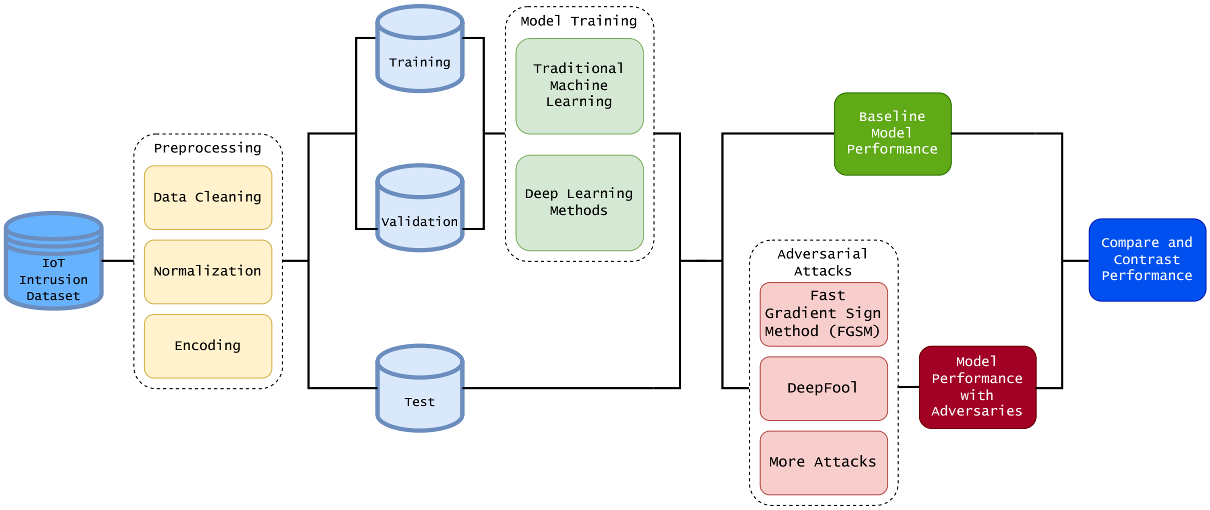


Fig. 1: Adversarial ML-IDS evaluation framework

from normal behavior based on statistical-based, knowledge-based, and machine learning-based methods.

ML-based AIDS trains ML models using historical system data [4]. It has become prevalent due to improved prediction performance and lower requirement for human knowledge. Various ML methods, including support vector machine, random forest, deep neural network, and recurrent neural network have been proposed [5]. Although ML methods provide a great intrusion detection performance, they can be vulnerable to small changes in the input data [21]. Adversarial attacks involve creating deliberately altered input data to misguide ML models in their decision-making, e.g., predicting malicious attacks as benign. These attacks can be classified into white-box, gray-box, and black-box attacks, depending on how much knowledge of the targeted model the attacker has [22]. White-box attacks assume complete knowledge of the model architecture. Gray-box attacks leverage white-box models and transfer them to other models. Black-box attacks rely solely on the model's outputs. These adversarial attacks can compromise the integrity of IDS predictions, leading to security breaches.

Several studies focused on analyzing the performance of ML-IDS under adversarial attacks. A recent survey by Alotaibi and Rassam [9] list various white-box, gray-box, and black-box adversarial attacks against ML-based IDS. They present an analysis of the adversarial attack and defense strategies. However, no attacks or defense mechanisms were implemented and evaluated by the authors. Qiu et al. [23] explore salient map adversarial attacks against network intrusion detection in IoT systems. They report that the adversary could modify less than 0.005% of the bytes in a malicious packet to achieve an average attack success rate of 94.3%. This paper is limited to only exploring the effect of one type of adversarial attack. Thakkar and Lohiya [7] discuss the broader challenges and future research directions for ML-based IDS. Their work highlights the ongoing development of more robust detection mechanisms. Additionally, Macas et al. [24] survey the landscape of deep-learning-enabled adversarial examples in cybersecurity systems, where they detail both the attack methods and the novel defenses. He, Kim, and Asghar [8] in their comprehensive survey, highlight the distinctive vulnerabilities of deep learning-based IDS to adversarial attacks. Unlike previous evaluation-based studies, we leverage an integrated approach by applying state-of-the-art adversarial techniques across realistic and up-to-date IoT datasets. Our work extends the understanding of IDS vulnerabilities by implementing a quantitative, rigorous, cross-dataset, multi-model analysis that assesses the transferability and impact of adversarial examples, thereby offering innovative insights into the development of resilient, next-generation IDS solutions.

III. Proposed Framework

A. Overview

Fig. 1 presents our adversarial attack ML-IDS evaluation framework. After data pre-processing (data cleaning, normalization, and encoding), we split the IoT intrusion data into training, validation, and test sets. We use the training set to train the selected ML and DL models, and validation set to search for the best hyperparameters. Then, we use the test portion to obtain the baseline performance of our models without any perturbation, i.e., adversarial attack. After that, we use trained deep neural network (DNN) to generate both white-box and black-box adversarial attacks on the original datasets. Furthermore, we transfer the DNN-based white-box attack samples to other pre-trained models to assess their performance on gray-box (transfer) attacks. We evaluate the performance of the models under adversarial attacks. Given the predictive performance of models with and without adversarial attacks, we ultimately compare the effects of attacks on different models on ML-IDS solutions.

B. Selected Machine Learning (ML) Methods

We select four traditional ML algorithms: Naive Bayes (NB), Decision Tree (DT), Random Forest (RF), and eXtreme Gradient Boosting (XGB). We also select three DL algorithms: Deep Neural Network (DNN), Convolutional Neural Network (CNN), and Recurrent Neural Network (RNN).

Naive Bayes (NB) [25] is an algorithm based on Bayes Theorem, assuming that all features are independent. Naive Bayes is good at dealing with categorical data and commonly used in classification problems due to its high efficiency.

Decision Tree (DT) [25] is used for classification and regression. It works by splitting the datasets into smaller subsets based on certain conditions. It is a simple and interpretable algorithm due to its creation of tree-like model of decisions.

Random Forest (RF) [25] is a DT-based ensemble technique which combines multiple DTs to reduce the risk of overfitting and increases the robustness to noise. Although RF requires higher computational costs compared to DT, it usually reaches higher accuracy and is more effective on large datasets.

eXtreme Gradient Boosting (XGB) [25] is an advanced implementation of gradient boosting algorithms. It optimizes traditional gradient boosting to reach faster speed and better performance. This implementation is efficient in handling large-scale data as well as reaching high accuracy.

Deep Neural Network (DNN) [25] mimics the way human brains operate, using layers of nodes to process data in complex ways. It is capable of learning from large and complicated data and making accurate predictions.

Convolutional Neural Network (CNN) [25] assumes dependent relationship among data points surrounding each other. The use of convolutional layer in CNN helps figure out the correlated feature among connected data points.

Recurrent Neural Network (RNN) [25] assumes data is sequence-dependent, making it ideal for processing time-series data. RNN is capable of maintaining a memory of previous inputs to capture the time-dependent feature and is thus efficient in doing prediction tasks based on historical data.

C. Adversarial Threat Model

Attack Surface: We employ the most common type of adversarial attack, evasion attack. This type of attack aims to evade trained models at the testing phase to cause incorrect predictions in inference time [26].

Adversarial Capabilities: We leverage white-box, gray-box, and black-box attacks. White-box attack has complete knowledge and access to the trained model. For white-box attacks, we create adversarial attack samples on trained DNN and evaluate on the same model. Gray-box attack has partial knowledge regarding the model and information during training [27]. We perform gray-box attack by transferring generated white-box attacks from DNN to other pre-trained ML and DL models. On the contrary, black-box attack has no knowledge of the model and only has access to the model outputs in response of specific inputs. We apply black-box attack on all pre-trained ML and DL methods.

TABLE I: Selected IoT intrusion datasets

Dataset	Year	Number of Features	Number of Attacks	Number of Samples
MQTTset [11]	2020	33	5	20M
WUSTL-IIOT [12]	2021	41	4	1M
X-IIoTID [13]	2021	59	18	0.9M
Edge-IIoTset [14]	2022	61	14	2.2M

Adversarial Goals: The selected adversarial attacks aim at misclassification. The attack manipulates the testing datasets so that the models classify it to the wrong class [26].

D. Selected Adversarial Attacks

Fast Gradient Sign Method (FGSM) [27] is a white-box attack which calculates the gradient of the loss function, and perturbations based on the sign of the calculated gradient are added to the original dataset. This method ensures minimal changes to the original dataset to deceive the model.

Basic Iterative Method (BIM) [28] is an extension of FGSM. Instead of applying perturbation to the original dataset in one step, BIM calculates the gradient multiple times and adds small perturbation to the dataset iteratively.

Projected Gradient Descent (PGD) [29] is a generalized version of BIM. PGD starts at a random perturbation value and applies gradient ascent to maximize the loss. PGD projects the inputs with perturbation back to their allowed range.

DeepFool (DF) [30] is a white-box attack which assumes a locally linear approximation of the decision boundary of the model and calculates the minimal perturbation that makes the input to cross the boundary to achieve misclassification.

Jacobian Saliency Map Attack (JSMA) [31] is a white-box attack that calculates the Jacobian matrix of the outputs of the model given the inputs to find out the predominant features. These features are then modified to deceive the model.

Elastic Net Attack (EN) [32] is a white-box attack that uses gradient-based approach to optimize the perturbation to the input datasets to maximize the impact on the model's predictions with the minimal changes to the original datasets.

Zeroth Order Optimization (ZOO) [33] is a black-box attack. It uses the model's output to estimate the gradients by finite difference methods. Based on estimated gradients, adversarial examples with optimal perturbation are generated.

IV. Experimental Analysis

A. Selected IoT Intrusion Datasets

Table I summarizes the selected IoT intrusion datasets based on their recency, number of features, number of attacks, and number of samples (in million). We cover a good range of up-to-date, diverse, and realistic IoT intrusion datasets.

MQTTset [11]: The MQTTset focuses on IoT networks with the MQTT protocol, the most common messaging protocol for IoT. The data collected are real logs from an MQTT network, while artificial cyber-attacks are generated to collect adversarial data. The focus on specific protocols in this dataset make it important to experiment with, as many others do not have support for dedicated protocols.

WUSTL-IIoT [12]: WUSTL-IIoT is an Industrial Internet of Things (IIoT) intrusion dataset aimed to emulate real-world industrial systems. The dataset is deliberately unbalanced to imitate real-world industrial control systems, consisting of 41 features and 1,194,464 observations.

X-IIoTID [13]: X-IIoTID is an IIoT intrusion dataset, taking into account the nature of the devices and connectivity protocols these large-scale systems employ. The dataset features are collected with the independence of devices and connectivity, generating a holistic intrusion data set to represent the heterogeneity of IIoT systems. It includes novel IIoT connectivity protocols, activities of various devices, as well as an abundance of attack scenarios.

Edge-IIoTset [14]: Edge-IIoT is an edge computing focused IIoT intrusion dataset, specifically in the realms of centralized and federated learning. The data is collected from IoT devices and IIoT connectivity relating to sensors, e.g., ultrasonic sensors, heart rate sensors, and fourteen attacks are identified that correspond to these devices. Various features are extracted to supplement the data of these attacks, including attacks, system resources, logs, and network traffic.

B. Experimental Setup

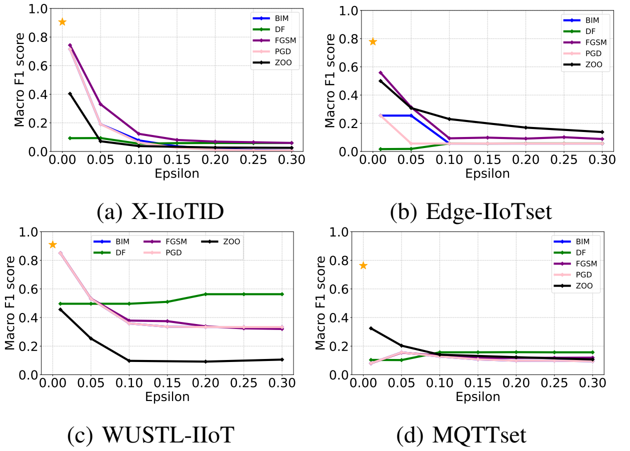
Hardware: We run our experiments on a Linux virtual machine server equipped with a 16-core CPU, 16 GB of RAM, and an NVIDIA RTX A6000 GPU.

Evaluation Metric: We select macro F_1 score due to unbalanced characteristic of our intrusion datasets. F_1 score is the harmonic mean of the precision and recall. Macro F_1 score calculates the F_1 score of each class and averages them.

Data pre-processing: We pre-process the raw system data by utilizing techniques including one hot encoding for categorical variables, z-score normalization as well as logarithmic transformations for standardizing the features, and removing lower-correlated columns to reduce the dimensionality. We also apply a 60-20-20 training-validation-test split.

ML and DL model training: For model training, we use prepackaged models from the scikit-learn machine learning API [34] for Python and leveraged PyTorch [35] to design the neural network architectures. We then employ hyper-parameter optimization techniques, e.g., grid search, to obtain the best model prediction performances over the validation data.

Adversarial attack generation: We use Adversarial Robustness Toolbox (ART) [36] to generate adversarial attacks due to its compatibility with a large number of attacks. For white-box adversarial attacks, we inject white-box attacks from III-D to the pre-trained DNNs. For elastic net attack, we test learning rate of $\{0.001, 0.01, 0.1\}$. For the remaining adversarial attacks, we experiment with perturbation amounts of $\{0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$, allowing us to obtain metrics for how potent they are across different model architectures and datasets. For the transfer (gray-box) attacks, we utilize the generated adversarial samples from pre-trained DNNs and transfer these instances to other target ML and DL models. This provides a more realistic adversarial attack setting since the generated samples were not developed with the target model architecture entirely known. For the black-box attacks, we use the same perturbation amounts and measure attack performance on the pre-trained ML and DL models.



(a) X-IIoTID

(b) Edge-IIoTset

(c) WUSTL-IIoT

(d) MQTTset

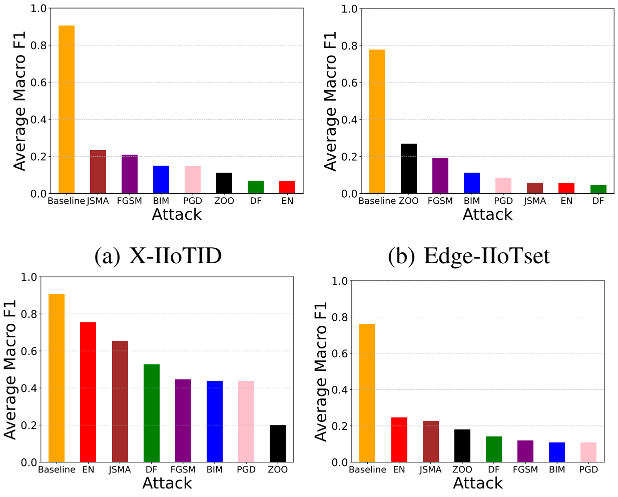
Fig. 2: DNN selected adversarial attack results (macro F_1 score) across varying perturbation amounts

C. Results

1) White-Box Attacks: Fig. 2 shows adversarial attack results under changing perturbation amount (epsilon) values. We mark the baseline (no adversarial attack) with orange star (top left corner at each sub-figure), and adversarial attacks are represented with different colors. Each sub-figure presents our results on a different dataset, e.g., X-IIoTID. While x-axis shows varying epsilon values, y-axis reports macro F_1 score. In these sub-figures, we also show our black-box attack (ZOO) result which is later explained in Section IV-C3.

We can observe that there is a trend of lower F_1 scores as epsilon increases across all datasets, with most of the substantial drops being prevalent within the interval $[0.01, 0.05]$. As epsilon increases further, we notice that the change in attack performance drops steeply, suggesting that there is a threshold as to how far each attack can deteriorate the model prediction performance. Additionally, BIM and PGD appear to closely follow each other in terms of performance deterioration across epsilon values due to their similar adversarial attack generation process. Despite the similarities across the datasets, some of them still deviate in terms of how each attack performed. Most prominently, we notice that WUSTL-IIoT (Fig. 2c) is more robust to the attacks with the most drastic performance drops occurring in PGD, BIM, and FGSM from 90.9% to 33.3%, which is considerably less of a difference compared to the others, i.e., 90.6% to 0.7% in X-IIoT, 77.8% to 1.67% in Edge-IIoTset, and 76.3% to 4.82% in MQTTset.

Fig. 3 presents DNN average F_1 scores (calculated over the selected epsilon values). While x-axis shows adversarial attacks, y-axis provides average macro F_1 score. For each



(a) X-IIoTID

(b) Edge-IIoTset

(c) WUSTL-IIoT

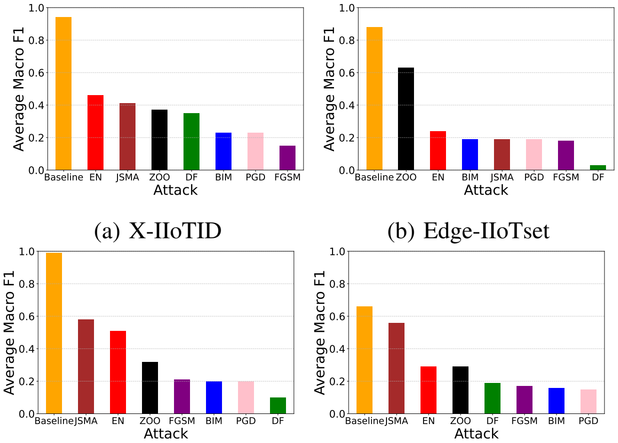
(d) MQTTset

Fig. 3: Deep neural network adversarial attack results

sub-figure, we represent the baseline (no adversarial attack) with an orange color, and adversarial attacks are represented with different colors. We observe similar levels of performance loss for BIM, PGD, and FGSM across all datasets, although the general trend appears to be that PGD is the strongest attack, i.e., the largest prediction performance loss. However, the margin between BIM and PGD performances are not drastic due to the similar nature of both attack algorithms. Interestingly, in X-IIoTID and Edge-IIoTset, EN and DF seem to be the strongest attacks. EN displayed a macro F_1 score degradation of 90.6% to an average 6.6% and 77.8% to an average of 5.6% of each dataset, respectively. Likewise, DF showed a macro F_1 score deterioration from an average of 90.6% to an average of 6.8% on X-IIoTID and an average of 77.8% to an average of 5% on Edge-IIoTset. This trend does not appear to be present in WUSTL-IIoT and MQTTset, whose respective ENs created performance drops from an average 90.9% to an average of 75.4% and 76.3% to an average of 24.7%. Furthermore, DF only decreased the performances on WUSTL-IIoT from averages of 90.9% to 52.7%, and from averages of 76.3% to 14.2% on MQTTset.

2) Gray-Box Attacks: Table II, III, IV, V present our transfer-attack results for each adversarial attack and target model. In these tables, we report the average F_1 scores calculated over selected epsilon values. We also select DT and RNN to illustrate their results in Fig. 4 and Fig. 5 respectively.

The results demonstrate several common patterns across each model and dataset. One noticeable trend is that the DF attack consistently achieves the highest performance in terms of degradation of model performance across all models and datasets. For instance, in the Edge-IIoTset (Table III), the DF attack reduces the F_1 score of the DT, RNN, and XGB models to 0.03, which are among the lowest of the scores for each respective model. This pattern is repeated in WUSTL-IIoT and



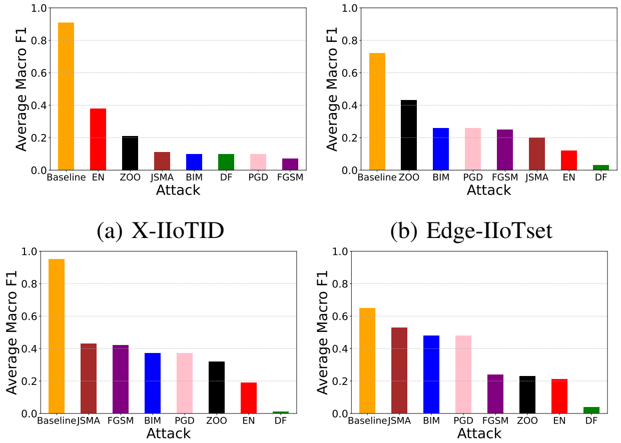
(a) X-IIoTID

(b) Edge-IIoTset

(c) WUSTL-IIoT

(d) MQTTset

Fig. 4: Decision tree adversarial attack results



(a) X-IIoTID

(b) Edge-IIoTset

(c) WUSTL-IIoT

(d) MQTTset

Fig. 5: Recurrent neural network adversarial attack results

TABLE II: X-IIoTID adversarial attack F_1 scores

X-IIoTID	**NB**	**DT**	**RF**	**XGB**	**CNN**	**RNN**
Baseline	**0.62**	**0.94**	**0.94**	**0.96**	**0.90**	**0.91**
DF [30]	**0.04**	0.35	0.17	0.12	0.34	0.10
FGSM [27]	0.19	**0.15**	**0.12**	**0.08**	**0.11**	**0.07**
BIM [28]	0.26	0.23	0.18	0.12	0.15	0.10
PGD [29]	0.25	0.23	0.18	0.12	0.15	0.10
JSMA [31]	0.05	0.41	0.71	0.67	0.40	0.11
EN [32]	0.06	0.46	0.49	0.35	0.65	0.38
ZOO [33]	0.26	0.37	0.40	0.35	0.23	0.21

TABLE III: Edge-IIoTset adversarial attack F_1 scores

Edge-IIoTset	**NB**	**DT**	**RF**	**XGB**	**CNN**	**RNN**
Baseline	**0.68**	**0.88**	**0.87**	**0.86**	**0.72**	**0.72**
DF [30]	**0.06**	**0.03**	**0.06**	**0.03**	**0.09**	**0.03**
FGSM [27]	0.23	0.18	0.23	0.11	0.25	0.25
BIM [28]	0.23	0.19	0.26	0.12	0.27	0.26
PGD [29]	0.23	0.19	0.26	0.12	0.27	0.26
JSMA [31]	0.13	0.19	0.33	0.17	0.20	0.20
EN [32]	0.09	0.24	0.33	0.22	0.14	0.12
ZOO [33]	0.42	0.63	0.71	0.76	0.42	0.43

TABLE IV: WUSTL-IIoT adversarial attack F_1 scores

WUSTL-IIoT	**NB**	**DT**	**RF**	**XGB**	**CNN**	**RNN**
Baseline	**0.63**	**0.99**	**0.98**	**0.99**	**0.93**	**0.95**
DF [30]	**0.19**	**0.10**	**0.02**	0.18	**0.08**	**0.01**
FGSM [27]	0.46	0.21	0.22	**0.02**	0.66	0.42
BIM [28]	0.46	0.20	0.19	0.04	0.79	0.37
PGD [29]	0.46	0.20	0.19	0.04	0.79	0.37
JSMA [31]	0.24	0.58	0.85	0.44	0.65	0.43
EN [32]	0.48	0.51	0.23	0.26	0.50	0.19
ZOO [33]	0.32	0.32	0.65	0.72	0.26	0.32

TABLE V: MQTTset adversarial attack F_1 scores

MQTTset	**NB**	**DT**	**RF**	**XGB**	**CNN**	**RNN**
Baseline	**0.46**	**0.66**	**0.8**	**0.72**	**0.68**	**0.65**
DF [30]	**0.09**	0.19	0.20	**0.11**	**0.23**	**0.04**
FGSM [27]	0.34	0.17	**0.19**	0.14	0.31	0.24
BIM [28]	0.36	0.16	0.21	0.52	0.52	0.48
PGD [29]	0.36	**0.15**	0.21	0.52	0.52	0.48
JSMA [31]	0.39	0.56	0.70	0.62	0.57	0.53
EN [32]	0.34	0.29	0.26	0.30	0.38	0.21
ZOO [33]	0.18	0.29	0.35	0.31	0.34	0.23

MQTTset, with DF being the most effective attack in terms of lowering the models' F_1 scores. Another commonality is that the EN and JSMA attacks tend to have the least impact on the models' performance compared to other attacks. We can observe this in Fig. 4, where the EN and JSMA attacks result in the highest F_1 scores for the Decision Tree on the X-IIoTID (Fig. 4a), WUSTL-IIoT dataset (Fig. 4c) and the MQTTset (Fig. 4d). In contrast, in Fig. 5, the RNN model maintains relatively high F_1 scores under the JSMA attack compared to all other attacks on the WUSTL-IIoT (Fig. 5c) and the MQTTset (Fig. 5d).

Despite these commonalities, we observe a wide range of differences in how each model responds to specific adversarial attack across the datasets. One example is the Naive Bayes model's extreme vulnerability to the DF attack. Naive Bayes model consistently suffers the most severe performance degradation under the DF attack compared to the other models. For instance, in the X-IIoTID, the NB model's F_1 score plummets from a baseline of 0.62 to 0.04 under the DF attack, showing 16× performance loss. This pattern is repeated in the other datasets, with the Naive Bayes model experiencing F1 score drops of 11×, 5.5×, and 5.1× under the DF attack in the Edge-IIoTset, WUSTL-IIoT, and MQTTset, respectively. This suggests the Naive Bayes model's simplistic assumptions about feature independence make it particularly susceptible to the decision boundary-based perturbations generated by the DF attack. In contrast, the Random Forest model maintains relatively high F_1 scores under the JSMA attack compared to other models and attacks. For example, in the WUSTL-IIoT dataset (Table IV), the Random Forest Model achieves an F_1 score of 0.85 under the JSMA attack, which is only a 0.13 decrease from its baseline performance of 0.98. Similarly, in the X-IIoT dataset (Table II), the Random Forest model's F_1 score under the JSMA attack is 0.71, the highest among all models and attacks. This can be attributed to the ensemble model structure of the RF, which combines multiple decision trees to make predictions. As a result, it is less sensitive to the feature-based perturbations generated by the JSMA attack, as the ensemble can still make accurate predictions even if some individual trees are fooled. Another interesting observation is the varying impact of the Fast Gradient Sign Method (FGSM) attack on different models and datasets. In the WUSTL-IIoT dataset (Table IV), the FGSM attack severely degrades the performance of the XGBoost model, reducing its F_1 score from a baseline of 0.99 to 0.02. This represents a staggering 49.5× decrease in prediction performance, the largest among all models and datasets. However, in the same dataset, the CNN model exhibits remarkable resilience against the FGSM attack, maintaining an F_1 score of 0.66, which is only a 0.27 decrease from its baseline of 0.93.

3) Black-Box Attacks: Fig. 2 illustrates the effects of ZOO attack compared with other white-box attacks applied to the DNN models. The general trend for all the datasets is that as epsilon increases, the performance of the models decreases in terms of their macro F_1 score, which follows the same pattern we observed from white-box and gray-box attack results.

Deep Neural Network (DNN): Fig. 3 shows that the impact of ZOO attacks on DNN models varies with the dataset used. In the case of the X-IIoTID, the DNN model's average macro F_1 score under a ZOO attack is 0.11, ranking it as the second most destructive attack observed. A similar pattern emerges in the WUSTL-IIoT dataset, where the ZOO attack reduces the average macro F_1 score to 0.2, making it the most destructive attack encountered. In contrast, for the Edge-IIoTset and MQTTset, the ZOO attacks result in average macro F_1 scores of 0.27 and 0.18, respectively, positioning them among the least destructive attacks on these models.

Other Models: According to Table II, III, IV, and V, the effectiveness of the ZOO attack is comparable to other white-box attacks in terms of model performance degradation. Nevertheless, the effectiveness of ZOO attack is generally less than transfer (gray-box) attacks. This discrepancy likely stems from the ZOO attack's limited knowledge about the models' architectures and their training processes. One exception is observed in the MQTTset, where the ZOO attack significantly lowers the average macro F_1 score, making it one of the most detrimental attacks. Despite this, the ZOO attack does not match the potency of the most impactful adversarial attack, DeepFool, in terms of reducing the average macro F_1 score.

V. Conclusion

The unprecedented growth of IoT systems has introduced a new set of security challenges. While machine learning-based intrusion detection systems (ML-IDS) have demonstrated potential in identifying novel and evolving threats, their vulnerability to adversarial attacks calls into question their dependability. To address this pressing concern, we develop an all-encompassing evaluation framework that examines the performance of ML-IDS when confronted with numerous adversarial attacks across a spectrum of realistic and diverse IoT datasets. Our experiments show that adversarial attacks cause models to suffer sharp degradations in their F_1 score in comparison to their baseline performance. Specifically, XGBoost macro F_1 score, which achieved a baseline of 99%, decreases to 2% under the Fast Gradient Sign Method attack on the WUSTL-IIoT dataset, i.e., 49.5× prediction performance loss. This shows the flaws in current ML-IDS solutions, laying the foundation for future research aimed at fortifying IoT systems under ever-evolving adversarial threat.

References

[1] B. B. Zarpelão, R. S. Miani, C. T. Kawakani, and S. C. de Alvarenga, "A survey of intrusion detection in internet of things," *Journal of Network and Computer Applications*, vol. 84, pp. 25–37, 2017.

[2] "Iot value set to accelerate through 2030: Where and how to capture it." https://tinyurl.com/3xvs9dhc.

[3] N. Mishra and S. Pandya, "Internet of things applications, security challenges, attacks, intrusion detection, and future visions: A systematic review," *IEEE Access*, vol. 9, pp. 59353–59377, 2021.

[4] E. Anthi *et al.*, "Adversarial attacks on machine learning cybersecurity defences in industrial control systems," *Journal of Information Security and Applications*, vol. 58, p. 102717, 2021.

[5] H. Liu and B. Lang, "Machine learning and deep learning methods for intrusion detection systems: A survey," *applied sciences*, vol. 9, no. 20, p. 4396, 2019.

[6] O. Gungor, T. Rosing, and B. Aksanli, "Hd-i-iot: Hyperdimensional computing for resilient industrial internet of things analytics," in *2023 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 1–6, IEEE, 2023.

[7] A. Thakkar and R. Lohiya, "A review on challenges and future research directions for machine learning-based intrusion detection system," *Archives of Computational Methods in Engineering*, vol. 30, no. 7, pp. 4245–4269, 2023.

[8] K. He, D. D. Kim, and M. R. Asghar, "Adversarial machine learning for network intrusion detection systems: A comprehensive survey," *IEEE Communications Surveys Tutorials*, vol. 25, no. 1, pp. 538–566, 2023.

[9] A. Alotaibi and M. A. Rassam, "Adversarial machine learning attacks against intrusion detection systems: A survey on strategies and defense," *Future Internet*, vol. 15, no. 2, p. 62, 2023.

[10] H. Khazane, M. Ridouani, F. Salahdine, and N. Kaabouch, "A holistic review of machine learning adversarial attacks in iot networks," *Future Internet*, vol. 16, no. 1, p. 32, 2024.

[11] I. Vaccari, G. Chiola, M. Aiello, M. Mongelli, and E. Cambiaso, "Mqttset, a new dataset for machine learning techniques on mqtt," *Sensors*, vol. 20, no. 22, p. 6578, 2020.

[12] M. Zolanvari *et al.*, "Wustl-iiot-2021 dataset for iiot cybersecurity research," *Washington University in St. Louis, USA*, 2021.

[13] M. Al-Hawawreh *et al.*, "X-iiotid: A connectivity-agnostic and device-agnostic intrusion data set for industrial internet of things," *IEEE Internet of Things Journal*, vol. 9, no. 5, pp. 3962–3977, 2021.

[14] M. A. Ferrag *et al.*, "Edge-iiotset: A new comprehensive realistic cyber security dataset of iot and iiot applications for centralized and federated learning," *IEEE Access*, vol. 10, pp. 40281–40306, 2022.

[15] D. E. Kouicem *et al.*, "Internet of things security: A top-down survey," *Computer Networks*, vol. 141, pp. 199–221, 2018.

[16] "Monthly number of internet of things (iot) malware attacks worldwide from 2020 to 2022." https://tinyurl.com/33w2m97n.

[17] M. Antonakakis, T. April, M. Bailey, M. Bernhard, E. Bursztein, J. Cochran, Z. Durumeric, J. A. Halderman, L. Invernizzi, M. Kallitsis, *et al.*, "Understanding the mirai botnet," in *26th {USENIX} security symposium ({USENIX} Security 17)*, pp. 1093–1110, 2017.

[18] E. Bout, V. Loscri, and A. Gallais, "How machine learning changes the nature of cyberattacks on iot networks: A survey," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 1, pp. 248–279, 2021.

[19] N. Chaabouni *et al.*, "Network intrusion detection for iot security based on learning techniques," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2671–2701, 2019.

[20] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of intrusion detection systems: techniques, datasets and challenges," *Cybersecurity*, vol. 2, no. 1, pp. 1–22, 2019.

[21] O. Gungor, T. Rosing, and B. Aksanli, "Adversarial-hd: Hyperdimensional computing adversarial attack design for secure industrial internet of things," in *Proceedings of Cyber-Physical Systems and Internet of Things Week 2023*, pp. 1–6, 2023.

[22] O. Gungor, T. Rosing, and B. Aksanli, "Stewart: Stacking ensemble for white-box adversarial attacks towards more resilient data-driven predictive maintenance," *Computers in Industry*, vol. 140, p. 103660, 2022.

[23] H. Qiu, T. Dong, T. Zhang, J. Lu, G. Memmi, and M. Qiu, "Adversarial attacks against network intrusion detection in iot systems," *IEEE Internet of Things Journal*, vol. 8, no. 13, pp. 10327–10335, 2020.

[24] M. Macas, C. Wu, and W. Fuertes, "Adversarial examples: A survey of attacks and defenses in deep learning-enabled cybersecurity systems," *Expert Systems with Applications*, p. 122223, 2023.

[25] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow.* " O'Reilly Media, Inc.", 2022.

[26] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "A survey on adversarial attacks and defences," *CAAI Transactions on Intelligence Technology*, vol. 6, no. 1, pp. 25–45, 2021.

[27] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[28] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*, pp. 99–112, Chapman and Hall/CRC, 2018.

[29] A. Madry *et al.*, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[30] S.-M. Moosavi-Dezfooli *et al.*, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582, 2016.

[31] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE European symposium on security and privacy (EuroS&P)*, pp. 372–387, IEEE, 2016.

[32] P.-Y. Chen *et al.*, "Ead: elastic-net attacks to deep neural networks via adversarial examples," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[33] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "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*, pp. 15–26, 2017.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[35] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[36] M.-I. Nicolae, M. Sinn, M. N. Tran, B. Buesser, A. Rawat, M. Wistuba, V. Zantedeschi, N. Baracaldo, B. Chen, H. Ludwig, *et al.*, "Adversarial robustness toolbox v1. 0.0," *arXiv preprint arXiv:1807.01069*, 2018.