Adversarial robustness of ML-based intrusion detection systems

Dmitry Sivkov
SIT departments
ITMO University
St.Petersburg, Russian Federation
sivkov@itmo.ru

Viktoriia Korzhuk
SIT departments
ITMO University
St.Petersburg, Russian Federation
vmkorzhuk@itmo.ru

Roman Safiullin
SIT departments
ITMO University
St.Petersburg, Russian Federation
romsaaf@mail.ru

Omar Farshad Jeelani
SIT departments
ITMO University
St.Petersburg, Russian Federation
omar@itmo.ru

Alisa Vorobeva
SIT departments
ITMO University
St.Petersburg, Russian Federation
vorobeva@itmo.ru

2026 14th International Symposium on Digital Forensics and Security (ISDFS) | 979-8-3315-7310-2/26/$31.00 ©2026 IEEE | DOI: 10.1109/ISDFS69419.2026.11459025

***Abstract*—Machine-learning intrusion detection systems (IDS) are increasingly deployed, yet their robustness to adversarial manipulation of flow features is less studied for gradient-boosting models. This paper evaluates a CatBoost-based IDS trained on the CSE-CIC-IDS2018 dataset under four white-box evasion attacks (FGSM, BIM, PGD, and Carlini–Wagner). We report accuracy and precision/recall/F1 to quantify degradation and discuss practical constraints on feature-space perturbations. On CSE-CIC-IDS2018, accuracy drops from 94.86% (clean) to 63.53% (FGSM), 63.51% (BIM), 59.67% (PGD), and 3.53% (CW), revealing severe robustness gaps.**

Keywords—IDS, Adversarial attacks, Robustness, ML, Anomaly-based IDS, ML in cyber security

I. Introduction

Intrusion detection systems (IDS) are critical components of network security architectures. Their role is to monitor network traffic for signs of malicious activity and alert the system administrators about potential security breaches. In recent years, there has been a significant shift towards using machine learning and deep learning-based IDS to detect and classify network attacks. These systems have shown promising results in improving the accuracy and efficiency of intrusion detection.

However, machine learning-based IDS are vulnerable to adversarial attacks, which can be used to evade detection by the IDS. Similar concerns regarding the robustness and security of ML models have been raised in other safety-critical domains, including AI-based healthcare systems, where ensuring data privacy and integrity is equally essential [1], [2]. Adversarial attacks are a class of attacks that exploit vulnerabilities in machine learning models modifying the input data in a specific way, so it causes the model to misclassify the data.

Key issues at the heart of this research include the susceptibility of ML-based IDS to evasion attacks. Evasion attacks aim to mislead the system by subtly altering input data, allowing malicious traffic to go undetected. Due to their simplicity, adversarial attacks pose significant challenges to the robustness and reliability of ML-based IDS.

Adversarial attacks against IDSs are particularly challenging due to the complex nature of network traffic data and the large number of features that need to be analyzed. Attackers can use sophisticated techniques to generate adversarial examples that are tailored to bypass specific IDSs or exploit weaknesses in their machine learning algorithms.

Adversarial attacks against deep learning models have attracted significant research attention in recent years. Researchers have shown that deep learning models, including those used for intrusion detection, can be fooled by small and imperceptible perturbations to input data. Adversarial attacks can have a devastating impact on IDS, as they can cause false positives or false negatives, leading to failure in detecting or flagging an intrusion.

The research investigates the impact of adversarial attacks on the ML-based IDS model. Four adversarial attack generation methods were used to generate adversarial samples against the IDS model: Fast Gradient Sign Method (FGSM), Carlini and Wagner (CW), Basic Iterative Method (BIM) and Projected Gradient Descent (PGD). The CSE-CIC-IDS2018 dataset [3] was used for training and testing the models.

A widely adopted benchmark dataset for evaluating Intrusion Detection System (IDS) models is the CSE-CIC-IDS2018 dataset [3], derived from the collaborative project between the Communications Security Establishment (CSE) and the Canadian Institute for Cybersecurity (CIC). It was created by capturing all network traffic during ten days of operation inside a controlled network environment on AWS where realistic background traffic and different attack scenarios were conducted. As a result, the dataset contains both benign network traffic as well as captures of the most common network attacks. The dataset is comprised of the raw network captures in PCAP format as well as csv files containing 80 statistical features of the individual network flows combined with their corresponding labels. The dataset comprises various attack types such as Denial of Service (DoS), FTP and SSH Brute Force, different Web attacks based on Damn Vulnerable Web Application (DVWA), Botnet attacks, Port Scans and various Infiltration attacks. [4]

An ML-IDS model was built based on Christoph Stumpf's "Machine learning based approach towards building an Intrusion Detection System" [4]. The result of that part of work is a classifier capable of categorising network flows as either benign or malicious.

979-8-3315-7310-2/26/$31.00 ©2026 IEEE

This research investigates the susceptibility of gradient boosting-based intrusion detection systems to gradient-based adversarial attacks. Specifically, we evaluate a CatBoost-based ML-IDS model against four widely adopted white-box attack methods and systematically analyze their impact on detection performance. The study provides a comparative assessment of attack effectiveness and examines the implications of adversarial perturbations within the network flow feature space. Based on the obtained results, we discuss potential defense mechanisms grounded in adversarial learning.

The adversarial evaluation assumes a white-box setting in which the attacker has full knowledge of the trained model, including its architecture and parameters, and can generate adversarial perturbations in feature space. The attacker is assumed to control traffic generation but does not modify the feature extraction pipeline. This setting represents a worst-case robustness assessment and provides an upper bound on potential vulnerability.

Research gap: Existing adversarial evaluations of intrusion detection systems largely focus on deep neural network IDS and often emphasize attack success rates, while systematic evidence for widely used gradient-boosting IDS and the operational feasibility of feature-space perturbations in flow-based settings remains limited.

Contributions of this paper are:

- A robustness evaluation of a CatBoost-based IDS on CSE-CIC-IDS2018 under FGSM, BIM, PGD, and CW attacks implemented with Foolbox.
- Reporting complementary metrics (accuracy, precision, recall, and F1-score) on clean and adversarial samples to compare attack severity.
- A feature-level perturbation analysis perspective and a discussion of operational constraints that distinguish mathematically effective from feasible perturbations in real networks.

The rest of the paper is organized as follows. Section II provides a detailed overview of related work on adversarial attacks against IDS models and adversarial training as a defense mechanism. Section III presents the methodology used in our research, including the dataset, model, and adversarial attack generation methods. Section IV presents the results of our experiments and discusses the impact of adversarial attacks on the models' performance. Finally, Section V concludes the paper, proposes the protection method and provides future research directions.

II. Literature Review

ML techniques have been popular in recent years for anomaly detection. The research field of adversarial machine learning is the focus of a paper by Jmila, H., and Khedher, M. I. [5]. In the context of network IDS, they also investigate the robustness of many frequently used ML classifiers against hostile instances. This study takes into account both gray-box and white-box attacks. White-box based adversarial examples were produced using an external classifier built on a DNN. To increase the robustness of various NIDS, the influence of a defense strategy based on Gaussian data augmentation is also investigated. Gaussian Data Augmentation: The defense strategy based on Gaussian data augmentation was investigated in the study by Jmila and Khedher [5]. This technique involves augmenting the training data with Gaussian noise, which aims to increase the robustness of various NIDS against adversarial attacks.

In the development of Network Intrusion Detection Systems (NIDS), neural networks (NNs) are becoming more and more common, although they can be subject to adversarial examples. TIKI- TAKA, a general framework for evaluating the robustness of cutting-edge deep learning-based NIDS against adversarial manipulations, was introduced by Zhang, C. Jiang et al. in [6]. It also incorporates defense mechanisms to improve the models' ability to resist attacks using such evasion techniques. The volume of data moving through systems and the increasing growth of DNN usage presents a significant problem for stopping adversarial assaults on DNN. The goal of the research by Matrawy A et al. [7] is to investigate the efficacy of various evasion attacks and the training of a resilient deep learning-based IDS utilizing various Neural networks.

By using neuron activations during test time to detect adversarial attacks made using four well-known evasion attack algorithms — Fast Gradient Sign, Basic Iterative Method, Carlini and Wagner attack, and Projected Gradient Descent — Pawlicki, M., Chora, M., Kozik, R.'s work [8] makes a significant contribution to cybersecurity. Sadly, recent work on adversarial machine learning has shown that deep learning models are inherently vulnerable to hostile changes made to their input data. In the work of Clements J., Yang Y., Sharma A., Hu, H., Lao Y. [9], they examine the possibility for adversarial actors to breach such vulnerabilities to compromise deep learning-based NIDS systems. The study shows, for example, that an attacker can create deceptive inputs that successfully mislead a target deep learning-based NIDS by changing as few as 1.38 on average of the input properties of an observed packet.

In order to create adversarial perturbed traffic records that attack intrusion detection systems by fooling and escaping detection, a framework of generative adversarial networks known as IDSGAN is suggested in [10]. In this study, NSL-KDD, an improved version of KDD'99, is used as a benchmark dataset to assess IDS. Adversarial attacks are discussed by Ren, K. et al. [11] in a broader context that goes beyond intrusion detection, in which the effects of adversarial attacks on image processing, natural language processing, game theory, and other fields were thoroughly discussed.

A novel attack framework, AIDAE, is presented in [12] to produce features to deactivate IDSs. The setup's multi-channel decoders are divided into continuous and discrete channels to produce, respectively, continuous and discrete features. Furthermore, using the same properly trained encoder, the produced features can maintain the correlation between continuous and discrete components. As a result, the created features can mimic the distribution of original normal features.

The study by Pawlicki, M., Chora, M., and R. Kozik [8] assesses the potential for degrading the performance of an intrusion detection algorithm that has been effectively optimized at test time by creating adversarial attacks using the four recently proposed approaches and then provides a method to identify those attacks. Those approaches being

Attack by Carlini and Wagner (CW), Projected Gradient Descent (PGD), Basic Iterative Method (BIM), and Fast Gradient Sign Method (FGM). The authors collected the test time neural activations of an ANN trained on a portion of the CICIDS2017 dataset as well as the test time neural activations of adversarial instances developed for this ANN. They trained and tested five different ML classifiers to find adversarial cases using these activations, and they were able to get a recall of 0.99.

III. Methodology and Experimental Setup

The entire research set-up was seperated into two parts. The first part of the research is aimed at building the ML-based IDS model focused on detecting malicious traffic from the CSE-CIC-IDS2018 dataset [3]. The accuracy of the model will later be used for comparison with the model's performance after adversarial attacks to check the drop in the performance. The second part of the research is focused on evaluating the resilience of the trained models against different adversarial attacks.

A. Building the Model

The computational resources used to train the classifier are given in Table I.

TABLE I. TABLE I. COMPUTATIONAL RESOURCES

Category	Resource
CPU	Intel Core i5 9600KF, 4.1 GHz
RAM	32 GB
GPU	Nvidia GeForce RTX 2070S, 8 GB RAM
SSD	100 GB

The machine learning estimator created in this project is based on Christoph Stumpf's [4], follows a supervised approach and is trained using the Gradient Boosting algorithm. Employing the CatBoost library a binary classifier is created, capable of classifying network flows as either benign or malicious. The performance metrics of the classifier are presented in Table II.

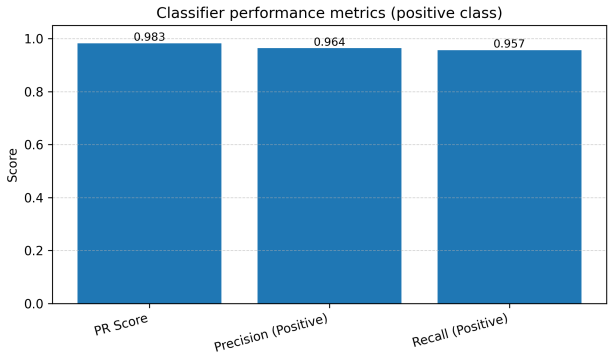

Fig. 1. Classifier performance metrics for the positive class

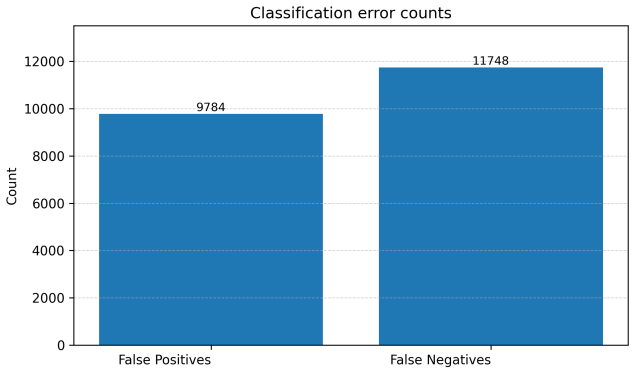

Fig. 2. False positives and false negatives observed on the evaluation set

TABLE II. CLASSIFIER PERFORMANCE METRICS

PR Score	Precision Positive	Recall Positive	False-Positives	False-Negatives
0.98266	0.964	0.957	9784	11748

The optimal parameter configuration obtained via the best performing model is as follows:

- usage of 1900 trees,
- a maximum depth of 10 per tree,
- a L2 regularization coefficient of 4.8139,
- a border count of 254 and,
- a random strength parameter of 5.

The data used to train the classifier is taken from the CSE-CIC-IDS2018 dataset [3] provided by the Canadian Institute for Cybersecurity. It was created by capturing all network traffic during ten days of operation inside a controlled network environment on AWS where realistic background traffic and different attack scenarios were conducted. As a result, the dataset contains both benign network traffic as well as captures of the most common network attacks. The dataset is comprised of the raw network captures in PCAP format as well as csv files created by using CICFlowMeter-V3 containing 80 statistical features of the individual network flows combined with their corresponding labels. A network flow is defined as an aggregation of interrelated network packets identified by the following properties:

- Source IP,
- Destination IP,
- Source port,
- Destination port,
- Protocol

The dataset contains approximately 16 million individual network flows and covers the following attack scenarios:

- Brute Force,
- DoS,
- DDos,
- Heartbleed,
- Web Attack,
- Infiltration,
- Botnet

Reproducibility summary: We use the pre-extracted CICFlowMeter feature set (80 numerical flow statistics). To ensure repeatability, we run all experiments with a fixed random seed and evaluate on a held-out evaluation subset. The CatBoost training configuration is summarized above. Adversarial samples are generated in a white-box setting using the Foolbox library, and the corresponding attack hyperparameters are reported in Table III.

B. Attacking the Model

The chosen methods for generating adversarial samples were implemented using Python's FoolBox library [13] [14]. All of the methods are examples of white-box attacks since they show the best performance and in case of real usage - an adversarial is likely to implement the attack close to one with white-box characteristics, since the datasets for implementing ML-based IDSs are public and the best-performing classifiers are also well-researched and known as an industrial standard.

The first attack - Fast Gradient Sign Method (FGSM), is one of the earliest methods of adversarial attacks. It was introduced in the paper Explaining and Harnessing Adversarial Examples [15], and has gained a lot of traction since. The idea behind FGSM is very simple: do the opposite of the gradient descent method in order not to minimize but to maximize the loss. This is the crux of the Fast Gradient Sign Method: use the sign of the gradient, multiply it by some small value, and add that perturbation to the original input to create an adversarial example.

Basic Iterative Method (BIM) [16] is a simple extension of the FGSM method. BIM is based on the iterative use of FGSM on the target model, generating the sample with the highest effect on accuracy so that the total perturbation does not exceed the specified range of values. BIM in comparison to FGSM takes an iterative approach by applying FGSM multiple times instead of being applied in a single large step.

Projected Gradient Descent Method (PGD) was introduced by Madry et al. in 2017 [17]. PGD operates by starting from a random point within the allowed perturbation region of the original input. It then performs gradient descent on the model's loss function, projecting the perturbed input back into the feasible set after each iteration. This iterative process ensures that the adversarial perturbation remains within the defined constraints around the original input, making the adversarial example both effective and imperceptibly different to human observers.

The Carlini and Wagner method [18] is an iterative method for generating adversarial samples while ensuring the perturbations are minimal and imperceptible. Basically, it generates the adversarial instance by finding the smallest noise added to an image that will change the classification to a class t. In the paper, the noise level is measured in terms of L2 distance.

The parameters for these attacks were chosen experimentally to show the best combination of performance and effect on the model's accuracy. The chosen parameters are presented in Table III.

TABLE III. CHOSEN ATTACK PARAMETERS

Attack	Stepsize	Steps	Random Start
FGSM	1.0	1	False
Basic Iterative Method	0.2	10	False
Carlini and Wagner	0.01	1000	-
Projected Gradient Descent Method	0,0333	40	True

C. Feature Perturbation Analysis

While the degradation in classification accuracy demonstrates the effectiveness of adversarial attacks, it does not explain how perturbations alter the network flow feature space to induce misclassification. To better understand the structural impact of adversarial manipulation, we introduce a feature-level perturbation analysis framework.

Let $x \in \mathbb{R}^{80}$ denote the original network flow feature vector and x_{adv} its adversarial counterpart.

The perturbation vector is defined as:

$$\delta = x_{adv} - x \tag{1}$$

1) Feature Change Frequency. For each feature j, the proportion of samples with perturbation magnitude exceeding a small threshold τ is computed as:

$$f_j = \left(\frac{1}{N}\right) \Sigma I\left(\left|\delta_{i,j}\right| > \tau\right) \tag{2}$$

2) Perturbation Magnitude. To characterize the global strength of adversarial modifications, we consider both the L2 and L∞ norms:

$$\left|\left|\delta\right|\right|_2 \tag{3}$$

$$\left|\left|\delta\right|\right|_\infty \tag{4}$$

The L∞ norm captures the largest single-feature deviation, whereas the L2 norm reflects the cumulative distribution of perturbations across the feature space.

3) Feature Group Interpretation. Since the CSE-CIC-IDS2018 dataset consists of 80 statistical flow-based features extracted via CICFlowMeter, perturbations can be analyzed in terms of feature groups rather than isolated variables. These typically include:

- Packet length statistics
- Inter-arrival time (IAT) statistics
- Flow duration
- Rate-based aggregates (bytes/s, packets/s)
- Active/idle timing features
- Transport-layer metadata

Iterative attacks such as BIM and PGD are expected to distribute perturbations across multiple statistical descriptors, while optimization-based attacks such as Carlini–Wagner typically concentrate modifications on a smaller subset of influential features. The significant degradation observed under CW suggests that the classifier may rely heavily on a

limited set of high-importance statistical features, making it vulnerable to targeted manipulation.

4) Realistic Feasibility Considerations. Although adversarial perturbations are computed in feature space, not all flow statistics can be independently manipulated in real network environments. Features derived from timing and traffic volume may be indirectly influenced by traffic shaping or packet injection strategies, whereas flow identifiers (e.g., protocol or endpoint metadata) are structurally constrained. Therefore, evaluating adversarial robustness in IDS contexts requires distinguishing between mathematically effective and operationally feasible perturbations.

This analysis framework provides insight into the structural weaknesses of ML-based IDS models beyond aggregate accuracy metrics and supports the development of constraint-aware defense mechanisms.

D. Operational Constraints

The evaluated adversarial attacks operate in unconstrained feature space, directly perturbing numerical flow descriptors. However, in real-world IDS deployments, many flow-based features are derived from structured network traffic and cannot be arbitrarily modified.
In particular:

- Structural constraints: protocol-related and endpoint metadata cannot be independently altered without changing session semantics.
- Statistical dependency constraints: many flow features (e.g., packet statistics and rate-based aggregates) are functionally interdependent.
- Physical feasibility constraints: timing and rate features must remain consistent with realizable traffic behavior.

Therefore, the presented results represent an upper bound on model vulnerability. Future work should incorporate constraint-aware adversarial generation to approximate realistic attacker capabilities.

IV. Experimental Results

Using four attack methods—FGSM, BIM, Carlini–Wagner (CW), and PGD—we generated adversarial samples for the proposed CatBoost-based IDS by applying bounded perturbations to the input feature vectors in a white-box setting. Fig. 3 reports the model accuracy under each attack.

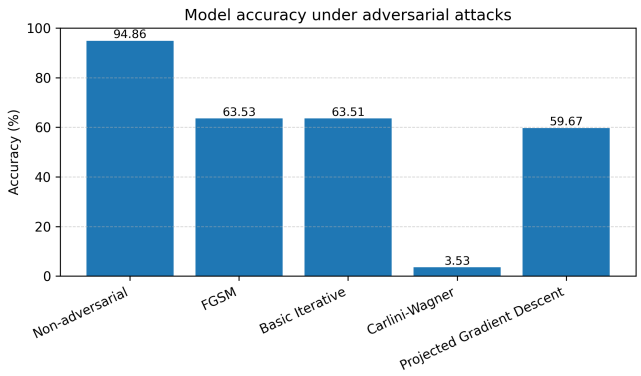

Fig. 3. Model accuracy under adversarial attacks (Table IV).

Next, we evaluated the model performance on clean and adversarial samples. Table IV summarizes accuracy, precision, recall, and F1-score for each attack. Compared to the clean baseline (accuracy 94.86%, F1-score 0.960), all attacks substantially degrade detection performance. The CW attack causes near-complete evasion (accuracy 3.53%, F1-score 0.016), while FGSM and BIM reduce F1-score to 0.678 and 0.673, respectively. PGD yields an F1-score of 0.641.

TABLE IV. MODEL PERFORMANCE ON ADVERSARIAL SAMPLES

Attack	Accuracy, %	Precision	Recall	F1-Score
Non-adversarial	94.86	0.964	0.957	0.960
FGSM	63.53	0.800	0.588	0.678
BIM	63.51	0.810	0.575	0.673
CW	3.53	0.024	0.012	0.016
PGD	59.67	0.763	0.553	0.641

Beyond accuracy, precision and recall highlight asymmetric failure modes under adversarial manipulation. For FGSM and BIM, precision remains relatively high (0.800–0.810), whereas recall drops to 0.575–0.588, indicating that the classifier increasingly misses malicious traffic while retaining moderate confidence on the positives it still predicts. PGD follows a similar pattern (precision 0.763, recall 0.553). In contrast, CW collapses both precision (0.024) and recall (0.012), confirming a severe degradation consistent with almost total evasion.

While aggregate metrics quantify the degradation, they do not explain how perturbations influence the feature space. Iterative attacks such as BIM and PGD are expected to distribute perturbations across multiple statistical descriptors, gradually shifting samples toward decision boundaries. In contrast, optimization-based attacks such as CW often concentrate modifications on a smaller subset of influential features, seeking minimal yet strategically effective changes.

The extreme degradation under CW suggests structural sensitivity of the classifier to targeted manipulation of high-importance statistical features. In flow-based intrusion detection, such features typically include rate-based aggregates, packet-length statistics, and timing descriptors. These observations motivate feature-level perturbation analysis in addition to performance-based robustness evaluation.

Finally, while adversarial training may improve resistance to the attack used during augmentation, potential trade-offs include increased computational cost and reduced clean-data accuracy. For gradient boosting-based IDS models, this direction remains an open area for empirical validation.

V. Conclusion and Further Research

We evaluated the adversarial robustness of a CatBoost-based intrusion detection model on the CSE-CIC-IDS2018 dataset. All tested white-box attacks substantially degraded detection performance, with Carlini–Wagner causing near-complete evasion (accuracy 3.53%, precision 0.024, recall 0.012, F1-score 0.016). These findings highlight that flow-based ML IDS can be highly vulnerable to feature-space

manipulation and that robustness assessment should consider both performance metrics and operational feasibility constraints.

- Constraint-aware robustness evaluation using perturbations that reflect realistic attacker capabilities.
- Feature-level perturbation analysis (change frequency and L2/L∞ norms) to identify the most sensitive flow statistics.
- Empirical validation of defenses for gradient-boosting IDS, including adversarial training and input preprocessing.

The next step of that concrete research is to provide the protection method against model evasion attacks. The main field of study for now is adversarial learning and adversarial training that could perform great not only against specifically crafted malicious input but also against some previously unknown/unseen malicious traffic.

In summary, future research in this field should focus on developing advanced attack methods, evaluating defense techniques, exploring attack transferability, using comprehensive evaluation metrics, analyzing the impact on network topologies, and investigating the broader impacts on machine learning-based security systems.

VI. Acknowledgment

The authors gratefully acknowledge the financial support provided by CentryNex Cybersecurity Limited, Paramount House, 1 Delta Way, Egham, Surrey TW20 8RX, United Kingdom. The authors also appreciate the institutional support that contributed to the successful completion of this research.

References

[1] O. F. Jeelani, M. Njie, and V. M. Korzhuk, "Enhancing data privacy and data security across healthcare," 2024.

[2] O. F. Jeelani, M. Njie, and V. M. Korzhuk, "Methods and algorithms of ensuring data privacy in AI-based healthcare systems and technologies," in Conf. Proc., Paris, France, vol. 11, Apr. 2024, p. 12.

[3] Canadian Institute for Cybersecurity, "CSE-CIC-IDS2018 Dataset," 2018. [Online]. Available: https://www.unb.ca/cic/datasets/ids-2018.html. Accessed: Dec. 14, 2023.

[4] C. Stumpf, "A machine learning based approach towards building an intrusion detection system," GitHub repository, 2019. [Online]. Available: https://github.com/cstub/ml-ids. Accessed: Jan. 14, 2024.

[5] M. I. Khedher and H. Jmila, "Adversarial machine learning for network intrusion detection: A comparative study," <Elsevier journal>, vol. 214, Art. no. 109073, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1389128622002146

[6] M. Zhang, C. Jiang, and M. Kamel, "Intrusion detection using hierarchical neural networks," <journal title>, vol. 26, no. 6, pp. 779–791, 2005. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9063416/

[7] A. Matrawy and R. Abou Khamis, "Evaluation of adversarial training on different types of neural networks in deep learning-based IDSs," in Proc. <IEEE conference/venue>, pp. 1–6, 2020. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9297344/

[8] R. Kozik, M. Pawlicki, and M. Choraś, "Defending network intrusion detection systems against adversarial evasion attacks," <Elsevier journal>, vol. 110, pp. 148–154, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X20303368

[9] J. Clements et al., "Rallying adversarial techniques against deep learning for network security," in Proc. <IEEE conference/venue>, pp. 1–8, 2021. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9660011/

[10] Z. Xue, Z. Lin, and Y. Shi, "IDSGAN: Generative adversarial networks for attack generation against intrusion detection," in <Springer book/proceedings title>, Springer, 2022, pp. 79–91. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-05981-0_7

[11] K. Ren et al., "Adversarial attacks and defenses in deep learning," <Elsevier journal>, vol. 6, no. 3, pp. 346–360, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S209580991930503X

[12] J. Chen et al., "Fooling intrusion detection systems using adversarial autoencoder," <Elsevier journal>, vol. 7, no. 3, pp. 453–460, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352864820302868

[13] J. Rauber, W. Brendel, and M. Bethge, "Foolbox: Fast adversarial attacks to benchmark the robustness of machine learning models in PyTorch, TensorFlow, and JAX," GitHub repository, 2017. [Online]. Available: https://github.com/bethgelab/foolbox/tree/master. Accessed: Jan. 14, 2024.

[14] J. Rauber, R. Zimmermann, M. Bethge, and W. Brendel, "Foolbox Native: Fast adversarial attacks to benchmark the robustness of machine learning models in PyTorch, TensorFlow, and JAX," J. Open Source Softw., vol. 5, no. 53, p. 2607, 2020, doi: 10.21105/joss.02607.

[15] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," arXiv preprint arXiv:1412.6572 [stat.ML], 2015.

[16] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," arXiv preprint arXiv:1607.02533 [cs.CV], 2017.

[17] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," arXiv preprint arXiv:1706.06083 [stat.ML], 2019.

[18] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," arXiv preprint arXiv:1608.04644 [cs.CR], 2017.