Advances in Adversarial Attacks and Defenses in Intrusion Detection System: A Survey

Mariama Mbow[1(✉)], Kouichi Sakurai[1], and Hiroshi Koide[2]

[1] Department of Informatics Graduate School of Information Science and Electrical Engineering, Kyushu University, Fukuoka, Japan
mbow.mariama.076@s.kyushu-u.ac.jp, sakurai@inf.kyushu-u.ac.jp
[2] Research Institute for Information Technology Cyber Security Center, Kyushu University, Fukuoka, Japan
koide@cc.kyushu-u.ac.jp

Abstract. Machine learning is one of the predominant methods used in computer science and has been widely and successfully applied in many areas such as computer vision, pattern recognition, natural language processing, cyber security etc. In cyber security, the application of machine learning algorithms for network intrusion detection system (NIDS) has seen promising results for anomaly detection mostly with the adoption of deep learning and is still growing. However, machine learning algorithms are vulnerable to adversarial attacks resulting in significant performance degradation. Adversarial attacks are security threats that aim to deceive the learning algorithm by manipulating its predictions, and Adversarial machine learning is a research area that studies both the generation and defense of such attacks. Researchers have extensively worked on the adversarial machine learning in computer vision but not many works in Intrusion detection system. However, failure in this critical Intrusion detection area could compromise the security of an entire system, and need much attention. This paper provides a review of the advancement in adversarial machine learning based intrusion detection and explores the various defense techniques applied against. Finally discuss their limitations for future research direction in this emerging area.

Keywords: Adversarial attack · Cyber security · Intrusion detection · Machine learning · Deep learning · Poisoning attack · Evasion attack

Network Intrusion Detection System (NIDS) [1,2] are playing an important role in cybersecurity for detecting malicious network traffic. NIDS uses signature or anomaly based detection to identify cyber-attacks. However with the growth of network traffic and attacks diversity [3], signature detection which can only detect existing attacks by using their signatures are being replaced by anomaly based detection which have potentially the capabilities to detect existing attacks as well as novel attacks. Among the various techniques applied to implement

C. Su and K. Sakurai (Eds.): SciSec 2022 Workshops, CCIS 1680, pp. 196–212, 2022.
https://doi.org/10.1007/978-981-19-7769-5_15

NIDS, Machine Learning (ML) based have been the predominant method and have seen a fast adoption due to their abilities to discriminate between abnormal and normal pattern over a data set. In the last decades there has been a wide research that apply machine learning (including deep learning) in NIDS settings [4,23,24]. However, serious security issues are now emerging with the discovery of the vulnerability of these algorithms [8].

Researchers [5–8] have shown that machine learning can be easily fooled when adding some perturbation during its training or prediction phase. These perturbations are called adversarial samples and they are specially crafted inputs that cause the learning model to wrongly classify/predict an input. For instance attackers can exploit the vulnerability of voice control system and influence the model to make wrong decision on recognizing voice command. In autonomous vehicles based machine learning the attacker can trick the model to make wrong decision on recognizing the traffic signs [10]. In intrusion detection the attacker might influence the classifiers to misclassify the attack traffic as benign and then bypass the security system. Failure in this critical cybersecurity area could compromise the security of an entire system. Then it is actually the security-critical area that face the biggest challenges from these threats [11].

Considering the limited reviews targeting the adversarial attacks against network intrusion detection system and the numerous papers being published recently, in this survey we aims to provide a comprehensive overview of the evolution of the works provided in this area with the following contributions:

1. We summarize and analyze the recent advance on adversarial machine learning applied to NIDS
2. By analyzing and comparing the different works proposed, we discuss open issues that can help as future direction in this evolving area.

The remainder of this paper is organized as follow: In Sect. 1 we discuss previous related works. Section 2 we discuss the background of basic concept on machine learning and adversarial attack taxonomy. In Sect. 3 we discuss the adversarial attack applied in NIDS. Section 4 discusses the adversarial defenses. Finally we propose some future direction in Sect. 5 and conclude in Sect. 6.

1 Related Work

Related works have been presented in [11,12]. In [12], authors worked on a review of adversarial machine learning in intrusion and malware detection. However they provided limited review on researches related to NIDS and mainly focused on the evasion attack and in white box scenario. Moisejevs et al. [11] provided an overview of adversarial attacks and defenses in intrusion detection. They attempted to focus on evasion and poisoning attacks in white box and black box scenario. However similar to [12], limited papers were reviewed and the most recent was in 2018. Recently there has been an increasing number of publications in adversarial machine learning [13] including applied in NIDS. The literature survey we provide differ from the previous in many ways. We include the more

recent works. In addition we review all adversarial machine learning scenario in NIDS including black box and white box and applied during training time (poisoning attack) or during test time (evasion attack). More details of these techniques will be discussed in Sect. 2.

To prepare this survey, studies are selected on multiple databases such as Springer and Elsevier, IEEE, Research Gate and Science Direct using keywords "Intrusion detection", "Adversarial Machine learning", "Adversarial Deep Learning". We survey a total of 29 papers that works on adversarial attacks or defense technique.

2 Background

In this section, we discuss the basic concept of machine learning and adversarial attack.

2.1 Machine Learning in NIDS

Machine learning is a part of artificial intelligence (AI) with a multidisciplinary research area that spans several fields. These fields include probability and statistics, computer science, algorithms, psychology and brain science. There are four (4) approaches used in Machine learning such as Supervised learning, Unsupervised learning, Semi-supervised learning, and Reinforcement learning. However Supervised and unsupervised learning are the most common type used in NIDS [14]. Machine learning models are mainly divided into shallow or traditional model and deep learning model. The most common traditional ML models applied in IDS include support vector machine (SVM), decision tree (DT), random forest (RF), k-means, artificial neural network (ANN), and ensemble method [1,14]. Recently Deep learning (DL) methods have greatly improved NIDS by overcoming the difficulty of feature selection and representation. The number of published works on DL based NIDS has rapidly increased [4,14]. The common DL models applies in NIDS include recurrent neural networks (RNNs), long short-term memory (LSTM) networks, convolutional neural network (CNN), AutoEncoder (AE), Deep neural network (DNN), Deep belief network (DBN).

2.2 Adversarial Machine Learning

Adversarial attacks represent a major limitation for the adoption of machine learning in many area. These attacks against the machine learning algorithms are security threats that aim to trick the learning model by purposely adding tiny perturbations to the data to easily subvert their predictions. This phenomenon has been explored for more than a decade in the traditional machine learning [25]. However the discovery of these adversarial examples against neural networks, by Szegedy et al. [8] and in subsequently [5,26,27], has renewed interest in the AI community [25].

These perturbations against the learning algorithms can be performed in mainly during the training time or test time. In the training time also called poisoning attack [28] the attacker alter the input data to induce wrong model prediction. This technique is performed with data manipulation, data injection or logic corruption [29]. In test-time called evasion attack [30], the attacker aims to evade the trained model by tricking the input data.

2.3 Modeling the Attack Scenario

Huang et al. classified these threats on the basis of three (3) axes: *the influence on the classifiers, specificity and security violation (or impact)*. This taxonomy has been further studied by Biggio et al. [15], to model the attack scenario for a comprehensive understanding of the attacker strategy. According to [15], the attack scenario can be modeled based on the attacker's goal, knowledge, capability and strategy.

Adversary's Goal: This goal defines which security violation (Integrity, Availability and Privacy), the attacker aims to target and its specificity which mean if the attack is targeted or untargeted. It can be categorized in 3 types:

- *Integrity violation* that occurs when the adversary attempts to evade the detector. For instance, the attacker may aim to misclassify malicious sample as benign and result in an increase of false negative.
- *Availability violation* which leads to a useless system by creating many misclassifications. Thus increasing the false negative and false positive rate.
- *Privacy violation* in which the attacker try to get information from the learner.

In term of deep learning, papernot et al. [35] define the integrity violation as primary adversary's goal.

Adversary's Knowledge: This describes how well the attacker knows his target. Depending on the type of information there are three types of knowledge: white box, gray box and black box.

- White box: It assumes the adversary has complete information related to the network model: training data, features, learning algorithm, as well as trained model.
- Grey box: It assumes the attacker has partial knowledge about the target. This is also called the semi-white box.
- Black box: It assumes the attacker has zero or limited knowledge about the target. The attacker only knows the output of the model

Adversary's Capability: It assumes the types of influence the adversary can perform against the target.

Adversary's Strategy: It determines the workflow pursued by the adversary to launch the attack. The attack can be performed during the training time (poisoning attack) or during the test time (evasion attack).

3 Adversarial Attack Against NIDS

In this section, we review different studies that applied the adversarial machine learning in network intrusion detection system (NIDS) domain. As mentioned in Sect. 2 the attack can be performed during the training time called Poisoning attack or during test time called Evasion. We will review both evasion and poisoning attack and note down if the attack is performed in black box of white box scenario where possible.

3.1 Poisoning Attacks

Data Manipulation: Ali et al. [37] performed poisoning attack on DNN based IDS for a SDN-compliant heterogeneous wireless communication network. Launched in a white box using relabeling techniques in which malicious traffic is labeled as benign and normal traffic as malicious. Results show that the proposed poisoning attack decrease significantly the DNN classifier performance.

Papadopoulos et al. [33] Performed a label flipping attack in a white box to attack a SVM based NIDS for IoT environment. The method significantly degrade the model performance.

Data Injection: Nguyen et al. [61] propose a backdoor against federated learning based IoT NIDS. The adversary inject gradually on the compromised devices small amount of malicious data in the normal traffic during the training model. As a result they successfully reduce the model accuracy.

3.2 Evasion Attack

(a) *Adversarial Deep Learning Against Intrusion Detection Classifiers*: Rigaki et al. [40] investigate a targeted and untargeted gray box attack against RF, SVM, DT and their Majority ensemble voting. They generated adversarial sample with FGSM and JSMA on a multilayer perceptron (MLP) model and then transferred [19]. All classifier were affected, with the SVM being the most vulnerable and RF being the most robust. They analyzed the effect of the FGSM and JSMA. Concluded FGSM modified all features whereas JSMA alter only 6% of the feature. This make the JSMA more realistic.

(b) *Deep Learning-Based Intrusion Detection With Adversaries*: Wang et al. [20] performed a white box attack against MLP assessed on NSLKDD dataset. They generated adversarial examples with JSMA, FGSM, DEEPFOOL and CW. All attacks successfully degrade the performance of the MLP classifier, with the CW less devastating. They noticed that JSMA attack can achieve 100% probability of fooling the model with very less features.

(c) *Adversarial Attack against LSTM-based DDoS Intrusion Detection System*: Huang et al. [16] propose the first study on adversarial LSTM-based DDoS detection under black box setting. They utilized genetic Algorithm (GA) and Probability Weighted Packet Saliency Attack (PWPSA), to generate

adversarial samples. In their experiment Both methods can fool the detector with high success rates.

(d) *Adversarial Machine Learning in Network Intrusion Detection Systems*: Alhajjar et al. [9] generate adversarial examples to evade 11 machine learning models (SVM, DT, NB, KNN, RF, MLP, GB, LR, LDA, QDA, BAG). They Explore the use of GAN, and evolutionary algorithms: particle swarm optimization (PSO) and genetic Algorithm (GA) as adversarial examples. Use the Monte Carlo (MC) simulation as baseline and transfer the attacks. The authors consider the constrained nature of the feature space in NIDS and design these algorithms to perturb the inputs without modifying the malicious functionality of the networks. The experiment results show these perturbations were able to fool all models with a high misclassification rate. SVM and DT were the most vulnerable.

(e) *Adversarial Attacks Against NIDS in IoT Systems*: Qiu et al. [21] propose a realistic and efficient novel adversarial attack method against DNN model in NIDS for IoT in a black box environment. Their proposed approach uses the model extraction technique to reproduce target model for crafting adversarial examples and with a small portion of the original train data to achieve a high efficiency. Subsequently, to identify the most significant feature that influence the detector with the least modifications, a saliency maps [22] is used. Then generate perturbations using the FGSM adversarial sample. The method is applied to target Kitsune, a NIDS for IoT. The experimental results show the attacker can successfully compromise the detection system with an average success rate of 94.31%.

(f) *Launching Adversarial Attacks against Network Intrusion Detection Systems for IoT*: Papadopoulos et al. [33] Performed a white box adversarial attack against both traditional machine learning and deep learning model to evaluate their robustness in NIDS for IoT. In their methodology, they studied both poisoning and evasion attack. The evasion is performed with the FGSM against an ANN based IDS implemented with Bot-IoT dataset. The experiment result show a significant performance degradation. Moreover, authors mentioned traditional machine learning are more vulnerable during training time. Therefore the poisoning attack is performed on SVM model with the label flipping method.

(g) *Adversarial Attacks to bypass a GAN based classifier trained to detect Network intrusion*: Piplai et al. [31] studied the effectiveness of adversarial attacks against adversarial training. They revealed that even training the model with an adversarial training method, the attacker can still fool the model. Adversarial training is a defense technique that aims to increase the robustness of the model against adversarial attacks.

(h) *Black-Box Attack Method against Machine-Learning-Based Anomaly Network Flow Detection Models*: Similarly to [9], Guo et al. [32] analyzed the constrained domain on adversarial attacks against NIDS. They performed a black box attack with limited number of query. An extension of BIM adversarial sample is used to craft adversarial sample in a substitute MLP model in a white box setting. Then used the transferability to achieve the black

box attack. The method is evaluated on KDD99 and CICIDS2018 dataset. On KDD99, they targeted SVM, MLP, KNN, and CNN. Subsequently three model were targeted on CICIDS2018: Resnet, CNN and MLP. The experimental results show the proposed black box method can bypass the detector with high probability.

(i) *Adversarial Attack Against DoS Intrusion Detection: An Improved Boundary-Based Method*: Peng et al. [34] studied the robustness of ANN-based DoS IDS in a black box environment. They proposed an improved boundary based method to generate the adversarial samples. The presented approach optimizes a Mahalanobis distance by influencing the feature of both continuous and discrete DoS samples. The experimental results revealed that with limited queries, their proposed method can craft adversarial DoS examples and bypass the detection model.

(j) *A Brute-Force Black-Box Method to Attack Machine Learning-Based Systems in Cybersecurity*: Zhang et al. [36] propose a brute-force attack method (BFAM) to generate adversarial examples. The BFAM overcome some limit of GAN such as the unstable training [7]. They targeted LR, DT, MLP, naive Bayes (NB) and RF. Experimental results show that the proposed BFAM method is computational efficient and outperforms adversarial attack method based on GAN. However, RF has been the most resilient classifier to the generated adversarial example.

(k) *Generative adversarial attacks against intrusion detection systems using active learning*: Shu et al. [41] propose GAN active learning (Gen-AAL) to compromise the ML based NIDS in a black box with limited training data. In the GAN model the Variational AutoEncoder (VAE) is proposed as the generator and the discriminator is a MLP to implement a substitute model which approximate the target model. The active learning is used to decrease the number of required label to train the model. The experimental results show the proposed method achieve an evasion success rate of 98% by only using 25 labels instance during the training.

(l) *Evading a Machine Learning-based Intrusion Detection System through Adversarial Perturbations*: Fladby et al. [42] investigate an evasion attack against stratosphere linux ips (Slips) in a gray box setting. Slips is a ML-based Network Behavioral Analysis (NBA) which use the Markov chains algorithms. In the proposed method, authors use a custom attack to target the property network flow periodicity. The simultaneous perturbation stochastic approximation (SPSA) optimization method is used to perturb the network flows with minimal magnitude. Experimental results show the proposed method was able to evade the detector.

(m) *Evaluating Deep Learning Based Network Intrusion Detection System in Adversarial Environment*: Peng et al. [48] evaluate the robustness of four ML based NIDS under adversarial attack: RF, Logistic regression, SVM, and DNN respectively. The attack are performed with four adversarial samples: Projected Gradient Descent attack (PGD), Momentum Iterative FGSM (MI-FGSM), L-BFGS attack, and Simultaneous Perturbation Stochastic

Approximation (SPSA). All models performance sharply decrease and with the MI-FGSM attack achieving the highest attack success rate.

(n) *Analyzing Adversarial Attacks against Deep Learning for Intrusion Detection in IoT Networks*: Ibitoye et al. [49] investigate a white box attack against NIDS in IoT network. Two deep learning models have been first used to implement the NIDS; Feedforward Neural Networks (FNN) and its variant Self-normalizing Neural Network (SNN). Then the models resilience are evaluated. The adversarial samples are generated with FGSM and two of its variant: BIM and PGD. Both model performance degraded, however the SNN has been more resilient than the FNN. Moreover, authors found that feature normalization make the model vulnerable to adversarial sample.

(o) *Evaluating Deep Learning-based NIDS in Adversarial Settings*: Mohammadian et al. [50] investigated the effect of features and their vulnerability in a white box evasion attack. The approach targets an IDS implemented with DNN and utilizes a FGSM to generate attack. The attack was assessed on two datasets: CICIDS2017 and CIC-DDoS2019. To evaluated the most suitable feature for generating adversarial sample, they group features into different categories based on their nature. Then they craft adversarial sample in different feature set. The experiments show there are no general conclusion regarding the most vulnerable feature in both dataset.

(p) *NIDSGAN*: Zolbayar et al. [51] studied the effectiveness of GAN against ML based NIDS. They introduce NIDSGAN, an attack algorithm that generate adversarial network traffic to fool the IDS in a white-box, black-box and restricted black box evasion attacks. The approach take into account the domain constraints in network traffic to develop a realistic attack. In the proposed method, GAN is associates with active learning. The active learning method is used to decrease the training data size and enhance the attack success rate and GAN generates the attack. The attack is evaluated in two DNN models: AlertNet [52] and DeepNet [53]. The experimental results show the proposed method can evade the detector with a success rate of 99% in white box, 85% in black box and 70% in restricted black box.

(q) *A Comparative Study on Contemporary Intrusion Detection Datasets*: Pacheco et al. [18] evaluate the effectiveness of adversarial examples against the UNSB-NB15 and Bot-IoT datasets. Four NIDS target model were implemented using MLP, DT, RF and SVM. The attacks are performed in a white box with three adversarial sample generations: JSMA, FGSM and CW. The findings results demonstrate all models performance were degraded with RF beign the most resilient and SVM being the most vulnerable. And the JSMA attack has been the least effective in both datasets.

(r) *Black Box Attacks on Deep Anomaly Detectors*: Kuppa et al. [54] propose a realistic black box attack with limited queries to evade the detector. In the proposed approach, the Mani fold Approximation Algorithm is applied to the target model and is used to minimize the query. Then adversarial samples are generated with the spherical local subspaces. They evaluate the approach on 7 NIDS model: Isolation Forests (IF), Adversarially Learned Anomaly Detection (ALAD), One Class Support Vector Machines

(OC-SVM), Deep Autoencoding Gaussian Mixture Model (DAGMM), Deep Support Vector Data Description (DSVDD), AnoGAN and AutoEncoder (AE). The experiments show an attack success rate over 70%. However the proposed approach is more suitable for case where normal and attack boundaries are not well defined and when the NIDS is threshold based decision.

Table 1 summarizes the attacks method explored in this section.

Table 1. Summary of contributions in adversarial attacks against NIDS

Ref	Year	Environment	Dataset	Strategy	Knowledge	Target model	Attack Algorithm	Result
[20]	2018	Traditional	NSL-KDD	evasion	white box	MLP	JSMA, FGSM, DEEPFOOL, CW	CW attack less effective with AUC=0.80, FGSM most effective with AUC=0.44
[42]	2019	Traditional	CTU-13 dataset	evasion, poisoning	black box	RF, MLP, KNN	custom attack	MLP the most vulnerable one in evasion attack(68% attack severity), KNN is the most vulnerable model in poisoning attack(72% attack success rate)
[9]	2021	Traditional	NSL-KDD, UNSW-NB15	evasion	white box	SVM, DT, NB, KNN, RF, MLP, GB, LR, LDA, QDA, BAG	MC, PSO, GA, GAN	- DT and SVM the most vulnerable models(>90% evasion rate) In attack success, evolutionary computation methods (PSO and GA) achieved the best attack evasion rate on both dataset
[43]	2017	Traditional	NSL KDD	evasion	gray box	RF, SVM, DT based on CART, Majority ensemble voting	JSMA	SVM most vulnerable (27% accuracy drop), RF most resilient(10% accuracy drop)
[33]	2021	IoT	Bot-IoT	evasion, Poisoning	white box	ANN, SVM	FGSM, label poisoning	significant accuracy drop on untargeted and targeted poisoning attack(55% and 38% accuracy drop respectively), Significant accuracy drop on untargeted evasion attack(81% accuracy drop)
[21]	2021	IoT	Kitsune	evasion	black box	DNN	IFGSM, Substitute Model, Saliency Maps	94.31% attack success rate
[16]	2020	Traditional	CICIDS2017	evasion	black box	LSTM	PWPSA	High attack success rate on both methods(>69%)
[34]	2019	Traditional	KDDcup99, CICIDS2017	evasion	black box	ANN	Improved boundary-based method	Performance drop of true class confidence from 90% to 49%
[31]	2020	Traditional	ieee bigdata 2019 cup	evasion	white box	GAN	FGSM	At least 41% success rate with lesser significant features
[45]	2020	Traditional	CTU-108-1 dataset	evasion	gray box	Stratosphere Linux IPS	custom attack, SPSA	Attack method were able to effectively confuse the NIDS
[32]	2021	Traditional	KDDcup99, CICIDS2018	evasion	black box	SVM, MLP, KNN,CNN, Resnet	Substitute BIM	All models achieved a high recall rate in baseline Significant drop of model performance under adversarial attack In KDD99 with MLP, more than 94.2% adversarial DoS evade the detector , KNN was the most resilient model. in CICID2018 with MLP an average of 72.2% of evasion rate
[37]	2020	Traditional	NSL-KDD	evasion	black box	LR, DT, MLP,NB,RF	BFAM, GAN	BFAM outperforms GAN in most cases and decreases significantly the detection rate of target classifiers Best classifier under adversarial attack was RF(
[39]	2020	SDN	NSL-KDD	Poisoning	white box	DNN	relabelling	The accuracy rate decreases by about 13%-17%
[44]	2020	Traditional	CICIDS2017	evasion	gray box	Gradient boosted DT	Gen-AAL	achieve 98.86% attack success rate
[51]	2019	Traditional	NSL-KDD	evasion	white box	DNN, SVM, RF, LR	PGD, MI-FGSM, L-BFGS, SPSA	MI-FGSM was the most effective attack with an attack success rate of: 31% on DNN, 29% on SVM, 21% on RF and 21% on LR
[53]	2022	Traditional	CICIDS2017, CIC-DDoS2019	evasion	white box	DNN	FGSM	Difficult to conclude the most suitable features for generating adversarial examples features are dependent and related to each other
[52]	2019	IoT	Bot-IoT	evasion	white box	FNN, SNN	FGSM, BIM, PGD	Both model performance were significantly degraded but SNN demonstrates better resilience than FNN BIM was the most effective attack with an accuracy drop from 95.1% to 18% on FNN
[54]	2022	Traditional	NSKDD, CICIDS2017	evasion	white box, gray box, black box	LR, SVM, DT,KNN, DNN	GAN	Attack success rate on average 99%, 85%, and 70% on white box, gray box and black box respectively
[18]	2021	IoT	UNSB-NB15, Bot-IoT	evasion	white box	MLP, DT, RF, SVM	JSMA, FGSM, CW	All attacks were able to effectively degrade the overall classifiers, CW was most efficient attack on the UNSWNB15 (decrease in accuracy by 42%) JSMA was the least efficient on both datasets RF was shown as the most resilient classifier SVM the least robust
[57]	2019	Traditional	CICIDS2018	evasion	black box	DAGMM ,AE, AnoGAN, ALAD, DSVDD, one-class SVM, IF	MAA, spherical local subspaces	methods can evade NIDS with high success rate (attack success rate > 70%)

4 Defending Against Adversarial Attacks

In this session we summarize existing works that propose a defense method against these adversarial machine on NIDS.

4.1 Defense Against Poisoning Attack

Data Transformation: Poisoning attacks are generally injecting during retraining phase of the target system. Therefore, Apruzzese et al. [39] propose a data transformation which consist of inverting the training data before storing to the database. Therefore the poisoned data will not have much effect during retraining.

Pruning and Fine-Tuning: Bachl et al. [58] Investigated the defense against backdoor attacks in ML based NIDS. RF and MLP models have been used to implements de NIDS in UNSW- NB15 and CIC-IDS-2017. They proposed a pruning and fine-tuning as defense method to decrease the backdoor efficacity.

In their findings, authors reveal the proposed methods are efficient for random forest but not for neural network. Also they suggested Partial Dependence Plots (PDPs) and Accumulated Local Effects (ALE) plots as an efficient method to visualize backdoor attack.

4.2 Defense Against Evasion Attack

Adversarial Retraining

(a) *Adversarial Training for Deep Learning-based Intrusion Detection Systems*: Debicha et al. [38] propose adversarial training as a defense method. The experimental findings show the adversarial training improve the robustness of the IDS against attacks. Moreover the performance of the NIDS was compared to the baseline NIDS implemented without adversarial training. However, the results finding show the adversarial training decrease the performance of the IDS accuracy in free adversarial.

(b) *Evaluation of Adversarial Training on Different Types of Neural Networks in Deep Learning-based IDSs*: Khamis et al. [17] propose adversarial training based on min-max optimization as a defense technique againts adversarial attacks. To validate the method, they first evaluated three deep learning classifiers: DNN, ANN, RNN in an adversarial setting with five attack algorithms: FGSM, BIM, PGD, CW and deepfool. Assessed on NSLKDD and UNSW-NB15 datasets. All classifiers were affected in both datasets with a significant decrease of the accuracy compared to the baseline models. However the adversarial trained has significantly improved the model resilience.

(c) *GAN For Launching and Thwarting Adversarial Attacks on NIDS*: Usama et al. [55] propose GAN based adversarial training. They first utilize GAN to compromise the NIDS performance in a black box setting while maintaining the functional behavior. The method was evaluated on DNN, LR, SVM, KNN, naive Bayes (NB), RF, DT, and gradient boosting (GB) using the KDD99 dataset as benchmark. The experimental results showed the GAN successfully evade the detector with a decrease of all performance metric. As Defense method, authors proposed GAN based adversarial training. The adversarial training has enhanced the performance.

(d) *Adversarial Attacks Against Deep Learning-Based NIDS and Defense Mechanisms*: Zhang et al. [60] propose TIKI-TAKA, a framework to evaluate the robustness of deep learning based NIDS. In their approach, MLP, LSTM and CNN model based NIDS are first evaluated under adversarial attack in a black box built with five adversarial samples: Natural Evolution Strategies (NES) [43], Pointwise Attack [44], Boundary Atttack [45], OPT-Attack [46] and HopSkipJumpAttack [47]. Experiments show all models were vulnerable with an evasion success rates up to 37%. Then Three Defense methods have been proposed model voting ensembling, ensembling adversarial training, and query detection. These methods can be used jointly or separately and have been effective to decrease the success rate of evasion attacks.

Ensemble Model: Debicha et al. [56] investigated the ensemble model and adversarial training as defense method. They first studied the adversarial transferability method on network traffic between Neural network and multiple traditional machine learning based NIDS and trained with two different training sets. In a white box setting using FGSM and PGD attacks. The generated adversarial samples are transferred to five traditional ML based NIDS target: SVM, LR, DT, RF, Linear Discriminant Analysis (LDA), and their ensemble model. The experimental results show the attack transferred from DNN to traditional ML can successfully decrease the accuracy of the models with the DT and RF being more resilient. As defense method, the ensemble model and adversarial training have been applied. However the ensemble model did not improve the model robustness. In contrast, the adversarial training has improved the models resilience.

Defensive Distillation: Apruzzese et al. [57] introduce a variant of defensive distillation technique with RF against adversarial attack. In their approach, authors propose the use of probability labels to train the model instead of class labels applied in previous models. The experiments demonstrate the proposed method can decrease the impact of adversarial attack.

Feature Removal: Apruzzese et al. [39] investigated feature removal and adversarial training. They first performed an integrity violation attack on three machine learning algorithms: MLP, RF and KNN. The attack was assessed over the CTU-13 dataset. The experiment was performed in a black box attack. In the adversarial setting scenario, a custom adversarial attack is implemented. All classifiers were severely affected. Then authors propose two defense methods against the evasion attack: the adversarial retraining and feature removal. Both defense mitigated the attack severity.

Graph-Structured Data: Pujol-Perich et al. [59] propose a Graph Neural Network (GNN) based NIDS to improve the NIDS performance and its robustness against adversarial attack. The proposed GNN has been first evaluated in adversarial free and with state-of-the-art ML model based NIDS: MLP, RF, Ada-boost and decision tree ID3. The GNN model achieve a F score of 99% and is comparable to state of the art models. For the adversarial setup, two custom attacks were implemented. The first attack is implemented by increasing the packet size of attack flow. The second attack is performed by incrementing the inter-arrival time attack flow. In both attacks the GNN model has been robust as the accuracy keep the same level as in adversarial free. In contrast to the state-of -the art model which were vulnerable with a performance degradation up to 50%. Authors argue that the GNN can not only capture relevant pattern on each feature but can also seize the important structural flow pattern of attack. This ability make the GNN resilient against adversarial attack.

Table 2. Summary of contributions in adversarial defense against NIDS

Ref	Year	Environment	Dataset	Strategy	Knowledge	Target model	Attack Algorithm	defense
[17]	2020	Traditional	NSL-KDD, UNSW-NB15	evasion	white box	ANN, CNN, RNN	FGSM, BIM, PGD, CW, Deepfool	Min-Max
[40]	2021	Traditional	NSL-KDD	evasion	white box	DNN	FGSM, BIM, PGD, CW, Deepfool	Adversarial training
[63]	2022	Traditional	CICIDS2018	evasion	black box	MLP, CNN, LSTM	NES, BOUNDARY, HOPSKIPJUMPATTACK, POINTWISE, OPT	model voting ensembling, ensembling adversarial training, query detection.
[58]	2019	Traditional	KDD99	evasion	black box	DNN, LR, SVM, KNN, NB, RF, DT, GB	GAN	Adversarial training
[59]	2021	Traditional	NSL-KDD	evasion	black box	DNN, SVM, DT, LR, RF, LDA, ensemble model	FGSM, PGD	Ensemble model, Adversarial training
[60]	2020	Traditional	CTU-13	evasion	white box	RF	custom attack	Defensive distillation with RF
[61]	2019	Traditional	UNSW- NB15, CIC-IDS-2017	Poisoning	black box	MLP, RF	custom	pruning, fine-tuning
[62]	2021	Traditional	CICIDS2017	evasion	white box	GNN, MLP, RF, Ada-boost, ID3	custom attack	Graph Neural Network
[42]	2019	Traditional	CTU-13 dataset	evasion, poisoning	black box	RF, MLP, KNN	custom attack	Adversarial retraining feature removal against evasion Data transformation

5 Discussion

In the previous sessions, we explored several works that studied the adversarial machine learning in NIDS and their defenses. We can notice a yearly increase of papers, that demonstrate a growing interest on the impact of adversarial machine learning in network intrusion detection. Based on the surveyed studies, some important observations can be drawn:

- The majority of the papers fall into a white box attack assuming the adversary has full capability and knowledge. In intrusion detection domain this assumption is not realistic. It is unlikely that an adversary get power on the model internal configuration. However, white box attack can be useful to improve the NIDS model robustness from the algorithm designer or defender's point of view.
- Very few papers have addressed the constraint in network traffic. Contrary to image classification and object recognition which belong to unconstrained domain, network security application belongs to constrained domain [9,32]. The adversarial situation in network traffic is therefore quite different due to the three characteristics that we might have in the data: (1) we can have in a single feature different value (binary, categorical, continuous). (2) features in a dataset can be correlated. (3) some feature are key features and cannot be controlled by adversaries, in other word their modification might lead to a lost of critical information and therefore weaken the attack. However due to the

constrained domain some feature modification might break the functionality of the network traffic. Therefore adversarial machine learning that perform well in other applications have limited success in network [9,21]. More research is needed in this area to understand the feasability of these attacks.
- There are not many studies on the defenses technique in NIDS. Most of studies propose an adversarial training, however adversarial training has certain limitation. They cannot detect attacks that differ from the ones in the training dataset.
- Most of the studies focused on traditional networks. Fewer investigated these attack in IoT networks. More research is needed in IoT area. They are emerging in various contexts (e.g. federated learning), and need protection against adversaries.

6 Conclusion

Adversarial machine learning is a challenging and growing research area. Several approaches in NIDS has been presented recently. This confirm that despite the high performance of ML and DL applied in NIDSs, they are vulnerable to adversarial perturbation. This survey presents a comprehensive view of the different methodology of adversarial attacks applied against ML-based NIDS. It also discusses the different defense techniques proposed (summarized in Table 2). Furthermore, this survey addresses the limitations of the reviewed literature and outlines some directions for future work.

Acknowledgments. This research is supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT). This research is also supported by JSPS KAKENHI Grant Number 21K11888 and Hitachi Systems, Ltd. The last author, Kouichi SAKURAI, is grateful to The Telecommunications Advancement Foundation (TAF) for their academic support on this research.

References

1. Thakkar, A., Lohiya, R.: A review of the advancement in intrusion detection datasets. Procedia Comput. Sci. **167**, 636–645 (2020)
2. Lazarevic, A., Kumar, V., Srivastava, J.: Intrusion detection: a survey. In: Kumar, V., Srivastava, J., Lazarevic, A. (eds) Managing Cyber Threats. Massive Computing, vol. 5, pp. 19–78. Springer, Boston (2005). https://doi.org/10.1007/0-387-24230-9_2
3. Hindy, H., et al.: A taxonomy of network threats and the effect of current datasets on intrusion detection systems. IEEE Access **8**, 104650–104675 (2020)
4. Ahmad, Z., Shahid Khan, A., Wai Shiang, C., Abdullah, J., Ahmad, F.: Network intrusion detection system: a systematic study of machine learning and deep learning approaches. Trans. Emerg. Telecommun. Technol. **32**(1), e4150 (2021)
5. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)

6. Huang, L., Joseph, A.D., Nelson, B., Rubinstein, B.I., Tygar, J.D.: Adversarial machine learning. In: Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence, pp. 43–58 (2011)
7. Goodfellow, I., et al.: Generative adversarial nets. In: Proceedings of the 27th International Conference on Neural Information Processing Systems, vol. 2, pp. 2672–2680 (NIPS 2014). MIT Press, Cambridge (2014)
8. Szegedy, C., et al.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2014)
9. Alhajjar, E., Maxwell, P., Bastian, N.: Adversarial machine learning in network intrusion detection systems. Expert Syst. Appl. **186**, 115782 (2021)
10. Liu, Q., Li, P., Zhao, W., Cai, W., Yu, S., Leung, V.C.: A survey on security threats and defensive techniques of machine learning: a data driven view. IEEE Access **6**, 12103–12117 (2018)
11. Moisejevs, I.: Adversarial attacks and defenses in intrusion detection systems: a survey. Int. J. Artif. Intell. Expert Syst. **8**(3), 44–62 (2019)
12. Martins, N., Cruz, J.M., Cruz, T., Abreu, P.H.: Adversarial machine learning applied to intrusion and malware scenarios: a systematic review. IEEE Access **8**, 35403–35419 (2020)
13. Carlini, N.: A complete list of all (arXiv) adversarial example papers. https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html. Accessed 30 Oct 2021
14. Liu, H., Lang, B.: Machine learning and deep learning methods for intrusion detection systems: a survey. Appl. Sci. **9**(20), 4396 (2019)
15. Biggio, B., Fumera, G., Roli, F.: Security evaluation of pattern classifiers under attack. IEEE Trans. Knowl. Data Eng. **26**(4), 984–996 (2013)
16. Huang, W., Peng, X., Shi, Z., Ma, Y.: Adversarial attack against LSTM-based DDoS intrusion detection system. In: 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI), pp. 686–693. IEEE (2020)
17. Abou Khamis, R., Matrawy, A.: Evaluation of adversarial training on different types of neural networks in deep learning-based IDSs. In: 2020 International Symposium on Networks, Computers and Communications (ISNCC), pp. 1–6. IEEE (2020)
18. Pacheco, Y., Sun, W.: Adversarial machine learning: a comparative study on contemporary intrusion detection datasets. In: ICISSP, pp. 160–171 (2021)
19. 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)
20. Wang, Z.: Deep learning-based intrusion detection with adversaries. IEEE Access **6**, 38367–38384 (2018)
21. Qiu, H., Dong, T., Zhang, T., Lu, J., Memmi, G., Qiu, M.: Adversarial attacks against network intrusion detection in IoT systems. IEEE Internet Things J. **8**(13), 10327–10335 (2020)
22. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
23. Berman, D.S., Buczak, A.L., Chavis, J.S., Corbett, C.L.: A survey of deep learning methods for cyber security. Information **10**(4), 122 (2019)
24. Buczak, A.L., Guven, E.: A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Commun. Surv. Tutor. **18**, 1153–1176 (2016)

25. Biggio, B., Roli, F.: Wild patterns: ten years after the rise of adversarial machine learning. Pattern Recogn. **84**, 317–331 (2018)
26. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 427–436 (2015)
27. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: DeepFool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2574–2582 (2016)
28. Muñoz-González, L., et al.: Towards poisoning of deep learning algorithms with back-gradient optimization. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, pp. 27–38 (2017)
29. Tabassi, E., Burns, K.J., Hadjimichael, M., Molina-Markham, A.D., Sexton, J.T.: A taxonomy and terminology of adversarial machine learning. In: NIST IR, pp. 1–29 (2019)
30. Biggio, B., et al.: Evasion attacks against machine learning at test time. In: Blockeel, H., Kersting, K., Nijssen, S., Železný, F. (eds.) ECML PKDD 2013. LNCS (LNAI), vol. 8190, pp. 387–402. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40994-3_25
31. Piplai, A., Chukkapalli, S.S.L., Joshi, A.: NAttack! adversarial attacks to bypass a GAN based classifier trained to detect Network intrusion. In: 2020 IEEE 6th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS), pp. 49–54. IEEE (2020)
32. Guo, S., et al.: A black-box attack method against machine-learning-based anomaly network flow detection models. Secur. Commun. Netw. **2021**, 1–13 (2021)
33. Papadopoulos, P., Thornewill von Essen, O., Pitropakis, N., Chrysoulas, C., Mylonas, A., Buchanan, W.J.: Launching adversarial attacks against network intrusion detection systems for IoT. J. Cybersecur. Priv. **1**(2), 252–273 (2021)
34. Peng, X., Huang, W., Shi, Z.: Adversarial attack against DoS intrusion detection: an improved boundary-based method. In: 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), pp. 1288–1295. IEEE (2019)
35. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A.: The limitations of deep learning in adversarial settings. In: 2016 IEEE European Symposium on Security And Privacy (EuroS&P), pp. 372–387. IEEE (2016)
36. Zhang, S., Xie, X., Xu, Y.: A brute-force black-box method to attack machine learning-based systems in cybersecurity. IEEE Access **8**, 128250–128263 (2020)
37. Ali, M., Hu, Y.F., Luong, D.K., Oguntala, G., Li, J.P., Abdo, K.: Adversarial attacks on AI based intrusion detection system for heterogeneous wireless communications networks. In: 2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC), pp. 1–6. IEEE (2020)
38. Debicha, I., Debatty, T., Dricot, J.M., Mees, W.: Adversarial training for deep learning-based intrusion detection systems. arXiv preprint arXiv:2104.09852 (2021)
39. Apruzzese, G., Colajanni, M., Ferretti, L., Marchetti, M.: Addressing adversarial attacks against security systems based on machine learning. In: 2019 11th International Conference on Cyber Conflict (CyCon), vol. 900, pp. 1–18. IEEE (2019)
40. Rigaki, M.: Adversarial deep learning against intrusion detection classifiers (2017)
41. Shu, D., Leslie, N.O., Kamhoua, C.A., Tucker, C.S.: Generative adversarial attacks against intrusion detection systems using active learning. In: Proceedings of the 2nd ACM Workshop on Wireless Security and Machine Learning, pp. 1–6 (2020)

42. Fladby, T., Haugerud, H., Nichele, S., Begnum, K., Yazidi, A.: Evading a machine learning-based intrusion detection system through adversarial perturbations. In: Proceedings of the International Conference on Research in Adaptive and Convergent Systems, pp. 161–166 (2020)
43. Ilyas, A., Engstrom, L., Athalye, A., Lin, J.: Black-box adversarial attacks with limited queries and information. In: International Conference on Machine Learning, pp. 2137–2146. PMLR (2018)
44. Schott, L., Rauber, J., Bethge, M., Brendel, W.: Towards the first adversarially robust neural network model on MNIST. arXiv preprint arXiv:1805.09190 (2018)
45. Brendel, W., Rauber, J., Bethge, M.: Decision-based adversarial attacks: reliable attacks against black-box machine learning models. arXiv preprint arXiv:1712.04248 (2017)
46. Liu, S., Sun, J., Li, J.: Query-efficient hard-label black-box attacks using biased sampling. In: 2020 Chinese Automation Congress (CAC), pp. 3872–3877. IEEE (2020)
47. Chen, J., Jordan, M.I.: HopSkipJumpAttack: a query-efficient decision-based attack. IEEE Secur. Priv. **2020**, 1277–1294 (2020)
48. Peng, Y., Su, J., Shi, X., Zhao, B.: Evaluating deep learning based network intrusion detection system in adversarial environment. In: 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC), pp. 61–66. IEEE (2019)
49. Ibitoye, O., Shafiq, O., Matrawy, A.: Analyzing adversarial attacks against deep learning for intrusion detection in IoT networks. In: 2019 IEEE Global Communications Conference (GLOBECOM), pp. 1–6. IEEE (2019)
50. Mohammadian, H., Lashkari, A.H., Ghorbani, A.A.: Evaluating deep learning-based NIDS in adversarial settings. In: ICISSP, pp. 435–444 (2022)
51. Zolbayar, B.E., et al.: Generating practical adversarial network traffic flows using NIDSGAN. arXiv preprint arXiv:2203.06694 (2022)
52. Vinayakumar, R., et al.: Deep learning approach for intelligent intrusion detection system. IEEE Access **7**, 41525–41550 (2019)
53. Gao, M., Ma, L., Liu, H., Zhang, Z., Ning, Z., Xu, J.: Malicious network traffic detection based on deep neural networks and association analysis. Sensors **20**(5), 1452 (2020)
54. Kuppa, A., Grzonkowski, S., Asghar, M.R., Le-Khac, N.A.: Black box attacks on deep anomaly detectors. In: Proceedings of the 14th International Conference on Availability, Reliability and Security, pp. 1–10 (2019)
55. Usama, M., Asim, M., Latif, S., Qadir, J.: Generative adversarial networks for launching and thwarting adversarial attacks on network intrusion detection systems. In: 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC), pp. 78–83. IEEE (2019)
56. Debicha, I., Debatty, T., Dricot, J.-M., Mees, W., Kenaza, T.: Detect & reject for transferability of black-box adversarial attacks against network intrusion detection systems. In: Abdullah, N., Manickam, S., Anbar, M. (eds.) ACeS 2021. CCIS, vol. 1487, pp. 329–339. Springer, Singapore (2021). https://doi.org/10.1007/978-981-16-8059-5_20
57. Apruzzese, G., Andreolini, M., Colajanni, M., Marchetti, M.: Hardening random forest cyber detectors against adversarial attacks. IEEE Trans. Emerg. Top. Comput. Intell. **4**(4), 427–439 (2020)

58. Bachl, M., Hartl, A., Fabini, J., Zseby, T.: Walling up backdoors in intrusion detection systems. In: Proceedings of the 3rd ACM CoNEXT Workshop on Big Data, Machine Learning and Artificial Intelligence for Data Communication Networks, pp. 8–13 (2019)
59. Pujol-Perich, D., Suárez-Varela, J., Cabellos-Aparicio, A., Barlet-Ros, P.: Unveiling the potential of graph neural networks for robust intrusion detection. ACM SIGMETRICS Perform. Eval. Rev. **49**(4), 111–117 (2022)
60. Zhang, C., Costa-Pérez, X., Patras, P.: Adversarial attacks against deep learning-based network intrusion detection systems and defense mechanisms. IEEE/ACM Trans. Netw. **30**, 1294–1311 (2022)
61. Nguyen, T.D., Rieger, P., Miettinen, M., Sadeghi, A.R.: Poisoning attacks on federated learning-based IoT intrusion detection system. In: Proc. Workshop Decentralized IoT Syst. Secur. (DISS), pp. 1–7 (2020)