Tiki-Taka: Attacking and Defending Deep Learning-based Intrusion Detection Systems

Tiki-Taka: Attacking and Defending Deep Learning-based Intrusion Detection Systems

Chaoyun Zhang
University of Edinburgh, UK
chaoyun.zhang@ed.ac.uk

Xavier Costa-Pérez
NEC Laboratories Europe, Germany
i2CAT Foundation, Spain
ICREA, Spain
xavier.costa@ieee.org

Paul Patras
University of Edinburgh, UK
paul.patras@ed.ac.uk

ABSTRACT

Neural networks are increasingly important in the development of Network Intrusion Detection Systems (NIDS), as they have the potential to achieve high detection accuracy while requiring limited feature engineering. Deep learning-based detectors can be however vulnerable to adversarial examples, by which attackers that may be oblivious to the precise mechanics of the targeted NIDS add subtle perturbations to malicious traffic features, with the aim of evading detection and disrupting critical systems in a cost-effective manner. Defending against such adversarial attacks is therefore of high importance, but requires to address daunting challenges.

In this paper, we introduce Tiki-Taka, a general framework for *(i)* assessing the robustness of state-of-the-art deep learning-based NIDS against adversarial manipulations, and which *(ii)* incorporates our proposed defense mechanisms to increase the NIDS' resistance to attacks employing such evasion techniques. Specifically, we select five different cutting-edge adversarial attack mechanisms to subvert three popular malicious traffic detectors that employ neural networks. We experiment with a publicly available dataset and consider both one-to-all and one-to-one classification scenarios, i.e., discriminating illicit vs benign traffic and respectively identifying specific types of anomalous traffic among many observed. The results obtained reveal that, under realistic constraints, attackers can evade NIDS with up to 35.7% success rates, by only altering time-based features of the traffic generated. To counteract these weaknesses, we propose three defense mechanisms, namely: model voting ensembling, ensembling adversarial training, and query detection. To the best of our knowledge, our work is the first to propose defenses against adversarial attacks targeting NIDS. We demonstrate that when employing the proposed methods, intrusion detection rates can be improved to nearly 100% against most types of malicious traffic, and attacks with potentially catastrophic consequences (e.g., botnet) can be thwarted. This confirms the effectiveness of our solutions and makes the case for their adoption when designing robust and reliable deep anomaly detectors.

CCSW'20, November 9, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-8084-3/20/11…$15.00
https://doi.org/10.1145/3411495.3421359

CCS CONCEPTS

• **Security and privacy → Artificial immune systems**; • **Computing methodologies → Neural networks**.

KEYWORDS

Adversarial Attacks, Network Intrusion Detection Systems, Deep Learning

ACM Reference Format:
Chaoyun Zhang, Xavier Costa-Pérez, and Paul Patras. 2020. Tiki-Taka: Attacking and Defending Deep Learning-based Intrusion Detection Systems. In *2020 Cloud Computing Security Workshop (CCSW'20), November 9, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3411495.3421359

1 INTRODUCTION

Network Intrusion Detection (NID) aims at identifying malicious traffic flows, so as to protect computers, networks, servers, and data from attacks, unauthorized access, modification, or destruction [6]. Given the unprecedented growth in the data traffic volume transiting both wired and wireless infrastructure, NID is becoming increasingly important to ensure system/service availability and protect individuals' safety and privacy online. As new cyberattacks proliferate, traditional intrusion detection methodologies that rely on pattern matching (e.g., IP address and port number) and classification are losing effectiveness [61]. In this context, machine learning-based solutions are gaining traction, as they rely increasingly less often on deep packet inspection (hence raising fewer privacy concerns) and may have better generalization abilities.

Stimulated by recent success in areas such as image classification, the limited extent of feature engineering involved, and the decreasing cost of parallel processing hardware, deep learning – a subset of machine learning – is making its way also in the networking domain [58]. This includes NID, where solutions based on Deep Neural Networks (DNNs) yield demonstrably superior detection accuracy (see, e.g., [25, 47]). However, due to their complex structures, DNNs also suffer from limited interpretability, which inevitably raises important questions: *Is deep learning a truly reliable option for NID? Is there any "Achilles' heel" that can be exploited to compromise the expected high detection accuracy of neural network-based NID models?* Answering these questions is crucial to guaranteeing the reliability of Network Intrusion Detection Systems (NIDS).

Unfortunately, DNNs have been proven vulnerable to adversarial examples [36] or backdoor attacks [52] in several applications [13, 21], whereby they can be fooled by subtle perturbations introduced in the input [54], which interfere with the correctness of the inferences made. Since such adversarial manipulations are

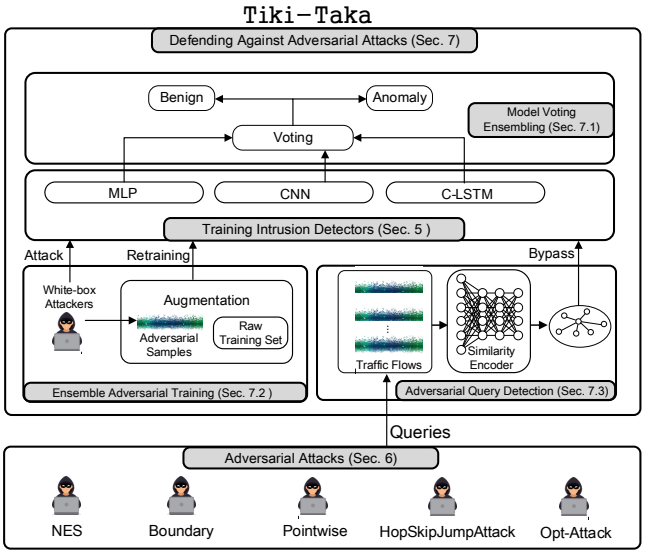

Figure 1: The Tiki-Taka framework for crafting and defending against adversarial attacks towards Network Intrusion Detection Systems (NIDS).

often extremely difficult to detect, deep learning-based NIDS are also at risk. Attackers, potentially unaware of the properties of an NIDS (i.e., black-box system), could generate adversarial samples by repeatedly changing small subsets of the traffic features, and make 'queries' to the NIDS. After each query, the attacker receives some feedback (e.g., an acknowledgment or lack of any response), which indicates the success or failure of the attack attempt. Based on this feedback, the attacker can adjust the perturbations on selected features (e.g., intervals between consecutive packets) of the traffic, or introduce new ones, without changing its essence, until succeeding in bypassing the NIDS [23, 46]. By this approach, malicious flows could then disguise into benign traffic and compromise their targets [28], while remaining undetected even by NIDS thought to be highly accurate. Cyberwarfare is exploding [17] and such **adversarial strategies offer cost-effective means to jeopardize healthcare systems, electronic voting, banking, industrial automation, and countless more**.

In this paper, we tackle the severe intrusion detection issues faced by classifiers under adversarial attacks. We first scrutinize the robustness of state-of-the-art deep learning NID models against different adversarial mechanisms, considering attacks in practical decision-based settings (i.e., attackers can only infer if the traffic generated was classified as benign or malicious, without knowledge about the exact class to which the traffic was mapped). We test the effectiveness and efficiency of each attack in two detection scenarios: one-to-all and one-to-one, i.e., aiming to discriminate malicious vs. benign traffic, and respectively to identify precise types of attacks. We then propose three solutions to defend against this new class of threats, which effectively reduce the success rate of each attack to a large extent. This enables more robust and reliable NID. In a nutshell we make the following **key contributions**:

[C1] We implement three types of DNN architectures based on state-of-the-art NID models, i.e., Multilayer Perceptron (MLP), Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM), and perform NID on the realistic cyber defense dataset CSE-CIC-IDS2018 [42]. The NID models implemented achieve over 98.7% detection accuracy based on a limited set of features, which matches the performance of previously reported NIDS implementations (details in Sec. 5).

[C2] To demonstrate the NIDS considered can be evaded, we employ five state-of-the-art attack strategies to generate adversarial samples (i.e., NES, Boundary, HopSkipJumpAttack, Pointwise, and Opt-Attack), bounding traffic feature manipulations to realistic domain constraints (i.e., leaving unchanged those features that may alter flow semantics). We conduct a comprehensive evaluation on the effectiveness of each adversarial attack and provide an in-depth analysis of their characteristics (see Sec. 6).

[C3] We propose three defense mechanisms to strengthen deep learning-based NIDS against adversarial attacks, namely: model voting ensembling, ensembling adversarial training, and query detection. Each defense method can either operate individually or jointly with the others. Experiments show these methods drastically reduce the attack success rates, significantly improving the robustness of the NIDS considered as we bring detection rates close to 100% (Sec. 7).

We name our general attack–defense framework Tiki-Taka[1] and illustrate the workflow of our methodology in Figure 1. To the best of our knowledge, **we are the first to introduce defense mechanisms against adversarial attacks targeting NIDS.**

2 RELATED WORK

DNNs are increasingly used for NID purposes, as they help minimize feature engineering efforts and operate with high detection accuracy [6]. However, recent research suggests that there exist loopholes that can degrade the performance of neural NIDS, as perturbation added to their input can trigger traffic misclassification [20, 58]. Thus, defending deep learning-based approaches from adversarial samples becomes a crucial issue for network security.

2.1 Deep Learning-based NID

Niyaz *et al.* employ sparse autoencoders for self-taught learning and extract important features from traffic flows [25]. They conduct NID on the NSL-KDD dataset [44] and achieve 98.84% F1 scores. Faker and Dogdu design an MLP to discriminate malicious traffic [16] in the CIC-IDS2017 dataset [42]. Although the model structure is simple, the MLP achieves significantly higher detection rates than Random Forest (RF), Gradient Boosting Tree (GPT), and Support Vector Machine (SVM) structures. Similar conclusions have been reached in [47], where Vinayakumar *et al.* compare the MLP with a large set of traditional machine learning approaches to NID, showing that deep learning yields better performance.

CNN-based approaches have been employed for NID as well [48]. Zhang *et al.* design a two-branch CNN and employ feature fusion, to resolve the class imbalance problem of the dataset used [59]. Their proposal detects a minor class of anomalies with higher accuracy, while being more efficient in terms of execution time. Recurrent

[1] Tiki-Taka is a football tactic that encourages short and fast ball passing, and tackling on the spot when losing ball possession. We use this to metaphorize the frequent queries passed to an NIDS in the attack process, with the detector subsequently regaining control through defenses.

Neural Networks (RNNs) are popular candidates for extracting temporal features of traffic flows [53]. Zhang *et al.* perform NID on raw packet-level traffic [60]. They combine CNNs and LSTMs to extract important spatial and temporal features, achieving higher detection rates than when using each of these components individually.

In our study, we select MLP, CNN, and LSTM as representative models to perform NID, then test them against adversarial attacks, and subsequently augment these models with a set of defense mechanisms that we propose for enhancing their robustness.

2.2 Attacking Deep Learning-based NIDS

The majority of existing methods that employ adversarial samples to compromise classifiers target image applications (e.g., [19, 30, 35]). Research on evading deep learning-based NIDS is scarce. Wang *et al.* employ four sets of white-box attack algorithms designed for image classification, to bypass MLP-based intrusion detectors trained on the NSL-KDD dataset [49]. Their experiments suggest that these attack algorithms are transferable to the NID domain and the MLP detectors are vulnerable to adversarial samples. However, attackers may not have access to the neural model underpinning the targeted NIDS, which make such settings more useful to NIDS designers to assess the robustness of their systems [28].

Yang *et al.* generate adversarial samples in black-box settings [51] using three types of approaches, namely surrogate models [8], Zeroth Order Optimization (ZOO) [10], and Generative Adversarial Networks (GANs) [3]. These methods can reduce the performance of MLP-based classifiers, thus becoming a threat to NIDS. Kuppa *et al.* consider a more realistic situation, performing black-box attacks against different deep learning-based detectors in decision-based and query-limited settings [28]. By learning and approximating the distribution of benign and anomalous samples, these methods can evade NIDS with high success rate.

2.3 Defending from Adversarial Samples

There exist a range of strategies for defending deep learning models from adversarial examples. Commonly used methods include Network Distillation [37], Adversarial Training [45], Adversarial Detecting [31], Input Reconstruction [40], Classifier Robustifying [2], Network Verification [26], and an ensemble of them [33, 54], which work either reactively or proactively. Network distillation methods employ a student neural network to learn knowledge from a more complex teacher network. With this approach, the student network generalizes better and becomes more robust to adversarial samples. Adversarial training retrains the neural networks by augmenting the original training set with adversarial samples, such that they can better defend against those inputs with subtle feature perturbations. Input reconstruction reduces the effectiveness of the perturbations by recovering the original input. Classifier robustifying employs various approaches (e.g., model ensembling) to improve the robustness of the original classifier. Network verification uses an additional classifier to identify adversarial samples.

While such approaches can be effective in the computer vision and natural language processing domains, *(i)* work by Carlini *et al.* demonstrates defense mechanisms against adversarial examples in imaging can be defeated by constructing new loss functions [7], while *(ii)* none of these are aimed at defending against adversarial samples in NIDS, which is the problem we tackle in this paper.

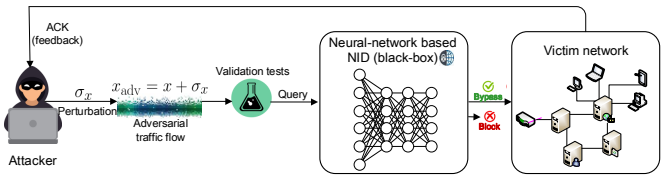

Figure 2: An illustration of the attack process against machine learning-based NID models.

3 THREAT MODEL

We focus on scenarios where attackers generate adversarial samples by adding small perturbations to the input given to NIDS, thereby aiming to cause misclassification and evade detection of their malicious traffic. As in [24], we denote by x the input to a classifier (i.e., features extracted from flows), an adversarial sample as $x_{\text{adv}} = x + \sigma_x$, and the targeted class as y_{adv}. The objective of the adversarial attacks can be formulated as finding x_{adv} such that $||x_{\text{adv}} - x||_\infty < \epsilon$ and x_{adv} is classified as y_{adv}. Here, σ_x is the perturbation added to the input and ϵ limits the perturbation scale.

3.1 Adversarial Settings

Typical attacks against machine learning models can be categorized into three classes: *(i)* white-box attacks, *(ii)* grey-box attacks, and *(iii)* black-box attacks. White-box and grey-box attacks assume the malicious actors have access to the training data or/and model structures. Such hypotheses apply in cases where system designers seek to improve the robustness of their NIDS, but less commonly to scenarios with external adversaries. At best, malicious actors could conduct white-box attacks against own models and subsequently seek to transfer these attacks onto victim NIDS. More often, potential hackers are forced to treat a NIDS as a black-box, since the details of a victim system's inner workings remain hidden and the only way in which the NIDS behavior can be learned is through a sequence of queries and the feedback received. This is also the primary practical threat model that we consider in this work, while the defense mechanisms we propose can also fend off adversarial samples adapted from white-box methods, as we reveal.

In general, an attacker may send a traffic flow towards the target network, which will be first examined by a NID model. This is known as a query process. Subsequently, the attacker will receive implicit/explicit feedback from the model, e.g., an ACK packet, which reflects whether the traffic flow was classified as anomalous. Based on the feedback, the attacker can adjust and apply subtle perturbations to the malicious traffic flows, thereby producing adversarial samples that eventually may compromise the effectiveness of the NIDS, which will end up classifying malicious traffic as benign. On the other hand, the attacker may not have confidence about the exact decision class decided by the NIDS, but whether the traffic was deemed malicious or benign (decision-based attacks). We illustrate this attack process against NID models in Figure 2.

3.2 Domain Constraints

Unlike adversarial attacks against image classifiers, adversarial samples against NIDS must respect certain domain constrains [28], such that the functionality and intactness of the samples is preserved when introducing perturbations σ_x. This means that *(i)* only a subset of features are amendable; and *(ii)* the features of adversarial samples do not violate the properties inherent to the original samples. To meet these requirements, here we confine consideration to 22 time-based features, to which we add perturbations, as also suggested in [28]. These include *(a)* Forward Inter Arrival Time – the time between two packets sent in the forward direction (mean, min, max, std); *(b)* Backward Inter Arrival Time – the time between two packets sent in reverse direction (mean, min, max, std); *(c)* Active-Idle Time – the amount of time a flow was idle before becoming active, and vice-versa (mean, min, max, std), *(d)* Average number of bytes and packets sent in forward and backward directions in the initial window or/and sub-flows. It is conceivable that an attacker may also attempt to mimic the time features of benign flows and conceal malicious content within payloads (e.g., SQL injection, cross-site scripting, etc.). In such cases, our Tiki-Taka framework can also accommodate payload-based features extracted, e.g., through word embedding or Text-CNNs [34]. Features outside the subset that may change flow semantics remain unchanged during the attack, which is inline with recent research confirming that adversarial samples can be constructed effectively by perturbing only a small subset of the input features [43].

In addition, we expect that *(i)* the Mean Absolute Percentage Error (MAPE) for each feature k does not exceed 20%, i.e., $100 \cdot |(x^{(k)} - x^{(k)}_{\text{adv}})/x^{(k)}| \leq 20\%$; *(ii)* the perturbed features preserve the mean property (e.g., the mean forward inter arrival time) plus/minus std features do not exceed their corresponding max and min features; *(iii)* the sign of each perturbed sample remains the same as that of the original; and *(iv)* if the std feature is zero, the corresponding mean, max, min and std features remain unchanged in the adversarial sample. Samples that violate these constrains are to be regarded as unsuccessful trials, since they alter the originally intended functionality of the flows. In crafting adversarial attacks for our study, we will also perform validation tests based on these constraints.

4 DATASET

We conduct all experiments (i.e., NID, black-box adversarial attacks, and attacks against NIDS augmented with the proposed defenses) using the publicly available CSE-CIC-IDS2018 dataset [42]. This encompasses 14 types of network intrusion traffic flows along with benign traffic. The attacks can be categorized into seven classes, namely Brute Force, Heartbleed, Botnet, Denial of Service (DoS), Distributed Denial of Service (DDoS), Web attacks, and infiltration. Table 1 summarizes the prevalence of each type of traffic. The infrastructure employed includes 50 machines, which attempt to intrude a victim network consisting of 420 end hosts and 30 servers.

A total of 80 features of the traffic flows are extracted to perform intrusion detection and we filter 65 of them for the purpose of our work. The features selected can be grouped into 8 classes, specifically *(a)* Forward Inter Arrival Time – the time between two packets sent in the forward direction (mean, min, max, std); *(b)* Backward Inter Arrival Time – the time between two packets sent in reverse direction (mean, min, max, std); *(c)* Flow Inter Arrival Time – the time between two packets sent in either direction (mean, min, max, std); *(d)* Active-Idle Time – the amount of time a flow was idle before becoming active (mean, min, max, std) and the amount of time a flow was active before becoming idle (mean, min, max, std); *(e)* Flags based features – the number of times the URG, PSH flags are set, both in the forward and backward direction; *(f)* Flow characteristics – bytes per second, packets per second, flow length (mean, min, max, std) and ratio between number of bytes sent downlink and uplink; *(g)* Packet count with flags FIN, SYN, RST, PUSH, ACK, URG, CWE and ECE; *(h)* Average number of bytes and packets sent in forward/backward directions in the initial window, bulk rate, and sub flows count. Our framework is readily extensible to other types of features, e.g., extracted from payloads [34].

We train all deep learning models, implement and defend against the adversarial attacks using the selected features.

Table 1: Statistics of the CSE-CIC-IDS2018 dataset employed.

Flow Type	Number of Instances	Ratio
Benign	14,097,779	83.6861%
Bot	286,191	1.6989%
DoS attack-SlowHTTPTest	139,890	0.8304%
DoS attack-Hulk	461,912	2.7420%
Brute Force-XSS	230	0.0014%
SQL Injection	87	0.0005%
Infiltration	161,934	0.9613%
DoS attack-GoldenEye	41,508	0.2464%
DoS attack-Slowloris	10,990	0.0652%
Brute Force-Web	611	0.0036%
FTP-Brute Force	193,360	1.1478%
SSH-Brute Force	187,589	1.1136%
DDoS attack-LOIC-UDP	1,730	0.0103%
DDoS attack-HOIC	686,012	4.0723%
DDoS attack-LOIC-HTTP	576,191	3.4203%
All of the above attacks	2,748,235	16.3139%
Total	16,846,014	100%

5 TRAINING INTRUSION DETECTORS

Training accurate deep network intrusion detectors is the initial important step of our study, as Tiki-Taka builds on the pre-trained NID models. To this end, we employ three well-known deep learning architectures, namely Multilayer Perceptron (MLP) [16], Convolutional Neural Network (CNN) [59], and CNN with Long Short-Term Memory (LSTM) layers, i.e., C-LSTM [60]. These models are

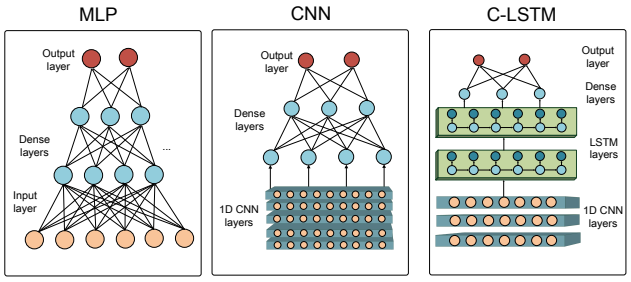

Figure 3: Architectures of the deep learning-based Network Intrusion Detection (NID) models employed in this study.

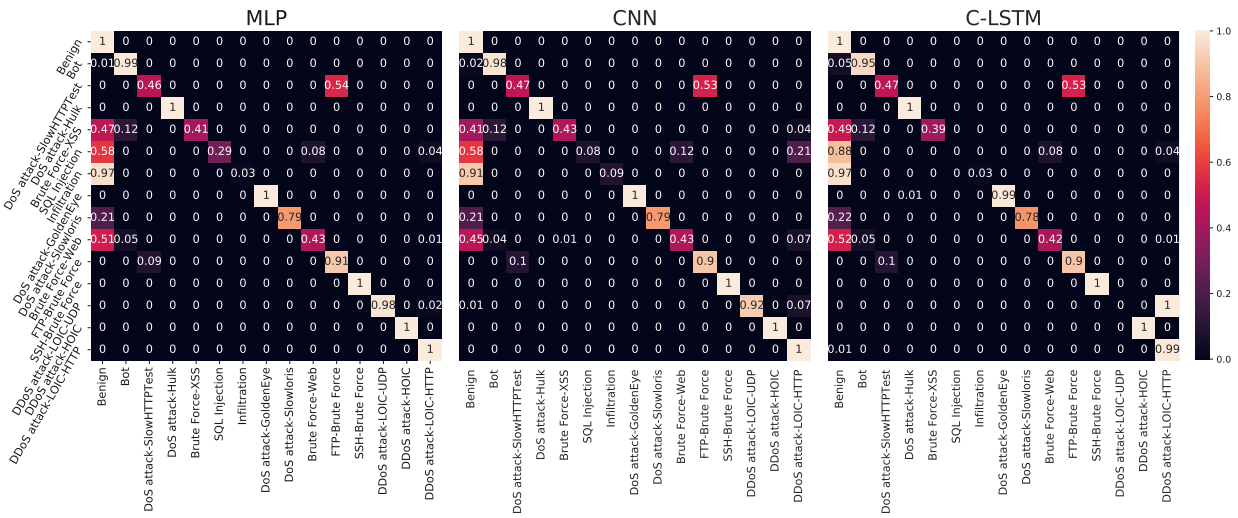

Figure 4: Confusion matrices of the MLP, CNN, and C-LSTM models for the one-to-one NID.

frequently used for NID purposes and have achieved notable performance. We illustrate the architecture of each model in Figure 3.

The MLP is the most simple deep learning architecture, which employs multiple stacks of fully-connected layers for features extraction. It is particularly suitable for handling traffic flows that have mixture type features and ranges. In our study, we construct an MLP with 3 hidden layers. Each layer has 200 units, except for the last hidden layer, which has 400 units. CNNs have good spatial perception abilities and have demonstrated remarkable precision in NID tasks [59]. In this work, we design a CNN with 10 one-dimensional CNN layers, each equipped with 108 filters, with filter size 5. Lastly, we replicate the C-LSTM employed in [60], with our C-LSTM operating on the features that characterize the traffic, instead of operating on raw flows. The C-LSTM combines CNN and LSTM structures to extract spatial and temporal features separately. Data will be first processed by a CNN with 5 hidden layers, then passed to a 2-layer LSTM for final predictions. Each LSTM layer has 160 units. We perform NID, black-box adversarial attacks, and then defend against them based on these models, as we detail next.

We consider NID in two different scenarios, namely *(i)* one-to-all detection and *(ii)* one-to-one detection. The one-to-all scenario groups all types of attacks into a single 'anomaly' class, which leads to a supervised binary classification problem. In contrast, one-to-one detectors separate each network attack (14 in total) into individual classes, and perform multi-class classification. In our study, the same neural network architectures are employed for both scenarios, except for changes in the final layers, as their number directly depends on the number of classes considered for identification. We train and validate all models using 80% of the dataset and test on 20% of it, as customary. All models are trained via minimizing the cross-entropy loss function through the Adam optimizer [27]. Super-sampling is employed to handle class imbalance between benign and malicious traffic, inherent to the dataset.

All models are trained and evaluated on a parallel computing cluster equipped with one or multiple Nvidia TITAN X, Tesla M40 or/and Tesla P100 GPUs. The neural models are implemented in Python using the TensorFlow [1] and TensorLayer [14] packages.

5.1 One-to-all NID Performance

We quantify the performance of the NIDS using four metrics, namely accuracy, precision, recall, and F1 score, as shown in Table 2. These metrics are frequently employed for evaluating binary classifiers.

Observe that all models achieve high detection performance, as all F1 scores are above 0.960. In addition, the three models considered perform similarly, since the difference between the F1 scores attained by each never exceeds 0.01. This matches the performance of state-of-the-art deep learning-based NID solutions, thus the models we use can be considered to be 'reliable'.

5.2 One-to-one NID Performance

One-to-one NIDS aim at classifying each traffic flow into 14 types of anomalies and benign. We employ the same neural networks and this time resort to normalized confusion matrices to assess their performance, as shown in Figure 4. The diagonal elements represent ratios of points for which the predicted label is equal to the true label, while off-diagonal elements indicate misclassification ratios [38]. Therefore, the elements of each row sum to 1. The higher the diagonal values in a confusion matrix, the higher the performance, indicating many correct predictions.

Table 2: The detection performance of MLP, CNN, and C-LSTM in the one-to-all scenario.

	Accuracy	Precision	Recall	F1 score
MLP	0.987	0.968	0.954	0.961
CNN	0.987	0.968	0.953	0.960
C-LSTM	0.987	0.967	0.952	0.960

Observe that all NID models achieve high detection accuracy for most types of anomalies, as diagonal values are close to 1. However, taking a closer look at the Brute Force-XSS, SQL Injection, Infiltration, and Brute Force-Web attacks, it appears the NID models tend to misclassify them as 'benign'. In addition, all DNNs face difficulties in dealing with DoS attack-SlowHTTPTest and FTP-Brute Force, as they mix them roughly 50/50. Further, the C-LSTM misclassifies

almost all DDoS attack-HOIC traffic as DDoS attack-LOIC-HTTP. This is perhaps less critical, since both attacks belong to DDoS category. Overall, the MLP, CNN, and C-LSTM attain 98.4%, 98.3%, and respectively 98.3% classification accuracy, which matches fairly closely the performance observed in the one-to-all scenario.

In what follows, we demonstrate that **although the NID solutions considered seem reliable in terms of detection accuracy, they can be easily compromised through a sequence of perturbations and queries, without requiring knowledge about the underlying models**.

6 ADVERSARIAL ATTACKS AGAINST NIDS

We consider five state-of-the-art black-box attack approaches, which we use to generate adversarial samples and compromise the pre-trained deep anomaly detectors discussed in Sec. 5. These include *(i)* Natural Evolution Strategies (NES) [24], *(ii)* Boundary Attack [5], *(iii)* Pointwise Attack [41], *(iv)* HopSkipJumpAttack [9], and *(v)* Opt-Attack [12], all of which were originally designed to compromise image classifiers. We quantify their performance in terms of different metrics and examine closely the role of different features in the adversarial sample generation process, as well as the decision mechanics of each NID model.

6.1 Black-box Adversarial Attack Methods

We first summarize the operation of each of the adversarial attack techniques we use against NIDS.

NES [24] are black-box gradient estimation methods for machine learning models. Estimated gradients can be used for projected gradient descent (as used in white-box attacks) to construct adversarial examples. This approach does not require a surrogate network, thus it is more query-efficient and reliable when crafting adversarial examples. Notably, NES work well in decision-based settings, which makes them suitable for attacks against NID models.

Boundary Attack [5] is a method that follows the decision boundary between adversarial and non-adversarial samples via rejection sampling. At each step, it employs constrained i.i.d. samples following a Gaussian distribution, starting from a large perturbation and successively reducing this until successful. This attack is highly flexible and can accommodate a set of adversarial criteria.

Pointwise Attack [41] is a simple decision-based attack method that greedily minimizes the L_0-norm between raw and adversarial samples. In image applications, it first introduces salt-and-pepper noise until misclassification, and then repeatedly iterates over each perturbed pixel, resetting it to the initial value if the perturbed image remains adversarial. We implement a similar approach to attack the NID models, but substitute the salt-and-pepper noise with additive Gaussian noise, to better suit network traffic.

HopSkipJumpAttack [9] is a hyperparameter-free, query-efficient attack method, which consists of three main steps: *(i)* estimation of the gradient direction, *(ii)* step-size search via geometric progression, and *(iii)* boundary search via a binary search approach. It is applicable to more complex settings, such as non-differentiable models or discrete input transformations, and achieves competitive performance against several defense mechanisms.

Opt-Attack [12] projects the decision-based attack into a continuous optimization problem and solves it via randomized zeroth-order gradient update. In particular, a Random Gradient-Free (RGF) method is employed to find appropriate perturbations and converge to stationary points. Since Opt-Attack does not rely on gradients, it can attack other non-differentiable classifiers besides neural networks, e.g., Gradient Boosting Decision Trees.

We employ a modified version of the mean absolute percentage error to quantify the deviation between each unmodified sample x and its adversarial version x_{adv}, i.e.,

$$\text{MAPE}(x, x_{\text{adv}}) = \frac{100\%}{N} \sum_{k=1}^{N} \left| \frac{x^{(k)} - x_{\text{adv}}^{(k)}}{x^{(k)}} \right|, \tag{1}$$

where N is the total number of perturbed features in x and $x^{(k)}$, $x_{\text{adv}}^{(k)}$ are the k^{th} features of the original and adversarial samples respectively. Smaller MAPE indicates higher similarity between the raw input x and the adversarial sample x_{adv}.

6.2 Attack Performance

We randomly select 50,000 malicious traffic flows from the test set, to craft adversarial samples. We quantify the performance of each attack approach using four performance metrics, namely Attack Success Rate (ASR), average benign confidence, MAPE, and average number of queries. The ASR is widely used to assess the effectiveness of adversarial attacks against DNNs [30] and is measured by the ratio between the number successful adversarial samples and the total attack attempts (in our case 50,000). An attack attempt is successful if and only if the underlying algorithm converges, and the adversarial samples meet the constraints discussed in Sec. 3.2. The average benign confidence denotes the probability that the model predicts an adversarial sample x_{adv} as benign. Higher confidence implies that the model is more confident about the decision made

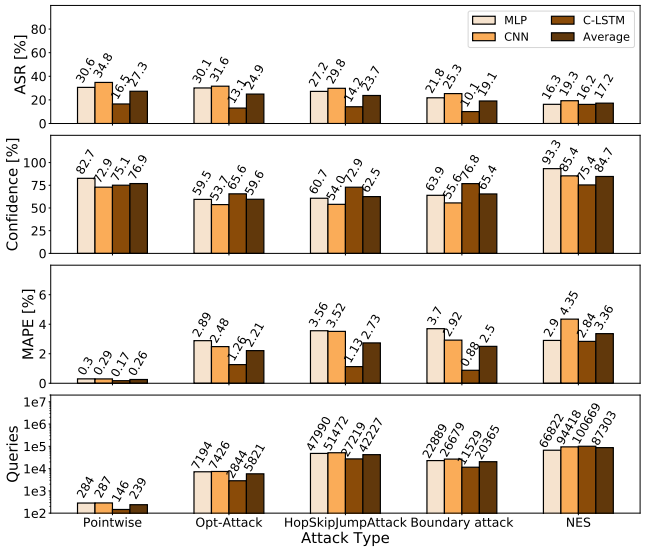

Figure 5: ASRs, Confidence, MAPE, and number of queries of all attack approaches against the three NID models considered, and their averages, in the one-to-one scenario.

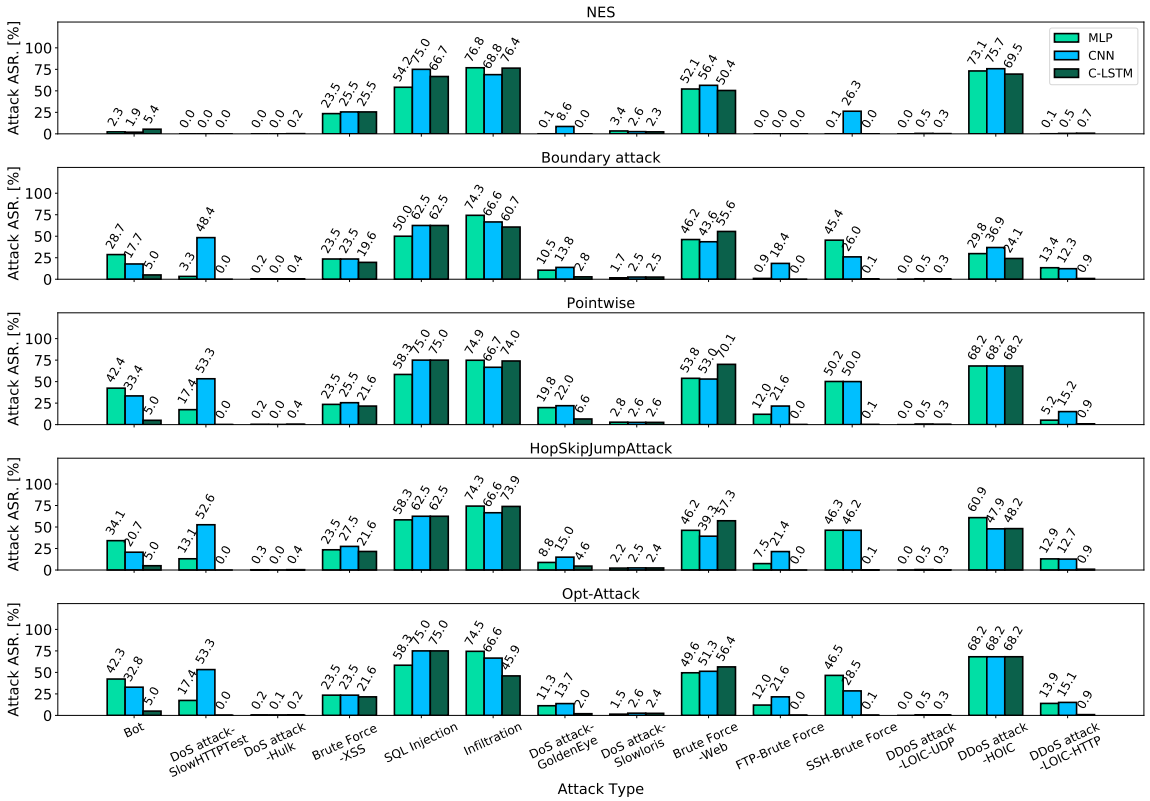

Figure 6: ASR with different types of attacks against all NID models considered in the one-to-one scenario.

over a sample. MAPE is defined in Eq. (1) and is computed over 22 features that allow perturbations. Recall that lower MAPE represent higher similarity between the raw and adversarial samples. The number of queries indicates how many attempts an attacker should perform in order to generate a successful adversarial sample. This can be used to measure the efficiency of an attack approach. Higher number of queries might trigger the NIDS, making the attack easier to be detected. Note that the MAPE, benign confidence, and number of queries are averaged over the successful attack attempts for each attack approach and NID model. All attacks are conducted using the original implementations and Python Foolbox [39].

Due to space constraints, in the rest of the paper we mainly focus on the *one-to-one* scenario.

Attack Performance in the One-to-one Scenario: We illustrate the statistics of each attack against the different NID models considered, in Figure 5. Observe that, except for the NES where the performance is similar among the different NID models, the ASR varies among models for all the other attack methods This is because the models work with large number of classes, which makes it difficult to craft adversarial samples to match the targeted 'benign' label. The Pointwise method obtains the highest ASR, lowest MAPE, and lower average number of queries. This suggests that this approach is effective and efficient in one-to-one settings. The C-LSTM appears to be the most robust model against adversarial samples, as all attack methods attain the lowest ASR values against this NID model. Although achieving the highest benign confidence with adversarial samples, NES obtain the lowest ASR on average. In general, they also require more queries to craft an adversarial sample.

In Figure 6, we show the ASR for each type of malicious traffic flow considered, in the same one-to-one scenario. Analyzing these results jointly with Figure 4, observe that anomalies with low detection rate (i.e., Brute Force-XSS, SQL Injection, Infiltration, Brute Force-Web) are easier to be disguised by attackers. This is because the models already have vague decision boundaries for these flow types, thus are easier to be gamed. Attackers obtain the lowest ASR when crafting adversarial samples based on DoS attacks-Hulk, -GoldenEye, -Slowloris, and DDoS attack-LOIC-UDP, as the NID models exhibit high detection rates over these anomalies.

6.3 Adversarial Samples Analysis

6.3.1 Feature-wise MAPE. We delve deeper into the adversarial samples generated, by showing in Figure 7 the average MAPE of each perturbed feature on all successful attack samples, across all NID models and attack approaches. Observe that for both detection scenarios, the Active/Idle Time (i.e., the time a flow was idle before becoming active and amount of time a flow was active before becoming active) are less affected, as the related features remain almost unchanged in the attack process. In contrast, those **features that characterize the average number of bytes and packets sent in the forward and backward directions in the initial window or/and sub flows, are perturbed more significantly**. This indicates that these features are the most influential in the decision of NID models, and therefore **more likely to be exploited by potential attackers**.

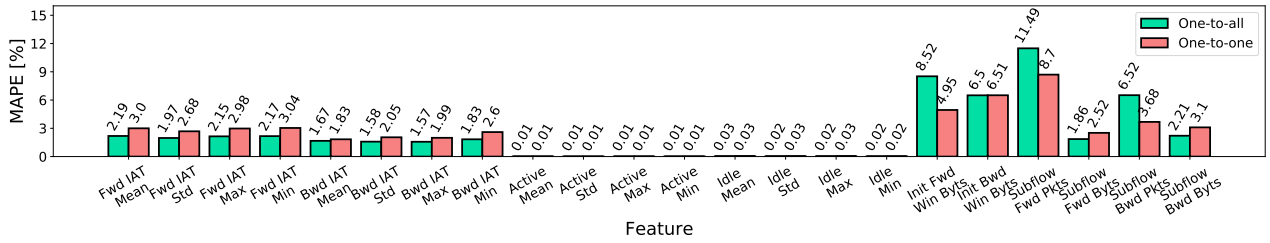

Figure 7: The MAPE between original and adversarial samples of each feature that allows perturbations.

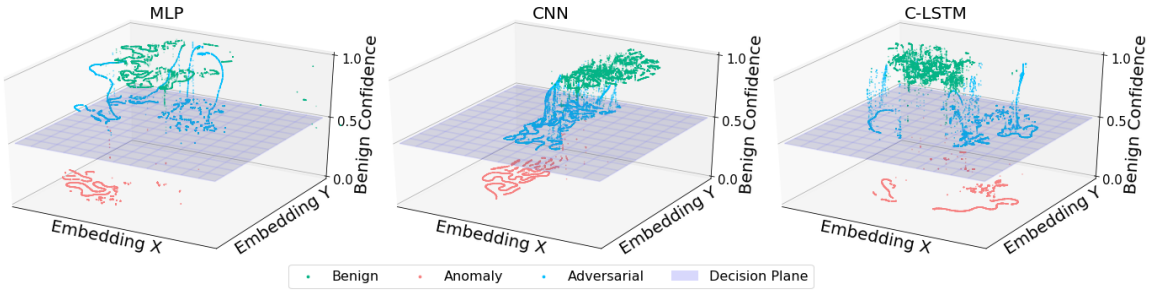

Figure 8: Two-dimensional (x, y axes) t-SNE embedding of the representations in the last hidden layer, along with the benign confidence (z axis) of each NID model. Data generated using 10,000 benign samples (blue), 10,000 adversarial samples (green) produced by all attack methods, and their corresponding original malicious samples (pink).

6.3.2 t-SNE Visualization. We also investigate the inner workings of each NID model, by visualizing the output embedding of their hidden layers, so as to understand better how a neural network 'thinks' of the benign, malicious, and adversarial samples. To this end, we adopt the t-distributed Stochastic Neighbor Embedding (t-SNE) [32] to reduce the dimension of the last hidden layer of each model to 2. In Figure 8, we plot the t-SNE embedding (x, y axes) of the hidden representations of 10,000 benign samples (blue), 10,000 adversarial samples (green) generated by each attack method, and their corresponding anomalous samples used to craft them (pink), along with their benign confidence (z axis). Note that a sample will be considered benign iff the benign confidence is greater than 0.5 (above the decision plane). Typically, the t-SNE approach organizes data points that have higher similarity into nearby embeddings [55]. It can therefore reflect how the model 'thinks' of the data samples, as similar data representations will be clustered together.

Observe that anomalous samples can be clearly distinguished from benign samples by their t-SNE embeddings for all NID models. The purpose of adversarial attacks is to cause misclassification by bringing malicious samples across the decision boundary. This is reflected in Figure 8, as the t-SNE embedding of adversarial samples are moved closer to the benign embedding cluster, while they remain anomalies in nature. This successfully confuses the NID models, making the adversarial samples indistinguishable. In addition, adversarial samples with higher benign confidence are in general closer to the benign embedding cluster.

7 DEFENDING AGAINST ADVERSARIAL ATTACKS

Defense mechanisms against adversarial attacks should improve the robustness of deep learning models to adversarial samples, such that they become less likely to be compromised and the ASR of different attacks is reduced. In general, countermeasures for adversarial examples can be categorized into two types [54]: *(i)* **Reactive** – detecting adversarial examples after DNNs have been trained; and *(ii)* **Proactive** – improving the robustness of DNN models against adversarial examples. In this paper, we propose three different defense mechanisms, and combine them to counteract the adversarial samples generated by the black-box attack methods discussed in the previous section. These defense mechanisms include:

(1) **Model Voting Ensembling (Proactive):** Ensembling pre-trained MLP, CNN, and C-LSTM using a voting mechanism, to construct stronger models that are less susceptible to misclassification of adversarial samples;
(2) **Ensemble Adversarial Training (Proactive):** Augmenting the training dataset with adversarial samples, and retraining the NID models, thereby reinforcing their capabilities against adversarial samples;
(3) **Adversarial Query Detection (Reactive):** Detecting the query process in the black-box attack process, so as to blacklist the attacker's IP address before they may succeed.

In what follows, we detail the proposed defense mechanisms, and demonstrate their effectiveness against adversarial attacks.

7.1 Model Voting Ensembling

The experiments we reported in Sec. 6 suggest that an attacker can successfully compromise a NID model with up to 35% ASR. However, only a small set of adversarial samples can bypass all three NID models simultaneously. This motivates us to construct a new ensembling model [56, 57] by combining all of these structures, to strengthen the barrier against potential attacks. Specifically, for each input traffic flow, we gather the decisions of all NID models individually, and make the classification using a voting process. A

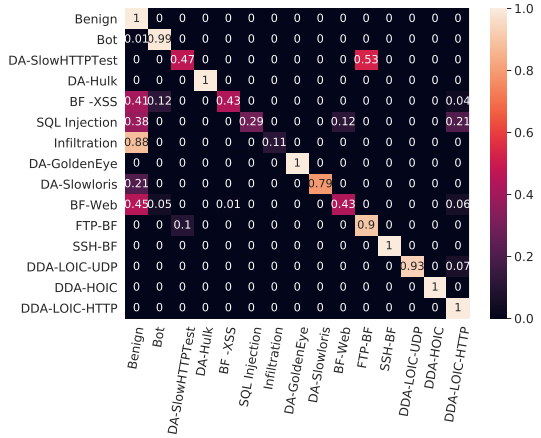

Figure 9: The confusion matrix of the ensembling model in the one-to-one detection scenario.

Table 3: One-to-all NID performance of ensembling model.

Model Ensembling	Accuracy	Precision	Recall	F1 score
	0.987	0.964	0.954	0.959

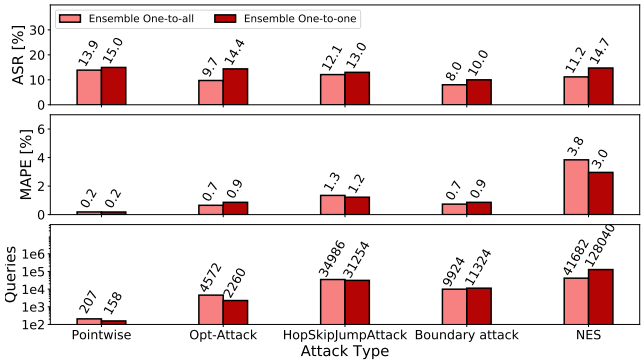

Figure 10: ASRs, MAPE, and number of queries statistics of all attack approaches against ensembling models in one-to-all and one-to-one scenarios.

flow will be classified as 'benign' iff all models reach consensus, i.e., all of them classify it as 'benign'. Otherwise, the traffic flow will be regarded as an 'anomaly'. We recognize several advantages of using such model voting ensembling as means of defense, namely:

(1) In order to construct a successful adversarial sample, attackers need to defeat all NID models simultaneously, which is much harder than compromising a single one;
(2) The voting mechanism makes the entire model non-differentiable, thus attack approaches that rely on model gradient estimation (e.g., NES) will be obstructed;
(3) The voting mechanism is easy to implement, as it does not require to re-train the original NID models.

The proposed model voting ensembling method is a proactive approach, as it improves the robustness of the pretrained models against adversarial samples. We show the NID performance of the ensembling model for the one-to-all scenario in Table 3 and the confusion matrix for the one-to-one scenario in Figure 9. Revisiting Table 2 and Figure 4, observe that **the ensembling model obtains very close performance compared to its individual components in both detection scenarios, while achieving lower false positive rates**, since it requires consensus to make the 'benign' classification decision.

We re-run the same five black-box attacks considered previously over the same set of 50,000 malicious samples and show statistics of their performance in Figure 10. Note that the benign confidence measure is abandoned, since the outputs of ensembling models are no longer probabilities. Jointly analyzing these results with Figure 5, observe that the ASR of each attack approach against the ensembling model has dropped relatively to when attacking each of the model's component (i.e., MLP, CNN, and LSTM). In the best case, the Boundary approach obtains 20.1% ASR in attacking the CNN-based NIDS in the one-to-all scenario, while its success rate is merely 8.0% when attacking the ensembling model. Regarding the one-to-one scenario, the reduction in ASR is also substantial. This indicates that the voting ensembling mechanism is an effective defense approach. Turning attention to MAPE, observe that adversarial samples crafted against ensembling models yield low MAPE, which suggests that **this defense mechanism applies hidden and tighter constraints to the adversarial samples, to prevent them from deviating excessively from the raw input samples, which in turn improves detection**.

We also show in Figure 11 the ASR on a malicious traffic type basis, when crafting adversarial samples against the ensembling model, for the one-to-one NID scenario. Observe that the voting ensembling mechanism successfully defends 9 type of adversarial samples, i.e., Bot, DoS attack-SlowHTTPTest, DoS attack-Hulk, DoS attack-GoldenEye, DoS attack-Slowloris, FTP-BruteForce, SSH-Bruteforce, DDoS attack-LOIC-UDP and DDoS attack-LOIC-HTTP, as their ASR is virtually 0%. **For attacks such as botnet this is critical, since any success could have catastrophic effects, which our ensembling technique thwarts.** For other types of malicious traffic, the ASR also drops by varying degrees, but not as significant, which calls for further defenses, as we show next.

7.2 Ensemble Adversarial Training (EAT)

As discussed in Sec. 3, white-box strategies are not commonly accessible to external adversaries seeking to compromise NIDS, as the training data, model structures and parameters are opaque. However, recent literature confirms that adversarial samples are adaptable across different attack methods and victim models [22, 45, 50]. Therefore, from the defenders' points of view, adversarial samples generated using white-box attacks can be exploited to improve the robustness of NID models, so as to defend against potential adversarial samples. Therefore, we employ the Ensemble Adversarial Training (EAT) as an additional defense approach [45], which augments the training data with adversarial examples generated by white-box attacks crafted on other static pre-trained NID models. Subsequently, the original NID models are reinforced by re-training on the augmented dataset. We expect that, with the proposed re-training, the NID models learn to classify adversarial samples better and thus become more resilient to attacks.

7.2.1 Reinforcing NID models with White-box Adversarial Samples. We randomly select 250,000 malicious flows to generate adversarial

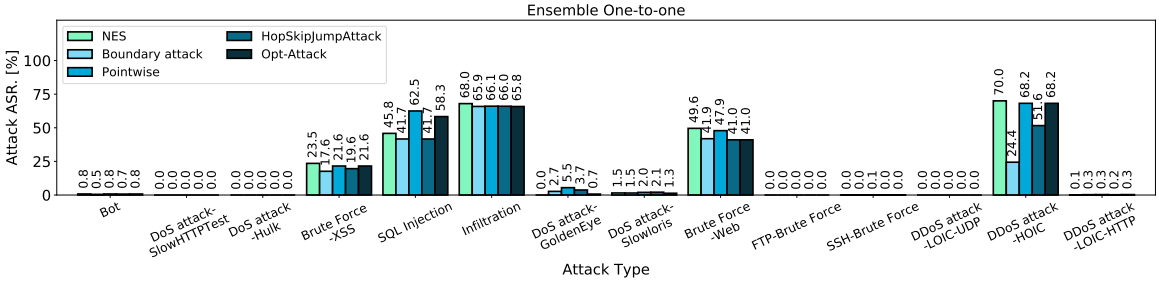

Figure 11: ASRs of each type of attack against the model voting ensemble technique in the one-to-one scenario.

samples using three state-of-the-art white-box attack approaches: Fast Gradient Sign Method [18], Iterative Attack (I-FGSM) [29] and Momentum Iterative Fast Gradient Sign Method (MI-FGSM) [15]. The FGSM-based approaches perform one step gradient update along the direction the gradient at each feature that allows perturbations, and introduce noise following that direction. I-FGSM extends the FGSM by running a finer optimization for multiple iterations to generate a valid adversarial sample. MI-FGSM introduces a momentum term into the iterative process of I-FGSM, which helps stabilizing the update directions and escaping from poor local maxima. This leads to more transferable adversarial samples.

Due to the *information asymmetry* between attackers and defenders, the defenders do not have knowledge about which features will be perturbed for attack purposes. We therefore relax the feature constraints (see Sec. 3.2) for perturbations in the white-box setting. However, the constraints over MAPE ($\leq 20\%$) are retained, to restrict the scale of the perturbations. Note that the adversarial samples generated by white-box attacks are not necessarily valid traffic flows, as they are only employed for training purposes. We gather successful adversarial samples generated by all white-box attack methods (i.e., FGSM, I-FGSM, and MI-FGSM), crafted with all NID models (i.e., MLP, CNN, and C-LSTM) in both detection scenarios (i.e., one-to-all and one-to-one) and combine these with the original training data, to build an augmented dataset for EAT.

We show the performance of each white-box attack in Figure 12. Observe that since the NID models are transparent, and looser constraints are applied to the adversarial samples, the ASR for all white-box attacks is significantly higher than their black-box counterparts. White-box attacks also require fewer queries to generate adversarial samples. Fortunately, attackers normally do not have access to the NID models.

7.2.2 NID Performance of Post-EAT Models. In Table 4, we report the detection performance on the same test set after EAT, for the one-to-all scenario. Compared to NID models prior to the EAT (See Tables 2 and 3), the detection performance of the newly trained models has dropped slightly in terms of accuracy, precision, and F1 score. However, the recall rate of each model has improved. This indicates that the models are prone to classifying some ambiguous samples as anomalies, which results in higher false positive and lower false negative rates. Similar phenomena are also observed in the one-to-one scenario. The accuracy for the MLP, CNN, C-LSTM, and the ensembling model appears worse than what was achieved prior to EAT. However, by taking a closer look at their confusion matrices in Figure 13, **post-EAT models achieve high detection rates on most anomalies they failed to detect previously** (i.e., Brute Force-XSS, SQL Injection, and Brute Force-Web). This suggests that EAT has improved the robustness of NID models, making them more sensitive to anomalous traffic that is difficult to classify.

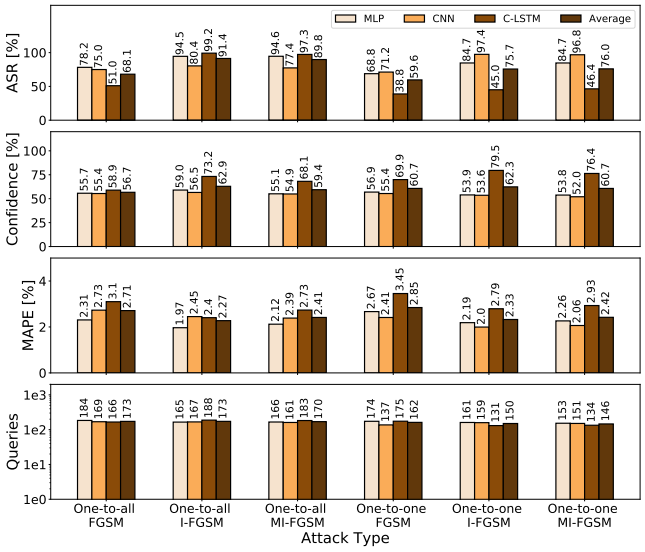

Figure 12: ASRs, Confidence, MAPE, and number of queries statistics of all white-box attack approaches against the three NID models considered in this study, and their average values for both NID scenarios.

Table 4: Performance of MLP, CNN, C-LSTM, and ensembling model after EAT in the one-to-all scenario.

Model	Accuracy	Precision	Recall	F1 score
MLP	0.987	0.968	0.953	0.960
CNN	0.986	0.959	0.954	0.956
C-LSTM	0.985	0.953	0.955	0.954
Ensembling	0.983	0.943	0.956	0.949

Table 5: Ratio of adversarial samples that can bypass each NID model after EAT.

Scenario	MLP	CNN	C-LSTM	Ensembling
One-to-all	40.04%	53.15%	48.43%	81.54%
One-to-one	42.45%	38.06%	38.26%	43.31%

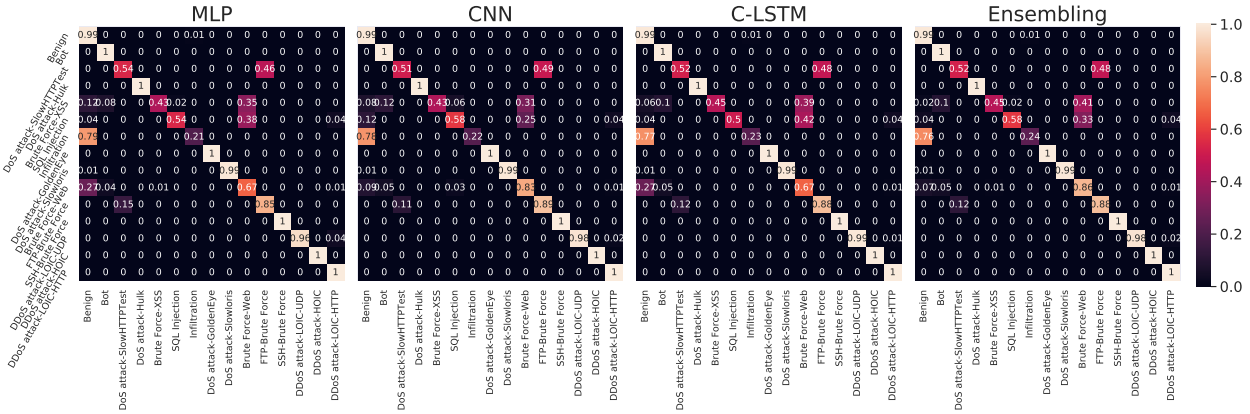

Figure 13: Confusion matrices of the MLP, CNN, C-LSTM, and ensembling model in the one-to-one NID scenario, after EAT.

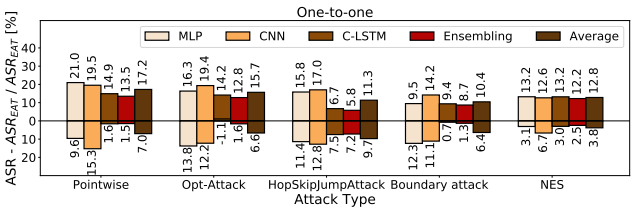

Figure 14: ASR of each attack after EAT (bars above x axis), and ASR reduction compared to when no further defense is applied (bars below x axis) for the one-to-one NID scenario.

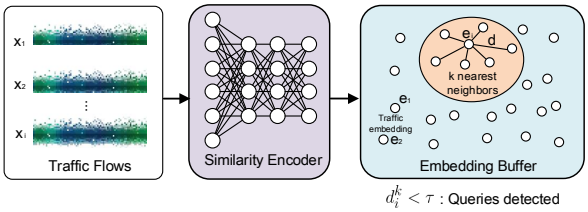

Figure 15: An illustration of the query detection defense mechanism using a deep similarity encoder.

7.2.3 Robustness to Old Adversarial Samples. In Table 5, we further show the ratio of adversarial samples crafted from the models before EAT, which can compromise the corresponding post-EAT models. Observe that EAT also makes each model more resilient to old adversarial samples, as those ratios are significantly below 100%. In particular, only 38.06% of adversarial samples crafted from the old C-LSTM can bypass the Ensemble Adversarial Trained C-LSTM. This means that EAT enables each model to learn to characterize adversarial samples generated using white-box attacks, and therefore fixes some 'bugs' present in the old setting.

7.2.4 The Effect of EAT. In Figure 14, we show the ASR for each attack after EAT (ASR_{EAT}, bars in the upper part of the plot), and the ASR reduction compared to the case before EAT was applied (ASR - ASR_{EAT}, bars in the lower part) for the one-to-one NID scenario. In the figure, positive numbers below the x-axis indicate that the ASR_{EAT} has dropped after EAT was employed. We observe that ASR of each attack drops for most of the models. This means that **EAT successfully improves the robustness of each model, making them more difficult to be compromised**. On average, the ASR drops to 6.70% and 5.78% for one-to-one and one-to-all NID scenarios, respectively. **This is particularly beneficial to practically mitigating DoS and brute-force type attacks.**

7.3 Adversarial Query Detection

Recall that all black-box attack methods rely on continuous queries to the victim model and feedback received. Based on the feedback, the attackers learn to adjust the perturbations added to the input, so as to compromise detection. The scale of perturbations is usually small, so that they to do not change the essence of the original input. Therefore, the queries in the same attack round are typically with high similarity. **This inherent similarity between queries can be harnessed to detect an attack**. Therefore, we explore query detection [11] as the final defense mechanism. Once queries have been discovered, the NIDS can blacklist the attackers' IP addresses, to prevent potential threats.

Specifically, for each IP address, we construct a buffer with size B to store the features of the traffic flows originating from that address in a pre-defined period. To reduce the dimension of the features saved and model the similarity degree between flows, we employ a deep similarity encoder (DSE) [4], encoding similar traffic flows in a lower-dimensional space with shorter l_2 distance. More precisely, for each new flow x sent from a given IP address, we compute the pairwise distance between the embedding of this flow and others in the buffer, calculating the k nearest neighbor average distance d_x^k. If d_x^k is lower than a threshold τ, i.e., $d_x^k < \tau$, this suggests that that IP address has sent an excessive number of similar traffic flows, which can be considered as queries in an ongoing attack. When this happens, the IP address can be blacklisted and thus the potential threat eliminated. We show the underlying principle of the query detection mechanism in Figure 15.

After an attack is detected, the buffer associated to the specific IP address can be cleared. In addition, when query detection suggests a potentially malicious actor, their IP address can be banned either immediately, or after subsequent queries, as suggested in [4]. This can minimize an attacker's knowledge of the time when their attack was detected, therefore reducing the probability of compromising the query detection mechanism.

7.3.1 Deep Similarity Encoder. The core component of the query detection-based defense mechanism is the deep similarity encoder

(DSE) [4], which is a neural network that reduces the dimension of the input. After embedding by a DSE, dissimilar flows will be far from each other in the encoded space, while similar queries will be close. Thus, queries and traffic flows become more distinguishable.

For the DSE, we employ a CNN similar to that in Figure 3, only replacing the last layer with 3 units. This means that the embedding of each traffic flow is a 3-dimensional vector. We denote $e_i = \text{DSE}(x_i)$ as the embedding of the input sample x_i. The DSE can be trained via minimizing the following contrastive loss function:

$$L(x_i, \tilde{x}_i, x_m, x_n; \theta) = ||e_i - \tilde{e}_i||_2^2 + \max(0, \varpi^2 - ||e_m - e_n||_2^2). \quad (2)$$

Here, $x_i, \tilde{x}_i$ are a pair of similar traffic flows, while x_m, x_n are traffic flows that are dissimilar. θ is the trainable parameter set of the DSE, and ϖ is a constant penalty, which regularizes the scale of $||e_m - e_n||_2^2$. We choose $\varpi = 0.5$ in our experiments. The first term of the function ensures that the l_2 distances of the similar traffic flows are minimized, while the second term guarantees that distances of dissimilar traffic pairs are maximized but limited to ϖ.

We train the DSE using the same training set sampled from the CSE-CIC-IDS2018 dataset as used by other NID models. For the purpose of training, we construct $\tilde{x}_i$ by adding Gaussian noise $\sigma_i \sim N(0, \alpha|x_i|)$ to each sample x_i, i.e., $\tilde{x}_i = x_i + \sigma_i$. Here, α controls the standard deviation of the Gaussian noise and we choose $\alpha = 0.15$. x_m, x_n are sampled from a training set distinct from x_i. After training, we use the full training set to randomly generate 13,153,902 pairs of similar and dissimilar flows.

7.3.2 Hyper-parameters Selection. There are three important hyper-parameters to be configured for query detection, namely *(i)* the detection threshold τ; *(ii)* the number of neighbors k used for detection; and *(iii)* the size of the buffer B, which stores the traffic flows sent from the same IP address. These parameters will significantly affect the performance of the query detection. First, we select $\tau = 0.00157$, since 10% of dissimilar pairs and 86.4% of similar pairs in the training set are below this threshold. This provides an appropriate decision boundary to discriminate normal traffic flows and attack queries. The values of k and B affect the robustness of the detection and also the computational and storage cost of the NIDS. We select $B = 500$ and $k = 200$, as these numbers allow efficient detection and yield 0 false positive rates when operating with traffic streams simulated with the entire training set.

7.3.3 Query Detection Defense Performance. In Figure 16 we show the ASR of each attack after the query detection (bars above the x-axis in the top sub-plot), the ASR reduction compared to when query detection is not employed (bars below the x-axis in the top sub-plot), and the average number of queries (bottom) when the attack is detected, for each attack method in the one-to-one NID scenario. Observe that the ASR has dropped significantly after the query detection was employed. In particular, the ASR of NES reaches 0 for all models, and the detection rate thus become 100%. On average, the query detection defense reduces the ASR of attacks by 8.56% and 12.38% in the one-one-to one and one-to-all scenario, respectively. Effectively, **the majority of the adversarial attack are detected during their query process.**

Taking a closer look at the average number of detected queries, we observe that NES, Boundary, Pointwise, and HopSkipJumpAttack attack attempts are detected at their 201st query. Recall that

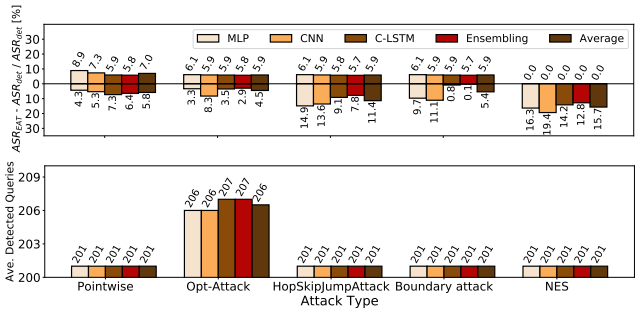

Figure 16: Performance statistics of query detection defense in the one-to-one scenario. Top: ASR of each attack after query detection (bars above x axis) and ASR reduction compared to when the query detection is removed (below x axis). Bottom: average number of queries when the query detector is triggered.

the k neighborhood size selected for query detection is 200, hence the detection alarm will only be triggered when the buffer has more than 200 samples. This means that the attack is detected immediately after the buffer has k neighbor samples. Regarding the Opt-Attack attack, this is detected always within 208 queries. This is due to the initial phase of the attack, when it injects a few benign traffic flows to learn the direction of perturbation to be added to the adversarial samples. These samples are normally dissimilar, which slightly increases the detection time. Note that, despite the efficiency of the query detection mechanism, a larger buffer size ($B = 500$) is still needed for tolerance, as the attacks may fill the buffer with queries (similar samples) and garbled traffic (dissimilar samples) alternately, to compromise the defense.

Overall, **by combining the model voting ensembling, EAT, and query detection mechanisms, our proposal can successfully prevent five mainstream black-box adversarial attacks from compromising deep learning-based NIDS**.

8 CONCLUSIONS

In this paper, we introduced Tiki-Taka, a framework for defending against adversarial attacks on deep learning-based NIDS. We trained three state-of-the-art deep learning models (MLP, CNN, and C-LSTM) on a publicly available dataset, then employed 5 classes of decision-based adversarial attacks to compromise the neural models. Experiments show that despite having high detection rates, deep learning-based NIDS are vulnerable to adversarial samples. To strengthen NIDS against such threats, we proposed three defense methods: model voting ensembling, ensembling adversarial training, and query detection. To our knowledge, these are the first defense mechanisms to be proposed against adversarial attacks targeting NIDS. Their combined use can reduce success rates of all attacks considered, bringing detection close to 100% on most malicious traffic and fending off particularly critical malicious traffic such as botnet and DoS. Future work will focus on handling infiltration traffic, which appears more resilient to NID models and defense approaches.

REFERENCES

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, et al. 2016. TensorFlow: A System for Large-Scale Machine Learning. In *OSDI*.

[2] Mahdieh Abbasi and Christian Gagné. 2017. Robustness to adversarial examples through an ensemble of specialists. In *Workshop in ICLR*.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *ICML*.

[4] Sean Bell and Kavita Bala. 2015. Learning visual similarity for product design with convolutional neural networks. *ACM Transactions on Graphics (TOG)* 34, 4 (2015), 98.

[5] Wieland Brendel, Jonas Rauber, and Matthias Bethge. 2018. Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models. In *ICLR*.

[6] Anna L Buczak and Erhan Guven. 2015. A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials* 18, 2 (2015), 1153–1176.

[7] Nicholas Carlini and David Wagner. 2017. Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods. In *Proc. ACM Workshop on Artificial Intelligence and Security*.

[8] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *IEEE S&P*.

[9] Jianbo Chen, Michael I Jordan, and Martin J Wainwright. 2020. HopSkipJumpAttack: A Query-Efficient Decision-Based Attack. In *IEEE S&P*.

[10] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. 2017. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proc. ACM Workshop on Artificial Intelligence and Security*.

[11] Steven Chen, Nicholas Carlini, and David Wagner. 2019. Stateful Detection of Black-Box Adversarial Attacks. *arXiv preprint arXiv:1907.05587* (2019).

[12] Minhao Cheng, Thong Le, Pin-Yu Chen, Huan Zhang, JinFeng Yi, and Cho-Jui Hsieh. 2019. Query-Efficient Hard-label Black-box Attack: An Optimization-based Approach. In *ICML*.

[13] Kenneth T. Co, Luis Muñoz González, Sixte de Maupeou, and Emil C. Lupu. 2019. Procedural Noise Adversarial Examples for Black-Box Attacks on Deep Convolutional Networks. In *ACM CCS*.

[14] Hao Dong, Akara Supratak, Luo Mai, Fangde Liu, Axel Oehmichen, Simiao Yu, and Yike Guo. 2017. TensorLayer: A Versatile Library for Efficient Deep Learning Development. In *Proc. ACM Multimedia*.

[15] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. 2018. Boosting adversarial attacks with momentum. In *IEEE CVPR*.

[16] Osama Faker and Erdogan Dogdu. 2019. Intrusion Detection Using Big Data and Deep Learning Techniques. In *Proc. ACM Southeast Conference*.

[17] Forbes. 2020. Cyberwarfare Will Explode In 2020 (Because It's Cheap, Easy And Effective).

[18] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *ICLR*.

[19] Chuan Guo, Jacob Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Weinberger. 2019. Simple Black-box Adversarial Attacks. In *ICML*.

[20] Douglas Heaven. 2019. Why deep-learning AIs are so easy to fool. *Nature* 574 (2019), 163–166.

[21] Sanghyun Hong, Pietro Frigo, Yigitcan Kaya, Cristiano Giuffrida, and Tudor Dumitras. 2019. Terminal Brain Damage: Exposing the Graceless Degradation in Deep Neural Networks Under Hardware Fault Attacks. In *USENIX Security*.

[22] R Huang, B Xu, D Schuurmans, and C Szepesvári. [n.d.]. Learning with a strong adversary. 2015.

[23] Olakunle Ibitoye, Rana Abou-Khamis, Ashraf Matrawy, and M. Omair Shafiq. 2019. The Threat of Adversarial Attacks on Machine Learning in Network Security – A Survey. arXiv:cs.CR/1911.02621

[24] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. 2018. Black-box Adversarial Attacks with Limited Queries and Information. In *ICML*. 2142–2151.

[25] Ahmad Javaid, Quamar Niyaz, Weiqing Sun, and Mansoor Alam. 2016. A deep learning approach for network intrusion detection system. In *Proc. EAI Intl Conference on Bio-inspired Information and Communications Technologies*. 21–26.

[26] Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. 2017. Reluplex: An efficient SMT solver for verifying deep neural networks. In *International Conference on Computer Aided Verification*.

[27] Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *ICLR* (2015).

[28] Aditya Kuppa, Slawomir Grzonkowski, Muhammad Rizwan Asghar, and Nhien-An Le-Khac. 2019. Black Box Attacks on Deep Anomaly Detectors. In *Proc. ACM International Conference on Availability, Reliability and Security*. 21.

[29] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2017. Adversarial examples in the physical world. *ICLR Workshop* (2017).

[30] Yandong Li, Lijun Li, Liqiang Wang, Tong Zhang, and Boqing Gong. 2019. NATTACK: Learning the Distributions of Adversarial Examples for an Improved Black-Box Attack on Deep Neural Networks. In *ICML*.

[31] Jiajun Lu, Theerasit Issaranon, and David Forsyth. 2017. Safetynet: Detecting and rejecting adversarial examples robustly. In *IEEE ICCV*.

[32] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[33] Dongyu Meng and Hao Chen. 2017. Magnet: a two-pronged defense against adversarial examples. In *ACM CCS*.

[34] Erxue Min, Jun Long, Qiang Liu, Jianjing Cui, and Wei Chen. 2018. TR-IDS: Anomaly-based intrusion detection through text-convolutional neural network and random forest. *Security and Communication Networks* (2018).

[35] Seungyong Moon, Gaon An, and Hyun Oh Song. 2019. Parsimonious Black-Box Adversarial Attacks via Efficient Combinatorial Optimization. In *ICML*.

[36] Milad Nasr, Reza Shokri, and Amir Houmansadr. 2019. Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning. In *IEEE S&P*. 0.

[37] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. 2016. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE S&P*.

[38] David Martin Powers. 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. (2011).

[39] Jonas Rauber, Wieland Brendel, and Matthias Bethge. 2017. Foolbox: A Python toolbox to benchmark the robustness of machine learning models. *arXiv preprint arXiv:1707.04131* (2017). arXiv:1707.04131

[40] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. 2018. Defense-GAN: Protecting classifiers against adversarial attacks using generative models. In *ICML*.

[41] Lukas Schott, Jonas Rauber, Matthias Bethge, and Wieland Brendel. 2019. Towards the first adversarially robust neural network model on MNIST. In *ICLR*.

[42] Iman Sharafaldin, Arash Habibi Lashkari, and Ali A Ghorbani. 2018. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. In *ICISSP*.

[43] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. 2019. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation* (2019).

[44] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A Ghorbani. 2009. A detailed analysis of the KDD CUP 99 data set. In *IEEE Symposium on Computational Intelligence for Security and Defense Applications*.

[45] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. 2018. Ensemble Adversarial Training: Attacks and Defenses. In *ICLR*.

[46] Muhammad Usama, Junaid Qadir, Ala Al-Fuqaha, and Mounir Hamdi. 2019. The Adversarial Machine Learning Conundrum: Can The Insecurity of ML Become The Achilles' Heel of Cognitive Networks? *arXiv preprint arXiv:1906.00679* (2019).

[47] R Vinayakumar, Mamoun Alazab, KP Soman, Prabaharan Poornachandran, Ameer Al-Nemrat, and Sitalakshmi Venkatraman. 2019. Deep Learning Approach for Intelligent Intrusion Detection System. *IEEE Access* 7 (2019), 41525–41550.

[48] R Vinayakumar, KP Soman, and Prabaharan Poornachandran. 2017. Applying convolutional neural network for network intrusion detection. In *IEEE International Conference on Advances in Computing, Communications and Informatics*.

[49] Zheng Wang. 2018. Deep learning-based intrusion detection with adversaries. *IEEE Access* 6 (2018), 38367–38384.

[50] Yi Wu, David Bamman, and Stuart Russell. 2017. Adversarial training for relation extraction. In *Proc. Empirical Methods in Natural Language Processing*.

[51] Kaichen Yang, Jianqing Liu, Chi Zhang, and Yuguang Fang. 2018. Adversarial Examples Against the Deep Learning Based Network Intrusion Detection Systems. In *IEEE MILCOM*.

[52] Yuanshun Yao, Huiying Li, Haitao Zheng, and Ben Y. Zhao. 2019. Latent Backdoor Attacks on Deep Neural Networks. In *ACM CCS*.

[53] Chuanlong Yin, Yuefei Zhu, Jinlong Fei, and Xinzheng He. 2017. A deep learning approach for intrusion detection using recurrent neural networks. *IEEE Access* 5 (2017), 21954–21961.

[54] Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. 2019. Adversarial examples: Attacks and defenses for deep learning. *IEEE Transactions on Neural Networks and Learning Systems* (2019).

[55] Chaoyun Zhang, Rui Li, Woojin Kim, Daesub Yoon, and Paul Patras. 2018. Driver Behavior Recognition via Interwoven Deep Convolutional Neural Nets with Multi-stream Inputs. *arXiv preprint arXiv:1811.09128* (2018).

[56] Chaoyun Zhang, Xi Ouyang, and Paul Patras. 2017. ZipNet-GAN: Inferring fine-grained mobile traffic patterns via a generative adversarial neural network. In *ACM CoNEXT*.

[57] Chaoyun Zhang and Paul Patras. 2018. Long-term mobile traffic forecasting using deep spatio-temporal neural networks. In *ACM MobiHoc*.

[58] Chaoyun Zhang, Paul Patras, and Hamed Haddadi. 2019. Deep learning in mobile and wireless networking: A survey. *IEEE Comms Surveys & Tutorials* (2019).

[59] Yong Zhang, Xu Chen, Da Guo, Mei Song, Yinglei Teng, and Xiaojuan Wang. 2019. PCCN: Parallel Cross Convolutional Neural Network for Abnormal Network Traffic Flows Detection in Multi-Class Imbalanced Network Traffic Flows. *IEEE Access* 7 (2019), 119904–119916.

[60] Yong Zhang, Xu Chen, Lei Jin, Xiaojuan Wang, and Da Guo. 2019. Network Intrusion Detection: Based on Deep Hierarchical Network and Original Flow Data. *IEEE Access* 7 (2019), 37004–37016.

[61] Qianru Zhou and Dimitrios Pezaros. 2019. Evaluation of Machine Learning Classifiers for Zero-Day Intrusion Detection–An Analysis on CIC-AWS-2018 dataset. *arXiv preprint arXiv:1905.03685* (2019).