

# TAD: Transfer Learning-based Multi-Adversarial Detection of Evasion Attacks against Network Intrusion Detection Systems

Islam Debicha<sup>a,b,\*</sup>, Richard Bauwens<sup>a</sup>, Thibault Debatty<sup>b</sup>, Jean-Michel Dricot<sup>a</sup>, Tayeb Kenaza<sup>c</sup>, Wim Mees<sup>b</sup>

<sup>a</sup> Cybersecurity Research Center, Université Libre de Bruxelles, 1050 Brussels, Belgium

<sup>b</sup>Cyber Defence Lab, Royal Military Academy, 1000 Brussels, Belgium

<sup>c</sup>Computer Security Laboratory, Ecole Militaire Polytechnique, Algiers, Algeria

## Abstract

Nowadays, intrusion detection systems based on deep learning deliver state-of-the-art performance. However, recent research has shown that specially crafted perturbations, called adversarial examples, are capable of significantly reducing the performance of these intrusion detection systems. The objective of this paper is to design an efficient transfer learning-based adversarial detector and then to assess the effectiveness of using multiple strategically placed adversarial detectors compared to a single adversarial detector for intrusion detection systems. In our experiments, we implement existing state-of-the-art models for intrusion detection. We then attack those models with a set of chosen evasion attacks. In an attempt to detect those adversarial attacks, we design and implement multiple transfer learning-based adversarial detectors, each receiving a subset of the information passed through the IDS. By combining their respective decisions, we illustrate that combining multiple detectors can further improve the detectability of adversarial traffic compared to a single detector in the case of a parallel IDS design.

**Keywords:** Intrusion detection system, Machine learning, Evasion attacks, Adversarial detection, Transfer learning, Data fusion.

## 1. Introduction

With the growth of information communication and in particular the growth of the Internet, the concern to secure this information and related systems has become increasingly important. Every day, massive amounts of sensitive information are created, transferred, stored, or updated around the world. This sensitive information can range from personal emails to banking transactions, from a simple holiday photo to military communication. This information has always been the target of malicious entities wanting to steal, modify or delete it. To achieve these goals, hackers and other malicious agents have discovered, exploited, and improved a wide range of cyberattacks.

This new era of cyber security has required a paradigm shift from simple defense mechanisms to sophisticated defense systems. While basic network protections, such as firewalls, may have been sufficient in the past, the increasing complexity of cyberattacks has made them insufficient if used alone. To guard against these complex attacks, Intrusion Detection Systems (IDS) are now the cornerstone of cyber security.

As the sophistication of attacks increases, weaknesses are also discovered and exploited at an increasing rate. Attacks that exploit weaknesses almost as soon as they are discovered are called zero-day attacks. With zero-day hacking attacks breaking records year after year [1], it becomes necessary for IDSs to be able to detect never-before-seen attacks. Machine learning-based IDSs are one of the solutions to this problem. Indeed,

numerous research papers have shown that the use of machine learning provides new approaches and detection mechanisms against zero-day attacks in all areas of cybersecurity [2, 3, 4, 5].

As always in the ongoing rivalry between cyber-attacks and defenses, as machine learning has become an important part of defense mechanisms, it has also become the target of attacks. The field of adversarial learning studies various methods of attack against machine learning and different ways of protecting models against these attacks. These new attacks take advantage of the fact that machine learning models have a discontinuous input-output mapping. This means that humanly imperceptible changes to the input of a model can cause a dramatic change in the classification result. These changes or perturbations in the input are called adversarial examples [6].

This change in cyber-target has led to an equivalent change in cyber defenses, as the focus has shifted to protecting IDSs themselves from attacks created to undermine their effectiveness. While adversarial learning features some defenses, it is clear that more research is needed to refine these techniques. Adversarial detection has shown promising results in the field of computer vision, but very limited work has been done regarding this method in the field of intrusion detection systems.

The main objective of this paper is to design and study the use of multiple strategically placed transfer learning-based detectors of adversarial attacks. The use of multiple detectors could lead to significantly better detection rates against adversarial attacks, as the information is distributed among them and the logical links between these pieces of information can be discovered independently. In our experimental work, the results show that the use of multiple adversarial detectors is more ad-

\*Corresponding author

Email address: islam.debicha@ulb.be (Islam Debicha)

vantageous in parallel IDS than in serial IDS. This is mainly due to the fact that the detectors are more diversified in the parallel architecture. Our four main contributions are the following: (1) Implementation of two IDS models based on deep learning (one in serial form and the other in parallel form) and evaluation of the effect of four adversarial attacks on their performance. (2) Proposal of a new adversarial detection scheme based on transfer learning to allow a better information flow between the IDS and the adversarial detectors. (3) Assessment of the performance of the proposed approach when exposed to evasion attacks that were not seen in the training phase to simulate the zero-day attack scenario. (4) Performance evaluation of the use of multiple adversarial detectors versus a single detector in both serial and parallel IDS designs.

The remainder of the paper is organized as follows. Preliminaries and related works are presented in Section 2. We present our evaluation methodology in Section 3. We present our performance evaluation in Section 4. We conclude in Section 5.

## 2. Preliminaries & Related Works

This section first presents a brief description of the concepts of adversarial learning, as well as the metrics used to evaluate the performance of adversarial attacks. We then present the attacks used in this paper and their methodology. Similarly, we describe the defense mechanisms that can be implemented to impede adversarial attacks. This section also presents the transfer learning techniques used in our detection mechanism. Furthermore, when considering multiple detectors, it is necessary to find a way to combine their respective evaluations into a final decision. Thus, this section presents the three fusion rules used in this work.

### 2.1. Adversarial Learning

Adversarial learning is the name given to the problem of devising attacks against machine learning as well as the defenses against those attacks [6]. There exists a large number of attacks against machine learning models that are often classified with regard to their respective goals [7]. Depending on the phase in which the attack is conducted, adversarial attacks can be divided into either poisoning or evasion attacks.

One of the main differences between poisoning and evasion attacks is the timing of the attack. In the case of the poisoning attack, the adversary targets the IDS during the training phase, whereas for the evasion attack, the adversary launches its attack on the IDS during the testing phase (i.e., only after the IDS has been trained and deployed). Clearly, the poisoning attack is more difficult to execute in a real-world setting, as it requires the attacker to re-train the IDS with its poisoned data. Evasion attacks, on the other hand, allow the attack to be performed without modifying the IDS.

Since evasion attacks are more practical and therefore more widely used against IDSs, our work will focus on evasion attacks and defenses against them. In particular, this work will only consider evasion attacks against Deep Neural Network-based IDS (DNN-IDS). Although DNNs seem to be the most

promising research area in terms of IDS performance [8], these models also seem to be the most vulnerable to adversarial attacks [6].

#### 2.1.1. Assumption of adversary knowledge

When attacking a system in a controlled environment, the first thing to decide is the adversary's knowledge. Indeed, we distinguish between an attacker with the full knowledge of the to-be-attacked model as well as any defenses in place (white-box) and an attacker with no prior knowledge of the attacked system (black-box).

In the case of IDSs, knowledge of the initial classifier is plausible because most IDSs use state-of-the-art models and learning methods. Knowledge of the defenses, on the other hand, is less plausible because there is no clear consensus on the best defense mechanism to implement in this case. In this work, we assume a kind of gray-box situation where the attacker has full knowledge of the IDS but no prior knowledge of the defenses that are employed to defeat it.

#### 2.1.2. Metrics

When performing adversarial attacks, several metrics can be considered to assess the results. Indeed, one of the commonly used metrics is the success rate of the attack: the difference between the performance of the model before and after the attack. The other commonly considered metric is the strength of the attack. Some attacks allow a strength metric to be defined, which increases the success rate but also potentially the detectability of the attack. The strength often refers to the maximum perturbation that can be applied to the initial sample. In this paper, we evaluate the model performance using the standard machine learning metrics: Precision, Recall, and F1-score to evaluate the performance of the IDSs and Detection Rate (Recall) to assess the performance of the adversarial detectors. These metrics are calculated as follows:

$$\text{Precision} = \frac{tp}{tp + fp} \quad (1)$$

$$\text{Recall} = \frac{tp}{tp + fn} \quad (2)$$

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

## 2.2. Adversarial attacks

### 2.2.1. Fast Gradient Sign Method (FGSM) [9]

In an FGSM attack, the objective is to maximize the classification error by maximizing the loss function. Since the attacker cannot modify the internal weights and biases of the attacked model, he instead modifies the input. These modifications are applied with the goal of maximizing the loss function of the victim model. In order to maximize this loss function, the alteration made to the original input vector is calculated with respect to the gradient (i.e., the direction of the steepest ascent) of the original model. To be able to calculate this gradient, FGSM requires internal knowledge of the attacked model, namely the weights and biases. A specially crafted perturbation is applied to each feature following the sign of the gradient.

### 2.2.2. Projected Gradient Descent (PGD) [10]

PGD is an iterative variant of the FGSM attack. The method uses iterative applications of the FGSM method, with a smaller step size. This iterative idea ensures that the resulting adversarial samples are minimally modified while being theoretically misclassified by the attacked classifier. The use of the iterative method is computationally expensive compared to FGSM, so a trade-off must be made between efficiency and computational power.

### 2.2.3. DeepFool (DF) [11]

In the paper presenting the DF attack, the authors introduce the notion of robustness for binary classifiers at a given point. This robustness would be quantified by the minimal perturbation introduced to this point in order to change its classification. This perturbation is equal to the distance between the original point and the closest decision boundary of the classifier. Once the robustness is estimated, the final perturbation vector  $r$  is multiplied by a constant of the form  $1+\epsilon$  in order to cross the decision boundary. This  $\epsilon$  can be considered as a meta-parameter of attack strength.

### 2.2.4. Carlini & Wagner (CW) [12]

While having the same goal as the FGSM attack, the CW attack uses a different approach. Here, the optimization problem is formulated in a new way. That new way of formulating the problem makes it solvable by using an optimization algorithm.

This implies finding a perturbation  $\delta$  that minimizes the distance metric  $D$  while changing the classification of  $x + \delta$  from the original classification of  $x$  to the desired class  $t$ . This  $\delta$  alteration must also ensure that the adversarial example still is in the set  $S$  of accepted samples. This clause limits the perturbation of the features so that it does not render the initial network attack invalid.

## 2.3. Defenses

According to Miller et al.[6], defenses against evasion attacks are divided into two categories: robust classification and anomaly detection.

### 2.3.1. Robust classification

The goal of the robust classification strategy is to correctly classify any adversarial examples, without compromising test performance in the absence of adversarial attacks. To achieve this, the training (and/or testing) process is typically modified (e.g., by preprocessing the data, using a robust training objective, or training with adversarial examples).

One of the earliest defense mechanisms proposed in the literature to counter evasive attacks is adversarial training [13, 9, 14]. Adversarial training is a method of data augmentation, which involves adding data to the training set used to build the classifier. This added data consists of adversarial examples created against the model. These adversarial samples are assigned the correct label so as to create, in theory, a more robust classifier that learns to correctly classify samples even if they have been modified by an adversarial attack.

Although this method gives some promising results in certain research areas, it is vulnerable to the blind-spot problem [6]. Indeed, if adversarial examples of a certain attack family were not added to the training set, the classifier might not be robust against them. It also means that this defense mechanism may not be as effective against zero-day attacks, which are one of the main concerns of the IDS field. Furthermore, as the IDS is retrained with adversarial examples, this could lead to a degradation of its performance in its initial classification task, as shown in [15].

Another adversarial defense proposed in the literature is defensive distillation[16, 17, 18] which was initially employed to reduce the dimensionality of DNNs, a distillation-based defense is proposed by Papernot et al.[16] where the objective is to smooth the decision surface of the model. In [17], Apruzzese et al. introduced a defensive scheme to limit the impact of adversarial disturbances by employing the defensive distillation technique on random forest-based IDS. Experimental evaluations demonstrated the effectiveness of the proposed method with respect to the literature. The authors also found that their defensive scheme does not affect the performance of the IDS in non-adversarial settings. Carlini and Wagner [19] showed that defensive distillation is not secure, i.e., it is no more resistant to targeted misclassification attacks than unprotected neural networks. It is therefore unclear whether the adapted version of defensive distillation for the random forest is more secure against strong attacks.

An alternative approach to countering adversarial attacks is to negate the impact of changes applied to a given feature by ruling it out. Thus, as a defense against a broad selection of adversarial attacks, one could disregard all features that are susceptible to being manipulated by the attacker [20, 21] without changing the logic of the network traffic. However, Apruzzese et al [22] showed that doing so has a detrimental impact on the performance of detectors in non-adversarial settings, e.g., by triggering more false alarms, which is highly undesirable for modern cyber security platforms. The reason for this reduction in quality is that the removed features have a significant impact on the underlying mechanisms of the baseline detectors; as such, excluding them results in a significant drop in performance, with most detectors averaging well below acceptable scores in real-world settings.

### 2.3.2. Anomaly detection

A considerable amount of research has been done in recent years on an alternative defense strategy against evasion attacks which is anomaly detection. One of the reasons given is the fact that robust classification of attacked samples is challenging, whereas their detection is somewhat easier [23]. One supporting argument is that if a good robust classifier is conceived, then a good detector is obtained for free - detection can be performed when the decision of the robust classifier disagrees with the decision of the non-robust classifier [6]. Our work falls into the category of anomaly detection.

This method proposes that instead of trying to classify the adversarial instance correctly despite the attack, the adversarial instance is detected and dealt with accordingly. Most often, this

action consists in rejecting the detected sample and not letting it pass through the system[24, 25, 26]

Many detection methods in the area of adversarial learning have been proposed. In [25], the authors used the neural activation pattern to detect adversarial perturbations in the image classification domain. A method similar to theirs was also used in [27] to detect adversarial traffic in the IDS domain. The method consists of taking a set of test samples (consisting of adversarial examples and clean samples) and getting the to-be-defended model to classify them. During this classification process, a neural activation vector is extracted for each sample. With these vectors, along with the initial sample label (adversarial or clean), a new labeled dataset is created. Using this new dataset, a detection model is trained to learn which activation patterns are most likely to be observed when classifying an adversarial sample.

It is worth mentioning that other works have studied the use of ensemble techniques to improve the detection rate against adversarial attacks, such as [17], [14] and [28], although in these papers the use of ensemble techniques has a different meaning than the one we consider in this paper. For them, the ensemble technique is to train each classifier on a specific attack or application to ensure proper tuning for that task, but this might lead to less generalization in the training process, making it difficult to handle unknown attacks. The overall scheme for them is to assemble multiple classifiers, each specialized in a particular attack or application, whereas for ours, each instance is inspected by all components of the ensemble scheme leading to a deeper inspection of each particular instance. It should also be noted that approaches [17], [14] and [28] seek to make the classifier robust by adding and/or filtering the training data used to train the model. This is not always possible, especially in the case of an already trained and deployed IDS. Our approach, on the other hand, does not affect the IDS but rather inspects it using other subnets, allowing this defense to be applied to already trained and deployed IDSs.

A similar design to that suggested in this paper can be found in [29] where the authors present an approach to detect adversarial perturbations in images by attaching a small sub-network to a deep neural network. This sub-network (detector) is trained specifically to detect adversarial perturbations. According to the authors, the detector performs surprisingly well on CIFAR10 [30] and ImageNET [31] datasets (both are image classification datasets). Their approach is to train a single detector using Feature Extraction as a transfer learning technique. The authors did not provide a way to determine the optimal positioning of the detector, rather they tried different positions where the single detector could be attached. In contrast, our work consists of designing an adversarial detector based on duplication + Fine-Tuning as a transfer learning technique. We show that this gives better results than Feature Extraction (As we will see in Section 3.4). To further improve the detection rate, we propose and investigate an alternative to a single detector by using multiple strategically placed detectors combined with a sophisticated fusion rule.

#### 2.4. Transfer Learning

Many real-world deep learning configurations involve the need to learn new tasks without compromising the performance of existing tasks. For instance, an object recognition robot may be shipped with a default range of abilities, but additional object models that are specific to the given environment need to be considered. In order to achieve this, new tasks should ideally be learned by sharing the parameters of old tasks without degrading the performance of the latter [32].

Given  $\theta_s$  a set of shared parameters for a DNN and  $\theta_o$  task-specific parameters for previously learned tasks, three typical approaches exist for learning new task-specific parameters,  $\theta_n$ , while leveraging previously learned  $\theta_s$  as shown in Fig. 1:

- Feature Extraction (FE): when learning  $\theta_n$ , we keep  $\theta_s$  and  $\theta_o$  unaltered, and the outputs of the shared layers are used as features for the new task learning.
- Fine-Tuning (FT): in this case,  $\theta_s$  and  $\theta_n$  are tuned for the new task, whereas  $\theta_o$  is unchanged.
- Duplication + Fine-Tuning (DFT): it is possible to mirror the original deep neural network and fine-tune it for each new task to build a dedicated neural network.

#### 2.5. Fusion Rules

In order to combine the individual decisions of each detector involved in our proposed design, we will use three fusion rules: majority voting, Simple Bayes averaging, and Dempster-Shafer combination [33].

The majority vote rule is a simple fusion rule and will serve primarily as a reference for the other two rules in our work. With this rule, the final decision will simply be the individual decision made by the majority of the detectors. The main limitation of this fusion rule is that it processes all models equally, meaning that all models contribute equally to the prediction. This poses a problem if certain models perform well in some situations and poorly in others.

When performing a Bayesian average, the activation score for each class will be summed over all the detectors and then divided by the number of detectors. The highest resulting average will be chosen as the final classification by the system. Using this method helps to account for the confidence of each detector in its classification.

The final fusion rule used in this work is the Dempster-Shafer rule of combination [34]. This rule combines evidence elements from different sources (e.g., detectors) to achieve a degree of belief for each hypothesis (classification decision in our case).

let  $\Omega = \{\omega_1, \dots, \omega_K\}$ , and  $\mathcal{P}(\Omega) = \{A_1, \dots, A_Q\}$  is its power set, where  $Q = 2^K$ . A mass function  $M: \mathcal{P}(\Omega) \rightarrow [0, 1]$  is a basic belief assignment (*bba*) if  $M(\emptyset) = 0$  and  $\sum_{A \in \mathcal{P}(\Omega)} M(A) = 1$ .

In case where two *bbas*  $M_1$  and  $M_2$  denote two elements of evidence (e.g., information from two detectors), we can combine them together using the Dempster-Shafer fusion rule, which results in  $M = M_1 \oplus M_2$  that is defined by Eq. 4.

$$M(A) = (M_1 \oplus M_2)(A) \propto \sum_{B_1 \cap B_2 = A} M_1(B_1)M_2(B_2) \quad (4)$$

### 3. Evaluation methodology

Anomaly detection is one of the key defenses proposed against adversarial learning in intrusion detection systems. Indeed, to perform any variant of evasion attacks on a trained model, the original features used by the model are modified to maximize the classification error. These modifications can be detected in a variety of ways, including using another neural network trained specifically to recognize the modified input packets. These new neural networks are called adversarial detectors (AdD).

This work aims to investigate the effectiveness of multiple adversarial detectors working together to detect adversarial perturbations during the inference phase. Each of these detectors is placed under a sub-network to receive different levels of information from the DNN-IDS. By combining their respective decisions, we could obtain a final classification affirming the legitimacy of each sample. The hypothesis here is that due to the added level of granularity, multiple detectors could help detect a wider range of perturbations and thus a wider range of evasion attacks. In addition, the use of transfer learning techniques would allow adversarial detectors to learn some important patterns from the IDS (since the two tasks are similar) while augmenting their knowledge based on the adversarial patterns.

To do this, we build two DNN-IDS in both serial (DNN-IDS-serial) and parallel (DNN-IDS-parallel) architectures, perform multiple attacks against these two IDSs, and then design our adversarial detectors by following several transfer learning techniques to finally combine their decisions as shown in Figures 3 and 5. This defense technique is then compared to adversarial learning.

#### 3.1. Network traffic datasets

The experimental evaluation considered in our paper is performed on two public network traffic datasets, NSL-KDD[35] and CIC-IDS2017[36]. NSL-KDD is chosen in this work because of its frequent use in the literature as a benchmark to compare different intrusion detection methods. CIC-IDS2017,

on the other hand, is chosen because of its recency, which ensures that it can be considered a good representation of large modern network environments.

#### 3.2. Adversarial examples generation

To craft adversarial samples, we use four different evasion attacks against each model. For this, the set of adversarial attacks that we use in the experiments is Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), Carlini & Wagner (CW), and DeepFool (DF).

To train and evaluate the adversarial detectors, we create multiple training and testing datasets based on the network traffic dataset as shown in Fig. 2. We create an attack-specific dataset for each of the adversarial attacks, comprising the original dataset re-labeled as clean and of the same dataset altered by the chosen attack and labeled as adversarial. This results in balanced datasets between non-altered and altered samples. The final dataset that is created is a balanced version of the attacks. Indeed, while the balanced dataset still contained the original dataset, the altered part is equally shared between the four adversarial attacks.

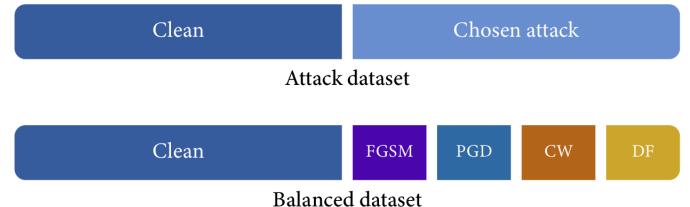


Figure 2. Balanced and attack-specific datasets composition

It bears mentioning that our attacks are applied directly to feature vectors (feature-space attacks) as opposed to raw traffic (problem-space attacks) [37]. Therefore, adversarial perturbations may not resemble truly realistic adversarial traffic [38, 39]. However, in an attempt to preserve the functionality of our adversarial samples, certain semantic and syntactic

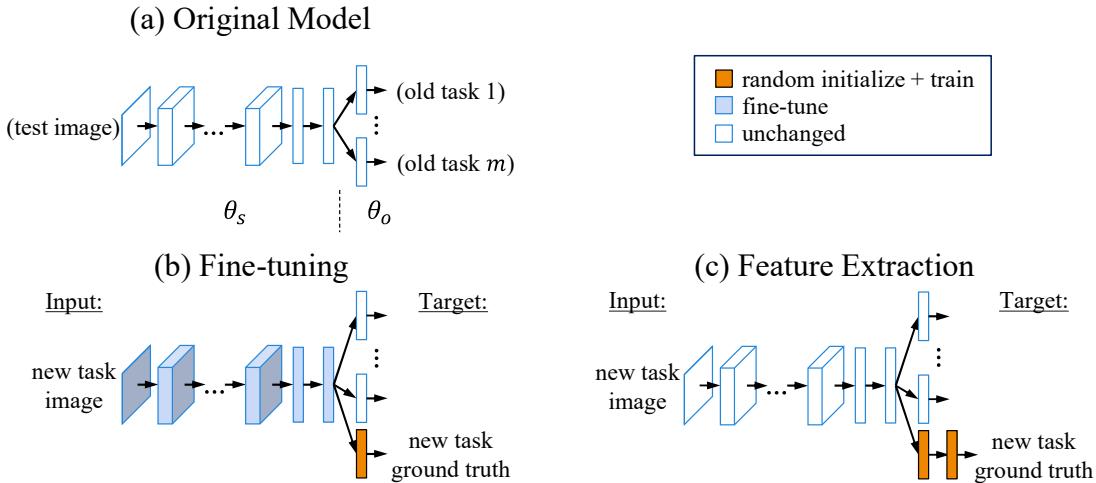


Figure 1. An illustration of two transfer learning techniques:Fine-Tuning and Feature Extraction, from [32]

constraints are taken into consideration [40]. A set of features are kept unmodifiable, such as protocol type, service, or flags. We also limit the maximum perturbation rate to 10% to ensure that the generated adversarial samples are close to the original samples.

### 3.3. Serial DNN-IDS design

The first step in this work is to design the DNN-IDS-serial. Its objective will be to classify benign and malicious traffic. Several design choices have to be tested before settling on an optimal design of the DNN: the number of hidden layers, the number of neurons per hidden layer, and the activation function.

The first design choice that was tested was the number of hidden layers to use. Indeed, while it was known that several hidden layers would be used in this DNN, the exact number of hidden layers is a tested parameter. We decided to test between two and four hidden layers. To test each configuration, we trained ten models of each configuration for fifty epochs each. During the training phase, we evaluated each model on our test dataset every ten epochs. This allowed us to keep only the best-performing model and avoid overfitting. We then averaged the performance obtained by each configuration. We found that using three ReLu-activated hidden layers was sufficient to achieve state-of-the-art performance on this dataset and that adding additional layers did not improve performance further.

Based on these results, we decided to use three ReLu-activated hidden layers. We also varied the number of neurons in each hidden layer between forty and two hundred and fifty-six. This variation in the number of neurons resulted in almost no change in overall performance. It was therefore decided that each hidden layer would have two hundred and fifty-six neurons.

### 3.4. Adversarial detectors design for DNN-IDS-serial

The choices made in this stage regarding our adversarial detectors revolved around the design of the detectors and their architecture. Regardless of these choices, we decided to give each detector the information of all layers placed before it in the DNN-IDS-serial. This means that  $AdD_{n \in [1, N]}$  will be composed by the  $[1, N]$  layers of DNN-IDS-serial in addition to the chosen detector design.

The design of the detectors refers to the number of layers as well as the number of neurons on each layer. We decided to test either one or two ReLu-activated layers, and either forty or two hundred and fifty-six neurons, as we did for our DNN-IDS-serial design. The architecture itself refers to the transfer learning technique used to train our adversarial detectors. The tested architectures are the following:

- Features Extraction: For  $AdD_{n \in [1, N]}$ , layers  $[1, N]$  of the DNN-IDS-serial are shared with the adversarial detector. Those layers are made untrainable by freezing the weights and biases. To that set of layers are added our detector's layers. Those new layers are then trained to detect the attacks. Theoretically, this method allows faster training of the detectors. Indeed, since the shared layers are already trained on a similar task, the provided weights

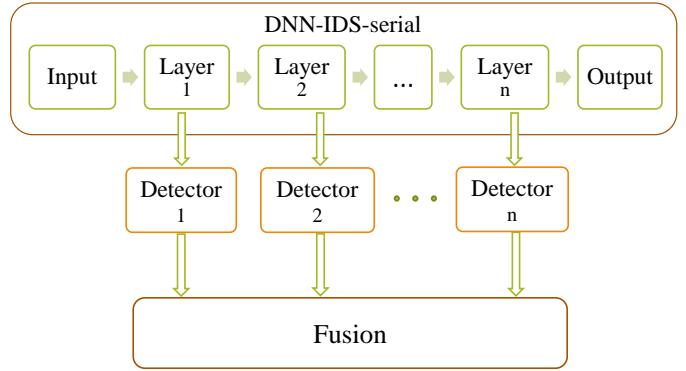


Figure 3. An illustration of the adversarial detectors in a serial DNN-IDS design

and biases should help to find optimal values for the detector's parameters faster than starting with random values.

- Fine-Tuning: For  $AdD_{n \in [1, N]}$ , layers  $[1, N]$  of the DNN-IDS-serial are shared with the adversarial detector. To that set of layers are added our detector's layers. The ensemble is then trained to detect the adversarial attacks. Since this method changes the DNN-IDS-serial's weights and biases, there is a non-zero risk of changing the DNN-IDS-serial's performance. Since both tasks are quite similar, this change might be beneficial for the original detection performance. But it is also possible that we observe a drop in IDS performance after training our adversarial detectors.
- Duplication + Fine-Tuning: For  $AdD_{n \in [1, N]}$ , layers  $[1, N]$  of the DNN-IDS-serial are duplicated and their copy is passed to the adversarial detector. To those layers are added our detector's layers. We then train the whole set of layers to detect adversarial attacks. This method benefits from the advantages of a Fine-Tuning architecture without risking a performance change for the DNN-IDS-serial. Indeed, since the layers are duplicated and not shared, the weights and biases of the original DNN-IDS-serial remain unchanged. This architecture would be a suitable solution if the performance change observed while using the Fine-Tuning architecture is detrimental to our DNN-IDS-serial.

To test the design and architecture of our adversarial detectors, we trained them for thirty epochs each and evaluated them in the same way as we evaluated the initial DNN-IDS-serial. The whole experiment was repeated 3 times to average out the results. FGSM attack was used to craft adversarial samples from the NSL-KDD dataset.

As shown in Table 1, we can see that the architectures allowing the re-training of the initial layers performed better at detecting the evasion attacks. This result is attributed to the fact that both tasks (detecting intrusion and evasion attacks) are quite similar. This means that the architecture allowing to fine-tune the DNN-IDS-serial benefits from its training and can cap-

Architecture	Layers	Neurons	AdD1	AdD2	AdD3	Mean	IDS F1-score change
DFT (this paper)	2	256	97,12%	97,50%	97,16%	97,26%	No change
		40	97,00%	96,79%	97,05%	96,94%	
	1	256	97,75%	97,12%	95,56%	96,81%	
		40	97,01%	97,00%	97,09%	97,03%	
FT	2	256	96,80%	95,93%	96,14%	96,29%	-1,04%
		40	96,31%	96,69%	96,72%	96,57%	-0,57%
	1	256	95,62%	96,52%	96,26%	96,13%	-0,69%
		40	97,20%	97,36%	96,13%	96,90%	-4,80%
FE (used in [33])	2	256	83,56%	82,02%	85,08%	83,55%	No change
		40	84,98%	83,49%	85,38%	84,62%	
	1	256	84,24%	84,84%	85,73%	84,94%	
		40	86,42%	85,37%	86,48%	86,09%	

Table 1. Comparaison of the detection rate of the adversarial detectors (AdD) in different design choices. For DFT and FE architectures, the IDS is not re-trained, hence no change in its performance.

italize on it. Unfortunately, an accuracy drop was observed in the DNN-IDS-serial when using the Fine-Tuning architecture. We notice also that the design choices of the adversarial detectors are of little influence on the detection rate. This is not surprising as it was already observed with the design of DNN-IDS-serial. We decided to use the Duplication + Fine-Tuning architecture with a design of 2 ReLu-activated layers of two-hundred and fifty-six neurons each for our adversarial detectors.

### 3.5. Parallel DNN-IDS design

In their original paper, the authors present Kitsune as a new IDS generation using an ensemble idea to split the learning process between multiple sub-models each training faster and using less computational and memory resources during the training process. To split the training process, the feature vector extracted from the training samples are first clustered together using their correlation to each other. Once clusters are decided, each cluster of features will be passed to one of the sub-models comprising the ensemble layer of Kitsune. The ensemble layer is comprised of a certain number of auto-encoders each taking one cluster of features and outputting the root-mean-square error (RMSE) between the original feature cluster and the one returned by the auto-encoder. Those RMSE are then used as input to another auto-encoder, the output layer, to compute the final RMSE which is then compared to a fixed threshold to perform a classification between benign and malicious samples. The Kitsune feature mapping, as well as ensemble & output layers are shown in Fig. 4. The reason we chose this architecture is to add a deep inspection layer to better detect adversarial perturbations by increasing the granularity.

#### 3.5.1. Feature clustering

As in the original Kitsune implementation, the network traffic features were split into multiple clusters. The number of clusters  $K$  obtained during this step is bound by an arbitrarily fixed parameter ( $K$  is fixed to nine (09) for NSL-KDD and five (05) for CIC-IDS2017). Since the features present in our datasets are not the same as the ones used in the original Kitsune paper, the relationship between them is not as direct, and using the original Kitsune clustering method yielded unusable feature clusters. We decided to propose two clustering methods.

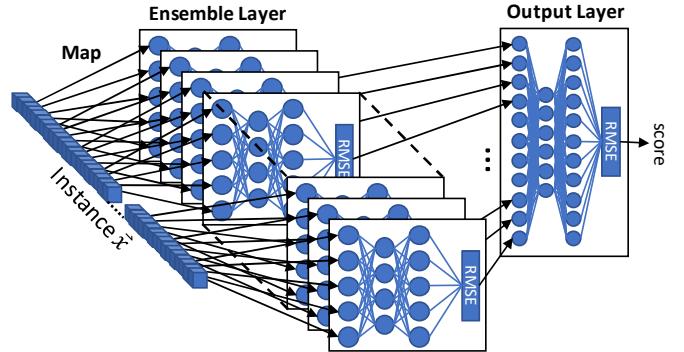


Figure 4. An illustration of Kitsune’s anomaly detection algorithm, from [41]

The first method, referred to as the distribution method, first performs the same clustering as the original Kitsune version. With highly uncorrelated features, the original Kitsune feature map may contain clusters containing only one feature. As passing only one feature to any neural network from the ensemble layer is not intended, the features belonging to those clusters are distributed between the other smaller clusters. This method will always yield clusters comprising a minimum of two features as well as a number of clusters inferior or equal to  $K$ .

The second method, referred to as the cut method, will first compute the number of features in each cluster by dividing the total amount of features by the number of clusters given by  $K$ . The features will then be ordered with regard to their correlation before being distributed between the clusters. This method will always yield  $K$  clusters of approximately the same size (plus or minus one feature).

By training the DNN-IDS-parallel with the two clustering methods, we observed that the distribution method performed better than the cut method. Therefore, the distribution method was chosen as the reference method throughout the rest of the work.

In the original Kitsune implementation, the ensemble and output layers were composed of auto-encoders. we decided to replace those with deep neural networks similar to the DNN-IDS-serial used in the first part of this work. Each model in our ensemble and output layers is composed of three hidden ReLu layers of two hundred and fifty-six neurons and a binuronal SoftMax layer. Each model of the ensemble layer was then trained for thirty epochs with a performance evaluation every ten epochs, keeping the best-performing model as the final trained model. From the predictions of those trained models for each instance, a new dataset was created. This new dataset was then used to train the model composing the output layer in the same manner as the models of the ensemble layer. Using this configuration, the final model achieved state-of-the-art performance on our tested datasets.

### 3.6. Adversarial detectors design for DNN-IDS-parallel

For the adversarial detector design, we decided to follow the same choices as for the DNN-IDS-serial. In Section 3.4, we compared three architectures: Fine-tuning, Features extraction,

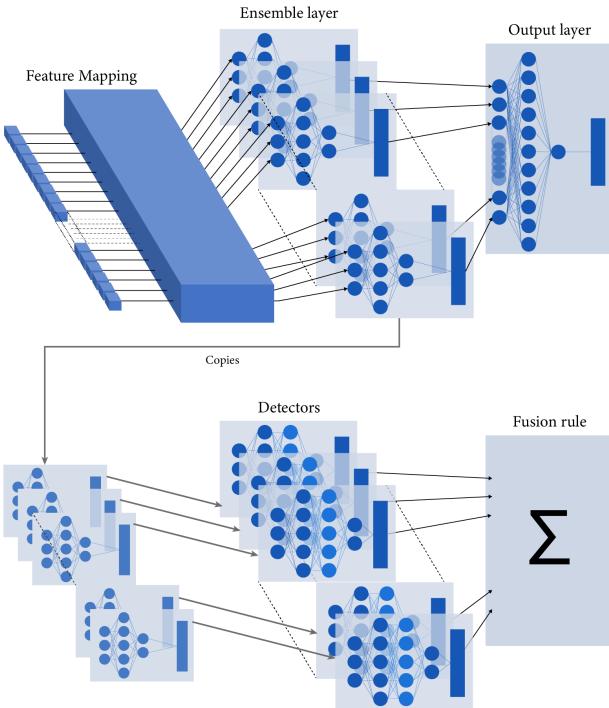


Figure 5. An illustration of the adversarial detectors in a parallel DNN-IDS design

and Duplication + Fine-tuning. The obtained results showed that being able to re-train any part of the DNN-IDS re-used or shared by the adversarial detectors led to a better detection rate. Since we also observed that using a Fine-Tuning architecture without duplication also led to a drop in accuracy for the DNN-IDS, we decided to use the Duplication + Fine-Tuning architecture. The multi adversarial detectors architecture in a parallel DNN-IDS design is illustrated in Fig. 5. Each adversarial detector is composed of:

- A copy of the input layer of the corresponding ensemble layer model.
- A copy of the hidden ReLu layers of the corresponding ensemble layer model, as well as their weights and biases.
- Two additional two hundred and fifty-six neurons hidden ReLu layers, randomly initialized.
- A mono-neuronal sigmoid output layer

The obtained array of detectors will then be trained in the same way as the ensemble layers: thirty epochs with an evaluation every ten epochs and keeping the best-evaluated detector.

#### 4. Performance evaluation

This section presents the performance evaluation of the experimental settings. We first investigate the effect of the four

adversarial attacks on the performance of the two DNN-IDSs. The adversarial detectors are then assessed by a cross-detection test. The final step is to evaluate the effectiveness of the three fusion rules. The results are then compared to another adversarial defense technique (adversarial training).

##### 4.1. Serial DNN-IDS

###### 4.1.1. Adversarial attacks effect on DNN-IDS-serial

In order to evaluate the impact of the four adversarial attacks, namely FGSM, PGD, CW, and DF on the performance of DNN-IDS-serial, we test the model with the four attack-specific datasets crafted as mentioned in Section 3.2.

The results presented in Fig. 6 confirm the effect of adversarial attacks on DNN-IDS-serial. Indeed, we notice a significant drop in the IDS performance once the samples are altered by any of the adversarial attacks. For example, using PGD on the NSL-KDD dataset resulted in a decrease in F1-score from 83% to 38%, while using FGSM on the CIC-IDS2017 dataset resulted in a decrease in F1-score from 98% to 48%. These results are compatible with those found in the literature [40, 42].

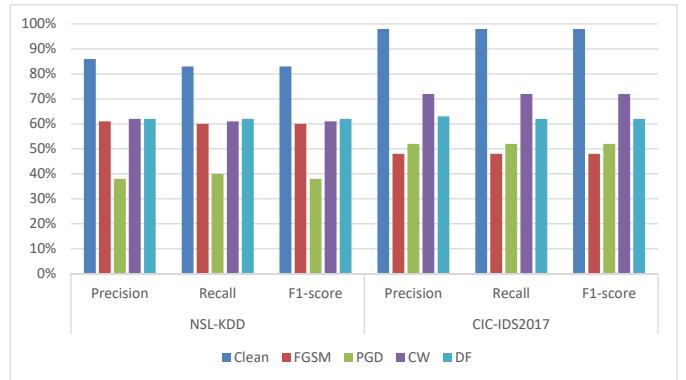


Figure 6. IDS performance difference between clean and adversarial network traffic in a serial DNN-IDS design

###### 4.1.2. Adversarial detectors performance

To evaluate the performance of adversarial detectors, we conduct a cross-detection test. For this test, we train a set of adversarial detectors on one of our adversarial training datasets and then evaluate it against all adversarial testing datasets. This allows us to see how an adversarial detector trained on a specific attack or a specific subset of attacks would perform against an unknown attack. We also compare the results between detectors to analyze any differences in performance among them.

As shown in Table 2, the first notable observation that can be made is that the detection rate differences between each detector in a set are minimal. If these minimal differences reflect the fact that every detector gives the same answer as the others for a vast majority of the samples, the fusion rules will yield results similar to those of any single detector. On the other hand, if those differences are only the results of the overall proportion of correctly identified samples, the fusion rules should yield better results than any single adversarial detector.

NSL-KDD						
AdD1						
Train\Test	FGSM	PGD	CW	DF	Balanced	Mean
FGSM	99.57%	84.06%	50.95%	70.61%	76.35%	76.31%
PGD	71.84%	83.79%	50.89%	70.02%	69.18%	69.14%
CW	53.62%	64.09%	58.82%	63.07%	60.12%	59.94%
DF	71.90%	80.57%	51.13%	70.42%	68.47%	68.50%
Balanced	94.38%	82.12%	58.13%	69.48%	76.13%	76.05%
					Grand Mean	69.99%
CIC-IDS2017						
AdD1						
Train\Test	FGSM	PGD	CW	DF	Balanced	Mean
FGSM	99.80%	78.91%	54.61%	71.54%	76.20%	76.21%
PGD	93.78%	86.46%	55.21%	74.76%	77.72%	77.58%
CW	62.70%	66.27%	72.84%	68.25%	67.18%	67.45%
DF	91.23%	81.99%	54.69%	78.63%	76.92%	76.69%
Balanced	98.96%	85.10%	71.87%	77.76%	83.56%	83.45%
					Grand Mean	76.28%
AdD2						
Train\Test	FGSM	PGD	CW	DF	Balanced	Mean
FGSM	99.50%	84.37%	50.88%	70.60%	76.37%	76.34%
PGD	66.74%	81.87%	51.02%	67.87%	66.94%	66.89%
CW	53.21%	63.66%	60.26%	63.72%	60.40%	60.25%
DF	70.38%	80.04%	50.77%	69.87%	67.76%	67.77%
Balanced	94.45%	83.00%	57.98%	70.05%	76.32%	76.36%
					Grand Mean	69.52%
AdD3						
Train\Test	FGSM	PGD	CW	DF	Balanced	Mean
FGSM	99.43%	85.07%	51.04%	70.74%	76.58%	76.57%
PGD	64.24%	83.37%	50.98%	68.19%	66.78%	66.71%
CW	59.64%	73.45%	60.02%	64.23%	64.34%	64.34%
DF	67.12%	77.38%	52.19%	69.85%	66.61%	66.63%
Balanced	92.22%	83.70%	56.35%	69.60%	75.54%	75.48%
					Grand Mean	69.94%

Table 2. Detection rate of the individual adversarial detectors in a serial DNN-IDS design on NSL-KDD and CIC-IDS2017

The detection rate results themselves reflect the fact that the CW attack is harder to detect than the other attacks. The detection rate difference could be amplified by the fact that our attack set comprises two FGSM-based attacks, FGSM itself and PGD, and that both could be detected similarly, resulting in a virtually unbalanced attack set. This hypothesis is also supported by the fact that the DF attack yields worst detection rates than both FGSM-based attacks.

#### 4.1.3. Fusion rules

The final step is to implement the three fusion rules we decided to use: Majority Voting, Bayes Simple Average, and Dempster-Shafer Combination rules. To test these fusion rules, we use the same method as when testing the cross-detection of individual adversarial detectors. The only difference is that the decisions of each adversarial detector are now combined using the chosen rule to obtain a final decision.

As shown in Table 3, the three fusion rules produced similar results to those of the individual detectors. This confirms our previous hypothesis that each detector is probably giving the same response as the other detectors for most samples. Even though each detector has access to the information of an additional layer compared to its predecessor, it seems that all detectors are obtaining similar levels of information about the tested instances and therefore giving the same results. This means that the use of multiple adversarial detectors in a serial DNN-IDS design does not represent a significant improvement over the use of a single detector in our experimental setting.

## 4.2. Parallel DNN-IDS

### 4.2.1. Adversarial attacks effect on DNN-IDS-parallel

To perform each adversarial attack on our DNN-IDS-parallel, we decided to craft adversarial input for each model in the ensemble layer. To do this, each attack was performed against each specific subset of features used as input by that model.

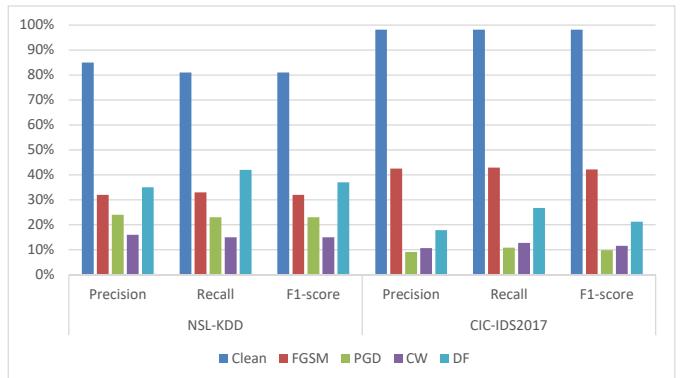


Figure 7. IDS accuracy difference between clean and adversarial network traffic in a parallel DNN-IDS design

As shown in Fig. 7, similarly to the DNN-IDS-serial, we observe a significant drop in the IDS performance after performing adversarial attacks against our DNN-IDS-parallel. For example, using CW on the NSL-KDD dataset resulted in a decrease in F1-score from 81% to 15%, while using PGD on the

NSL-KDD						
Majority Vote Rule						
Train\Test	FGSM	PGD	CW	DF	Balanced	Mean
FGSM	99.80%	84.58%	50.93%	70.76%	76.57%	76.53%
PGD	68.65%	83.13%	50.98%	69.26%	68.11%	68.03%
CW	52.87%	64.71%	60.43%	62.94%	60.38%	60.26%
DF	69.62%	79.71%	50.98%	70.04%	67.53%	67.58%
Balanced	96.41%	84.12%	57.59%	70.60%	77.19%	77.18%
					Grand Mean	69.92%

CIC-IDS2017						
Majority Vote Rule						
Train\Test	FGSM	PGD	CW	DF	Balanced	Mean
FGSM	99.88%	78.90%	54.53%	71.61%	76.22%	76.23%
PGD	89.09%	85.62%	53.89%	69.32%	74.81%	74.54%
CW	79.61%	78.04%	72.12%	67.48%	74.41%	74.33%
DF	92.91%	81.44%	54.20%	77.05%	76.71%	76.46%
Balanced	96.36%	84.54%	68.83%	77.23%	81.95%	81.78%
					Grand Mean	76.67%

Simple Bayes Average Fusion Rule						
Train\Test	FGSM	PGD	CW	DF	Balanced	Mean
FGSM	99.80%	84.60%	50.93%	70.77%	76.58%	76.54%
PGD	70.22%	83.57%	51.06%	69.32%	68.59%	68.55%
CW	53.20%	68.62%	60.16%	64.88%	62.02%	61.78%
DF	74.24%	81.10%	51.24%	70.41%	69.20%	69.24%
Balanced	96.70%	83.20%	58.44%	70.22%	77.20%	77.15%
					Grand Mean	70.65%

Dempster-Shafer Combination Rule						
Train\Test	FGSM	PGD	CW	DF	Balanced	Mean
FGSM	99.83%	84.64%	50.94%	70.78%	76.61%	76.56%
PGD	70.67%	83.63%	51.09%	69.17%	68.69%	68.65%
CW	53.33%	68.64%	60.16%	64.96%	62.07%	61.83%
DF	74.91%	81.22%	51.25%	70.44%	69.43%	69.45%
Balanced	96.58%	83.00%	58.47%	70.10%	77.10%	77.05%
					Grand Mean	70.71%

Table 3. Detection rate of the fusion of the adversarial detectors in a serial DNN-IDS design on NSL-KDD and CIC-IDS2017

CIC-IDS2017 dataset resulted in a decrease in F1-score from 98% to 10%. These results are compatible with those found in the literature [43].

#### 4.2.2. Adversarial detectors performance

When training the detectors on the various datasets, we observed greatly varying detection rates between the detectors as shown in Table 7. Indeed, some detectors achieve an accuracy as low as 50% while others can often reach as high as 100% detection rate. We notice also that many individual detectors are having low detection rate against CW and DF attacks compared to FGSM and PGD. The differences in detection rate observed between the detectors are believed to come from the fact that most attacks alter specific features and disregard others. This would lead to some detectors rarely seeing any altered features, thus classifying almost every sample as legitimate and achieving only a 50% detection rate. On the other hand, some detectors would receive the often-altered features and thus be able to detect an adversarial attack. As the best-performing detector will vary depending on the training dataset used, we expect that the use of multiple detectors will be an improvement over the use of a single detector.

#### 4.2.3. Fusion rules

Considering the results obtained on individual detectors, it is now obvious that the way in which we will combine those individual classifications will be of great importance for the overall detection rate of our system. Indeed, from the three proposed

fusion rules (cf. 2.5), the majority vote rule is expected to be the least effective. Since some attacks only alter a small number of features, it is to be expected that only a few adversarial detectors will be able to detect these attacks while the majority will only see unaltered legitimate features, thus resulting in a false classification. While the Dempster-Shafer combination rule also uses the multiplicity of belief as a decisive factor, we believe that the effect will be less important than for the majority vote rule. The Dempster-Shafer combination rule only eliminates outcomes if they are absent from the possible outcomes decided by one of the detectors. This means that for this effect to eliminate one of the outcomes, it is required that one of the detectors returns a classification with a 100% confidence. As the Dempster-Shafer combination rule is a more sophisticated rule and takes the detector uncertainty into consideration, it is expected that this rule will outperform both other rules.

The same cross-detection rate as before was computed, but this time combining the different classifications with each tested fusion rule. This method allows us to compare the different fusion rules but also to observe the possible generalization of detection through multiple attacks.

From Table 4, we notice that the fusion of the adversarial detectors using the Dempster-Shafer combination rule is outperforming the individual detectors in terms of detection rate. This confirms the relevance of using multiple, strategically placed adversarial detectors.

The contrast in observed results could come from the differences in the way every attack is performed. Indeed, both the





trained only on non-adversarial samples, and a robust adversarial model trained on both non-adversarial and adversarial samples. When subjecting a sample to the two models, one can assume that if the two models classify it differently, the sample is adversarial. In theory, if the submitted sample is non-adversarial, the base model and the robust model should classify it correctly. On the other hand, if the sample is adversarial, the base model should classify it incorrectly, while the robust model should classify it correctly.

We implemented this defense by training a second parallel DNN-IDS model on a dataset consisting of equal amounts of adversarial and non-adversarial samples. We then submitted a likewise test dataset to both models and recorded their predictions. By comparing these two sets of predictions, we obtained a final set of predictions classifying each sample as adversarial or non-adversarial. As shown in Table 5 and Table 6, this method yielded a detection rate of 71.69% and 74.05% for NSL-KDD and CIC-IDS2017 respectively. As we demonstrated earlier, our proposal using the Dempster-Shafer combination rule obtained a strictly higher detection rate on both datasets.

These differences in results can be explained by the contrasting objectives of the two strategies. Our proposed defense is specifically aimed at detecting adversarial samples. On the other hand, the adversarial training detection method is a by-product of training a robust classifier which should remove the need for detection (in theory, at least). In addition, the difference in result could stem from the fact that our proposed defense has a higher level of granularity by inspecting each subset of features separately and then reaching a consensus decision using an appropriate fusion rule.

## 5. Conclusion and Future Work

In this paper, we proposed a new defense approach for evasion attacks against network-based intrusion detection systems. To evaluate it, we implemented two deep learning-based IDS models (one serial and one parallel) and assessed the effect of four known adversarial attacks on their performance, namely: Fast Gradient Sign Method, Projected Gradient Descent, Carlini & Wagner, and DeepFool. A transfer learning technique was employed in the design of the adversarial detection scheme to enable a better information flow between the IDS and the adversarial detectors. Simulation of zero-day attack scenarios allowed for a more reliable evaluation of the performance of the proposed defense when exposed to evasion attacks that were not seen in the training phase. To further improve the detection rate, we proposed and investigated an alternative to a single detector by using multiple strategically placed detectors combined with a sophisticated fusion rule. We show that combining multiple detectors can further improve the detection rate over a single detector in a parallel IDS design. The reported results confirm the relevance of the proposed defense with respect to existing techniques in the literature.

In future work, we would like to improve the fusion rules used in our design. Indeed, while the use of the Dempster-Shafer fusion rule gave some promising results, it was not fully

successful in all cases, which could mean that a more sophisticated fusion rule is needed to model the complex relationship between detectors classification. In addition, we would like to further investigate the relationship between detectability and the strength of adversarial attacks.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## CRediT authorship contribution statement

**Islam Debicha:** Conceptualization, Validation, Methodology, Writing - Original Draft. **Richard Bauwens:** Software, Visualization, Writing - Original Draft. **Thibault Debatty:** Validation, Writing - Review & Editing, Supervision. **Jean-Michel Dricot:** Validation, Writing - Review & Editing, Supervision. **Tayeb Kenaza:** Validation, Writing - Review & Editing, Supervision. **Wim Mees:** Funding acquisition, Supervision.

## References

- [1] P. H. O'Neillarchive, 2021 has broken the record for zero-day hacking attacks, MIT Technology Review (2021). URL <https://www.technologyreview.com/2021/09/23/1036140/2021-record-zero-day-hacks-reasons/>
- [2] G. Apruzzese, M. Colajanni, L. Ferretti, A. Guido, M. Marchetti, On the effectiveness of machine and deep learning for cyber security, in: T. Minárik, R. Jakschis, L. Lindström (Eds.), 10th International Conference on Cyber Conflict, CyCon 2018, Tallinn, Estonia, May 29 - June 1, 2018, IEEE, 2018, pp. 371–390. doi:[10.23919/CYCON.2018.8405026](https://doi.org/10.23919/CYCON.2018.8405026) URL <https://doi.org/10.23919/CYCON.2018.8405026>
- [3] Y. Xin, L. Kong, Z. Liu, Y. Chen, Y. Li, H. Zhu, M. Gao, H. Hou, C. Wang, Machine learning and deep learning methods for cybersecurity, IEEE Access 6 (2018) 35365–35381. doi:[10.1109/ACCESS.2018.2836950](https://doi.org/10.1109/ACCESS.2018.2836950). URL <https://doi.org/10.1109/ACCESS.2018.2836950>
- [4] S. Mahdavifar, A. A. Ghorbani, Application of deep learning to cybersecurity: A survey, Neurocomputing 347 (2019) 149–176. doi:[10.1016/j.neucom.2019.02.056](https://doi.org/10.1016/j.neucom.2019.02.056). URL <https://doi.org/10.1016/j.neucom.2019.02.056>
- [5] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat, S. Venkatraman, Deep learning approach for intelligent intrusion detection system, IEEE Access 7 (2019) 41525–41550. doi:[10.1109/ACCESS.2019.2895334](https://doi.org/10.1109/ACCESS.2019.2895334). URL <https://doi.org/10.1109/ACCESS.2019.2895334>
- [6] D. J. Miller, Z. Xiang, G. Kesidis, Adversarial learning targeting deep neural network classification: A comprehensive review of defenses against attacks, Proc. IEEE 108 (3) (2020) 402–433. doi:[10.1109/JPROC.2020.2970615](https://doi.org/10.1109/JPROC.2020.2970615). URL <https://doi.org/10.1109/JPROC.2020.2970615>
- [7] I. Corona, G. Giacinto, F. Roli, Adversarial attacks against intrusion detection systems: Taxonomy, solutions and open issues, Inf. Sci. 239 (2013) 201–225. doi:[10.1016/j.ins.2013.03.022](https://doi.org/10.1016/j.ins.2013.03.022). URL <https://doi.org/10.1016/j.ins.2013.03.022>
- [8] J. Lansky, S. Ali, M. Mohammadi, M. K. Majeed, S. H. T. Karim, S. Rashidi, M. Hosseinzadeh, A. M. Rahmani, Deep learning-based intrusion detection systems: A systematic review, IEEE Access 9 (2021) 101574–101599. doi:[10.1109/ACCESS.2021.3097247](https://doi.org/10.1109/ACCESS.2021.3097247). URL <https://doi.org/10.1109/ACCESS.2021.3097247>

- [9] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.  
URL <http://arxiv.org/abs/1412.6572>
- [10] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, in: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings, OpenReview.net, 2018.  
URL <https://openreview.net/forum?id=rJzIBfZAb>
- [11] S. Moosavi-Dezfooli, A. Fawzi, P. Frossard, Deepfool: A simple and accurate method to fool deep neural networks, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, IEEE Computer Society, 2016, pp. 2574-2582. doi:10.1109/CVPR.2016.282.  
URL <https://doi.org/10.1109/CVPR.2016.282>
- [12] N. Carlini, D. A. Wagner, Towards evaluating the robustness of neural networks, in: 2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017, IEEE Computer Society, 2017, pp. 39-57. doi:10.1109/SP.2017.49.  
URL <https://doi.org/10.1109/SP.2017.49>
- [13] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, R. Fergus, Intriguing properties of neural networks, in: Y. Bengio, Y. LeCun (Eds.), 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings, 2014.  
URL <http://arxiv.org/abs/1312.6199>
- [14] G. Apruzzese, M. Andreolini, M. Marchetti, A. Venturi, M. Colajanni, Deep reinforcement adversarial learning against botnet evasion attacks, *IEEE Trans. Netw. Serv. Manag.* 17 (4) (2020) 1975–1987. doi:10.1109/TNSM.2020.3031843.  
URL <https://doi.org/10.1109/TNSM.2020.3031843>
- [15] I. Debicha, T. Debatty, J. Dricot, W. Mees, Adversarial training for deep learning-based intrusion detection systems, in: ICONS 2021, The Sixteenth International Conference on Systems, 2021, pp. 45–49.  
URL [https://www.thinkmind.org/index.php?view=article&articleid=icons\\_2021\\_3\\_40\\_40009](https://www.thinkmind.org/index.php?view=article&articleid=icons_2021_3_40_40009)
- [16] N. Papernot, P. D. McDaniel, X. Wu, S. Jha, A. Swami, Distillation as a defense to adversarial perturbations against deep neural networks, in: IEEE Symposium on Security and Privacy, SP 2016, San Jose, CA, USA, May 22-26, 2016, IEEE Computer Society, 2016, pp. 582–597. doi:10.1109/SP.2016.41.  
URL <https://doi.org/10.1109/SP.2016.41>
- [17] G. Apruzzese, M. Andreolini, M. Colajanni, M. Marchetti, Hardening random forest cyber detectors against adversarial attacks, *IEEE Trans. Emerg. Top. Comput. Intell.* 4 (4) (2020) 427–439. doi:10.1109/TETCI.2019.2961157.  
URL <https://doi.org/10.1109/TETCI.2019.2961157>
- [18] K. Grosse, N. Papernot, P. Manoharan, M. Backes, P. D. McDaniel, Adversarial examples for malware detection, in: S. N. Foley, D. Gollmann, E. Snekkernes (Eds.), Computer Security - ESORICS 2017 - 22nd European Symposium on Research in Computer Security, Oslo, Norway, September 11-15, 2017, Proceedings, Part II, Vol. 10493 of Lecture Notes in Computer Science, Springer, 2017, pp. 62–79. doi:10.1007/978-3-319-66399-9\_4.  
URL [https://doi.org/10.1007/978-3-319-66399-9\\_4](https://doi.org/10.1007/978-3-319-66399-9_4)
- [19] N. Carlini, D. A. Wagner, Defensive distillation is not robust to adversarial examples, *CoRR* abs/1607.04311 (2016). arXiv:1607.04311.  
URL <http://arxiv.org/abs/1607.04311>
- [20] F. Zhang, P. P. K. Chan, B. Biggio, D. S. Yeung, F. Roli, Adversarial feature selection against evasion attacks, *IEEE Trans. Cybern.* 46 (3) (2016) 766–777. doi:10.1109/TCYB.2015.2415032.  
URL <https://doi.org/10.1109/TCYB.2015.2415032>
- [21] C. Smutz, A. Stavrou, Malicious PDF detection using metadata and structural features, in: R. H. Zakon (Ed.), 28th Annual Computer Security Applications Conference, ACSAC 2012, Orlando, FL, USA, 3-7 December 2012, ACM, 2012, pp. 239–248. doi:10.1145/2420950.2420987.  
URL <https://doi.org/10.1145/2420950.2420987>
- [22] G. Apruzzese, M. Colajanni, M. Marchetti, Evaluating the effectiveness of adversarial attacks against botnet detectors, in: A. Gkoulalas-Divanis, M. Marchetti, D. R. Avresky (Eds.), 18th IEEE International Symposium on Network Computing and Applications, NCA 2019, Cambridge, MA, USA, September 26-28, 2019, IEEE, 2019, pp. 1–8. doi:10.1109/NCA.2019.8935039.  
URL <https://doi.org/10.1109/NCA.2019.8935039>
- [23] N. Carlini, D. A. Wagner, Adversarial examples are not easily detected: Bypassing ten detection methods, in: B. M. Thuraisingham, B. Biggio, D. M. Freeman, B. Miller, A. Sinha (Eds.), Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec@CCS 2017, Dallas, TX, USA, November 3, 2017, ACM, 2017, pp. 3–14. doi:10.1145/3128572.3140444.  
URL <https://doi.org/10.1145/3128572.3140444>
- [24] J. Lu, T. Issaranon, D. A. Forsyth, Safetynet: Detecting and rejecting adversarial examples robustly, in: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, IEEE Computer Society, 2017, pp. 446–454. doi:10.1109/ICCV.2017.56.  
URL <https://doi.org/10.1109/ICCV.2017.56>
- [25] D. J. Miller, Y. Wang, G. Kesidis, When not to classify: Anomaly detection of attacks (ADA) on DNN classifiers at test time, *Neural Comput.* 31 (8) (2019) 1624–1670. doi:10.1162/neco\_a\_01209.  
URL [https://doi.org/10.1162/neco\\_a\\_01209](https://doi.org/10.1162/neco_a_01209)
- [26] I. Debicha, T. Debatty, J. Dricot, W. Mees, T. Kenaza, Detect & reject for transferability of black-box adversarial attacks against network intrusion detection systems, in: International Conference on Advances in Cyber Security, Springer, 2021, pp. 329–339. doi:10.1007/978-981-16-8059-5\_20.  
URL [https://link.springer.com/chapter/10.1007/978-981-16-8059-5\\_20](https://link.springer.com/chapter/10.1007/978-981-16-8059-5_20)
- [27] M. Pawlicki, M. Choras, R. Kozik, Defending network intrusion detection systems against adversarial evasion attacks, *Future Gener. Comput. Syst.* 110 (2020) 148–154. doi:10.1016/j.future.2020.04.013.  
URL <https://doi.org/10.1016/j.future.2020.04.013>
- [28] G. Apruzzese, M. Andreolini, M. Marchetti, V. G. Colacino, G. Russo, Appcon: Mitigating evasion attacks to ML cyber detectors, *Symmetry* 12 (4) (2020) 653. doi:10.3390/sym12040653.  
URL <https://doi.org/10.3390/sym12040653>
- [29] J. H. Metzen, T. Genewein, V. Fischer, B. Bischoff, On detecting adversarial perturbations, in: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings, 2017.  
URL <https://arxiv.org/abs/1702.04267>
- [30] A. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images (2009).  
URL <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.222.9220&rep=rep1&type=pdf>
- [31] J. Deng, W. Dong, R. Socher, L. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA, IEEE Computer Society, 2009, pp. 248–255. doi:10.1109/CVPR.2009.5206848.  
URL <https://doi.org/10.1109/CVPR.2009.5206848>
- [32] Z. Li, D. Hoiem, Learning without forgetting, *IEEE Trans. Pattern Anal. Mach. Intell.* 40 (12) (2018) 2935–2947. doi:10.1109/TPAMI.2017.2773081.  
URL <https://doi.org/10.1109/TPAMI.2017.2773081>
- [33] M. A. Mohandes, M. A. Deriche, S. O. Aliyu, Classifiers combination techniques: A comprehensive review, *IEEE Access* 6 (2018) 19626–19639. doi:10.1109/ACCESS.2018.2813079.  
URL <https://doi.org/10.1109/ACCESS.2018.2813079>
- [34] A. P. Dempster, A generalization of bayesian inference, *Journal of the Royal Statistical Society: Series B (Methodological)* 30 (2) (1968) 205–232. doi:<https://doi.org/10.1111/j.2517-6161.1968.tb00722.x>.
- [35] M. Tavallaei, E. Bagheri, W. Lu, A. A. Ghorbani, A detailed analysis of the KDD CUP 99 data set, in: 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, CISDA 2009, Ottawa, Canada, July 8-10, 2009, IEEE, 2009, pp. 1–6. doi:10.1109/CISDA.2009.5356528.  
URL <https://doi.org/10.1109/CISDA.2009.5356528>
- [36] I. Sharafaldin, A. H. Lashkari, A. A. Ghorbani, Toward generating a new intrusion detection dataset and intrusion traffic characterization, in:

- P. Mori, S. Furnell, O. Camp (Eds.), Proceedings of the 4th International Conference on Information Systems Security and Privacy, ICISSP 2018, Funchal, Madeira - Portugal, January 22-24, 2018, SciTePress, 2018, pp. 108–116. doi:10.5220/0006639801080116.  
URL <https://doi.org/10.5220/0006639801080116>
- [37] F. Pierazzi, F. Pendlebury, J. Cortellazzi, L. Cavallaro, Intriguing properties of adversarial ML attacks in the problem space, in: 2020 IEEE Symposium on Security and Privacy, SP 2020, San Francisco, CA, USA, May 18-21, 2020, IEEE, 2020, pp. 1332–1349. doi:10.1109/SP40000.2020.00073.  
URL <https://doi.org/10.1109/SP40000.2020.00073>
- [38] G. Apruzzese, M. Andreolini, L. Ferretti, M. Marchetti, M. Colajanni, Modeling realistic adversarial attacks against network intrusion detection systems, Digital Threats: Research and Practice (2021). doi:10.1145/3469659.  
URL <https://doi.org/10.1145/3469659>
- [39] D. Han, Z. Wang, Y. Zhong, W. Chen, J. Yang, S. Lu, X. Shi, X. Yin, Evaluating and improving adversarial robustness of machine learning-based network intrusion detectors, IEEE J. Sel. Areas Commun. 39 (8) (2021) 2632–2647. doi:10.1109/JSAC.2021.3087242.  
URL <https://doi.org/10.1109/JSAC.2021.3087242>
- [40] M. A. Merzouk, F. Cuppens, N. Boulahia-Cuppens, R. Yaich, Investigating the practicality of adversarial evasion attacks on network intrusion detection, Annals of Telecommunications (2022) 1–13doi:10.1007/s12243-022-00910-1.  
URL <https://doi.org/10.1007/s12243-022-00910-1>
- [41] Y. Mirsky, T. Doitshman, Y. Elovici, A. Shabtai, Kitsune: An ensemble of autoencoders for online network intrusion detection, in: 25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-21, 2018, The Internet Society, 2018.  
URL [http://wp.internetsociety.org/ndss/wp-content/uploads/sites/25/2018/02/ndss2018\\_03A-3\\_Mirsky\\_paper.pdf](http://wp.internetsociety.org/ndss/wp-content/uploads/sites/25/2018/02/ndss2018_03A-3_Mirsky_paper.pdf)
- [42] R. A. Khamis, A. Matrawy, 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 2020, Montreal, QC, Canada, October 20-22, 2020, IEEE, 2020, pp. 1–6. doi:10.1109/ISNCC49221.2020.9297344.  
URL <https://doi.org/10.1109/ISNCC49221.2020.9297344>
- [43] J. Clements, Y. Yang, A. A. Sharma, H. Hu, Y. Lao, Rallying adversarial techniques against deep learning for network security, in: IEEE Symposium Series on Computational Intelligence, SSCI 2021, Orlando, FL, USA, December 5-7, 2021, IEEE, 2021, pp. 1–8. doi:10.1109/SSCI50451.2021.9660011.  
URL <https://doi.org/10.1109/SSCI50451.2021.9660011>



**Islam Debicha** received his master's degree in Computer Science with a focus in network security in 2018. He is pursuing a joint Ph.D. in machine learning-based intrusion detection systems at ULB and ERM, Brussels, Belgium. He works at ULB cybersecurity Research Center and Cyber Defence Lab on evasion attacks against machine learning-based intrusion detection systems. He is the author or co-author of 5 peer-reviewed scientific publications. His research interests include defenses against adversarial examples, machine learning applications in cybersecurity, data fusion, and network security.



and adversarial learning.

**Richard Bauwens** received his Master's degree in Cybersecurity with a focus on System Design and Analysis from the Université Libre de Bruxelles (ULB). He works on security automation and threat detection for a private company in Brussels. His work and research interests include security automation, fileless malware, automatic threat detection and handling,



**Thibault Debatty** obtained a master's degree in applied engineering sciences at the Royal Military Academy (RMA) in Belgium, followed by a master's degree in applied computer science at the Vrije Universiteit Brussel (VUB). Finally, he obtained a Ph.D. with a specialization in distributed computing at both Telecom Paris and the RMA. He is now an associate professor at the RMA, where he teaches courses in networking, distributed information systems, and information security. He is also president of the jury of the Master of Science in Cybersecurity organized by the Université Libre de Bruxelles (ULB), the RMA, and four other institutions.



**Jean-Michel Dricot** leads research on network security with a specific focus on the IoT and wireless networks. He teaches communication networks, mobile networks, the internet of things, and network security. Prior to his tenure at the ULB, Jean-Michel Dricot obtained a Ph.D. in network engineering, with a focus on wireless sensor network protocols and architectures. In 2010, Jean-Michel Dricot was appointed professor at the Université Libre de Bruxelles, with tenure in mobile and wireless networks. He is the author or co-author of more than 100+ papers published in peer-reviewed international Journals and Conferences and served as a reviewer for European projects.



**Tayeb Kenaza** received a Ph.D. degree in computer science from Artois University, France, in 2011. Since 2017, he is the Head of the Computer Security Laboratory at the Military Polytechnic School of Algiers. He is currently a senior lecturer in Computer Science Department at the same school. His research and publication interests include Computer Network Security, Wireless Communication Security, Intelligent Systems, and Data Mining.



**Wim Mees** is Professor in computer science and cyber security at the Royal Military Academy and is leading the Cyber Defence Lab. He is also teaching in the Belgian inter-university Master in Cybersecurity, and in the Master in Enterprise Architecture at the IC Institute. Wim has participated in and coordinated numerous national and European research projects as well EDA and NATO projects and task groups.