PhD student : DEBICHA Islam

Email: debichasislam@gmail.com

UNIVERSITÉ LIBRE DE BRUXELLES

Adversarial Training for Deep Learning-based Intrusion Detection Systems

Islam Debicha, Thibault Debatty, Jean-Michel Dricot, Wim Mees

IARIA

About the presenter

Currently doing a joint PhD between ERM and ULB about Intrusion detection.

Worked before as a network security engineer.

subjects of interest: machine learning & network security.

Outline

1. Introduction
2. What is an adversarial attack?
3. Effect of adversarial attacks on Intrusion detection systems
4. Adversarial Training as a defense
5. Future work

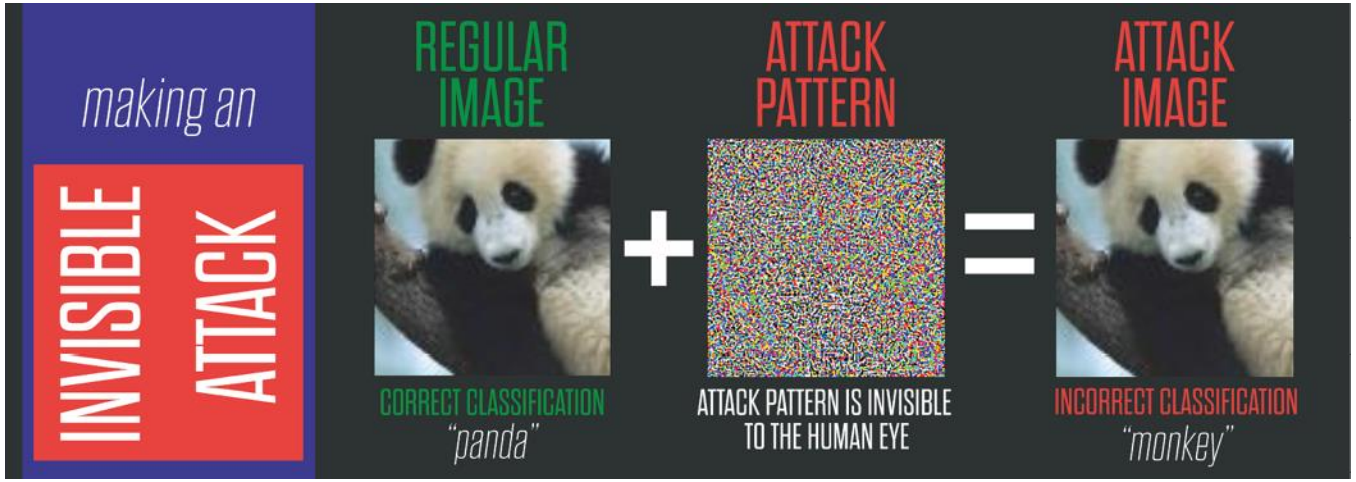
Adversarial Attacks against Intrusion Detection Systems

- **Deep Learning** is the state-of-the-art classification method used for **anomaly-based intrusion detection.**

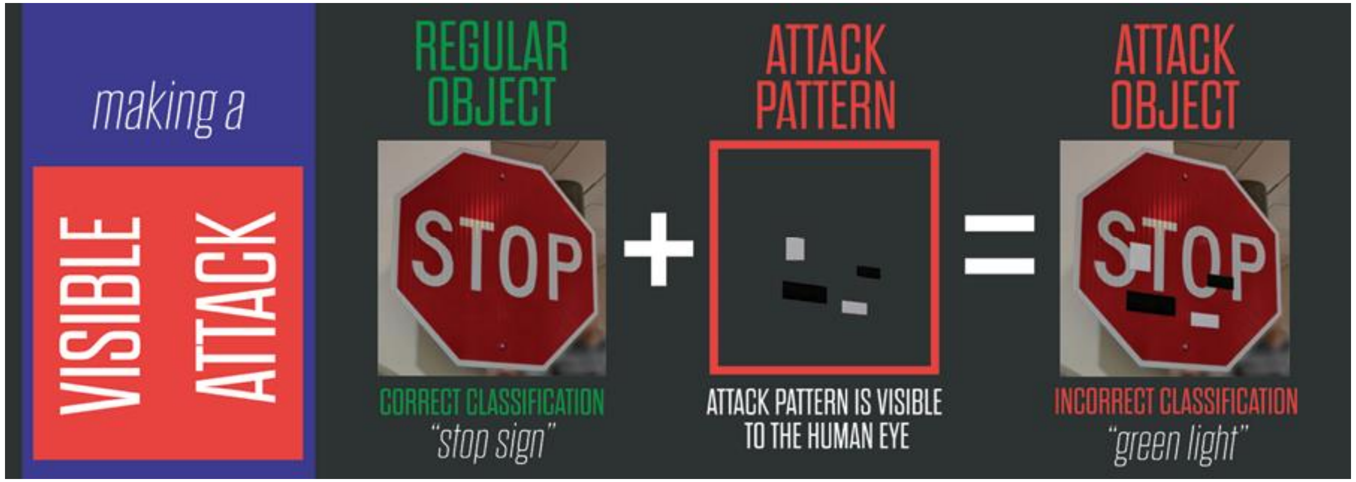
- Recent research has revealed that Deep Learning is **vulnerable** to specifically crafted attacks called "**Adversarial Attacks**".

- Most of these attacks were created for **computer vision**, therefore it is interesting to evaluate the **effectiveness** of these attacks against **intrusion detection systems** and possible defenses.

Adversarial attacks against computer vision systems



For the human eye, the **two images are identical**, but the specially crafted distortion, albeit small, leads the system to **classify them differently**.

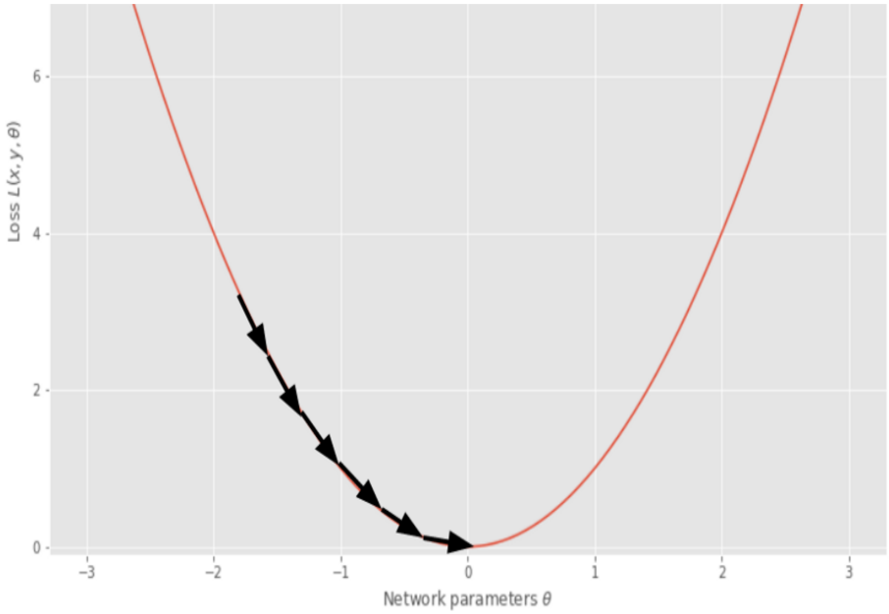


Imagine how dangerous would it be if an autonomous car recognized a ”stop sign” as ”green light” !!

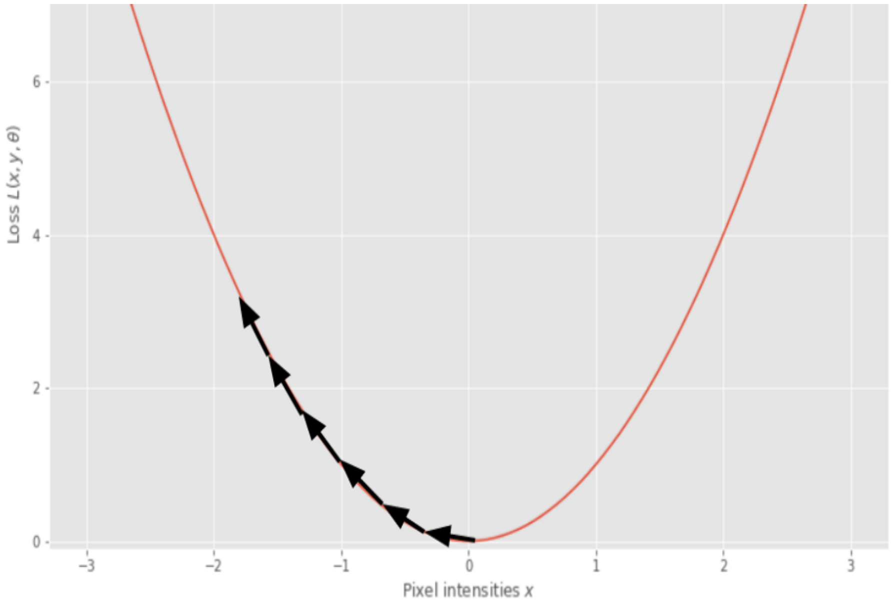
Source: https://www.belfercenter.org/publication/AttackingAI

What is an Adversarial Attack?

- The secret behind deep learning success is **gradient descent**.
- Given X and Y, we keep **changing model parameters $\boldsymbol{\theta}$** to make the **loss function J(X,$\boldsymbol{\theta}$,Y)** as small as possible.
- An attacker, in the other hand, will keep **changing input data X** to make the **loss function J(X,$\boldsymbol{\theta}$,Y)** as big as possible.



gradient descent : changing θ



gradient ascent (Attack) : changing X

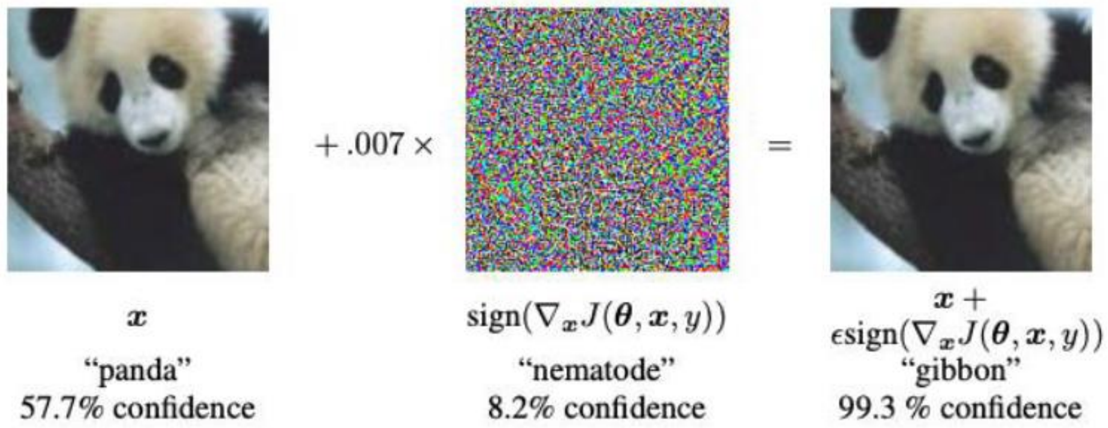
FGSM attack as an example

- The fast gradient sign method (FGSM) works by using the **gradients** of the neural network to create an adversarial example.
- For an input image, the method **uses the gradients of the loss with respect to the input image** to create a new image that **maximizes the loss**.

$$adv\_x = x + \epsilon * \text{sign}(\nabla_x J(\theta, x, y))$$

where

- adv_x : Adversarial image.
- x : Original input image.
- y : Original input label.
- ϵ : Multiplier to ensure the perturbations are small.
- θ : Model parameters.
- J : Loss.

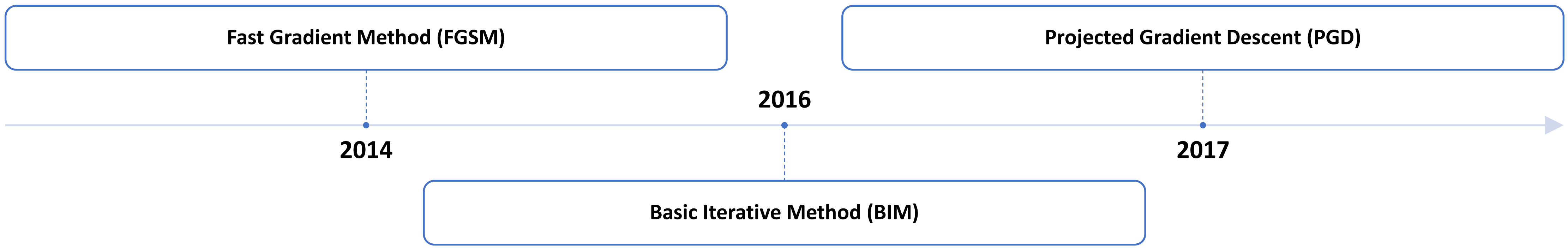


Source: Explaining and Harnessing Adversarial Examples, Goodfellow et al, ICLR 2015.

NSL-KDD dataset to test adversarial attacks

- NSL-KDD dataset is widely used to **evaluate the performance** of an intrusion detection system.
- This dataset covers several attacks organized into four classes according to their nature: Dos, Probe, R2L and U2R.
- Records in the NSL-KDD dataset have 41 features in addition to a class label. These features are grouped into three categories: Basic features, Content features and Traffic features.

Adversarial attacks used in this experiment

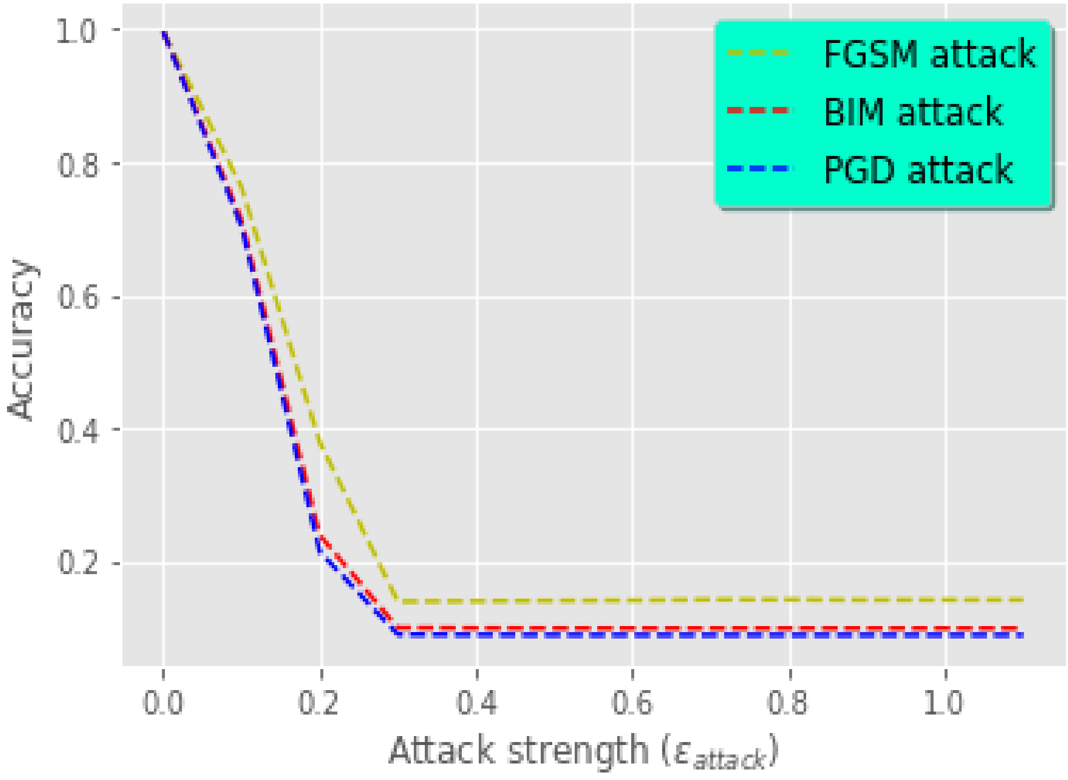


They all share a parameter called “**epsilon**” which determine the **strength of the attack.**

$$\|\boldsymbol{x}' - \boldsymbol{x}\|_p < \epsilon$$

i.e. the distance between the adversarial and original example is less than some small epsilon under a particular norm p.

Effect of adversarial attacks on Intrusion detection systems



Effect of adversarial attacks on deep learning-based intrusion detection system.

- The model accuracy was **99.61%** before the attacks.
- With sufficient distortion, adversarial attacks can **defeat** intrusion detection systems and **lead them into misdetection**.

Adversarial Training as a defense

- The current state-of-the-art defense against adversarial attack is adversarial training.
- Adversarial training is simply **putting adversarial samples inside the training loop**.
- In adversarial training we are minimizing the following loss function where Δ is a set of perturbations to which we want our model to be invariant.

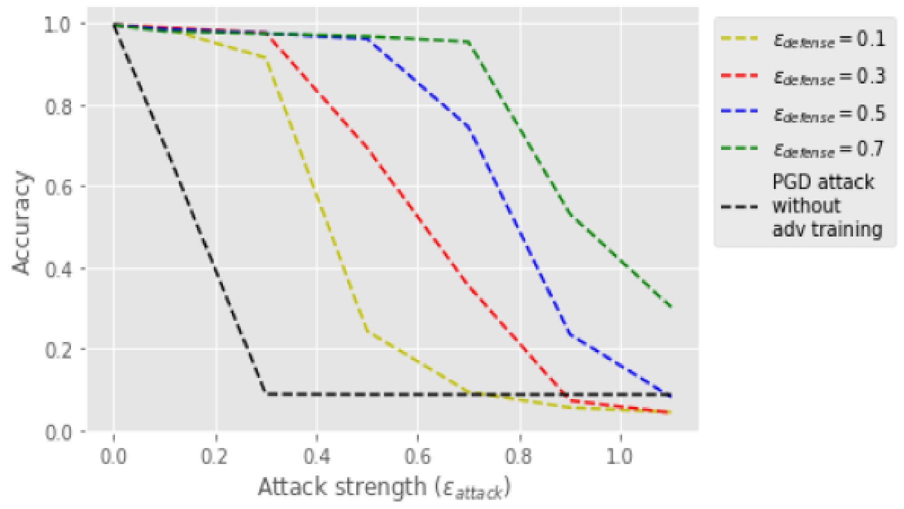
$$\min_{\theta} \max_{\delta \in \Delta} \mathcal{L}(x + \delta, y; \theta)$$

Adversarial Training as a defense

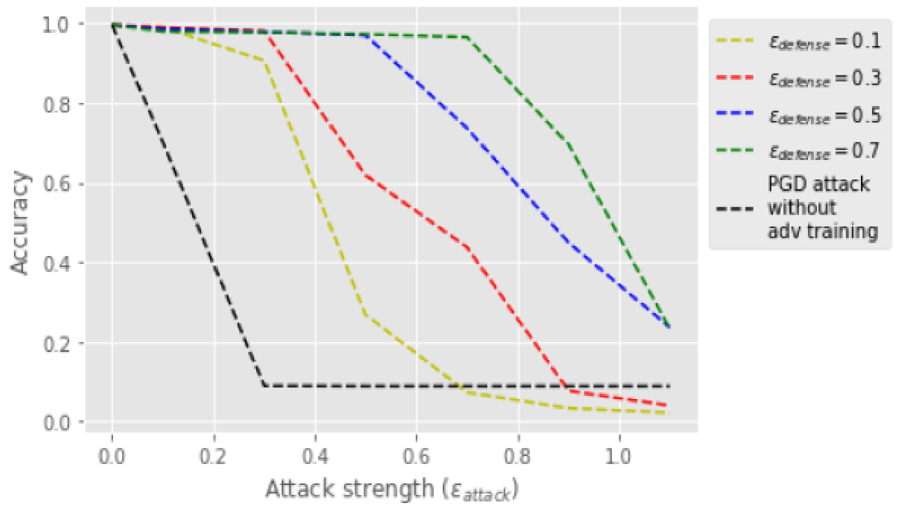
Adversarial training means using an attack method to create adversarial records and mix them with clean training record.

But:

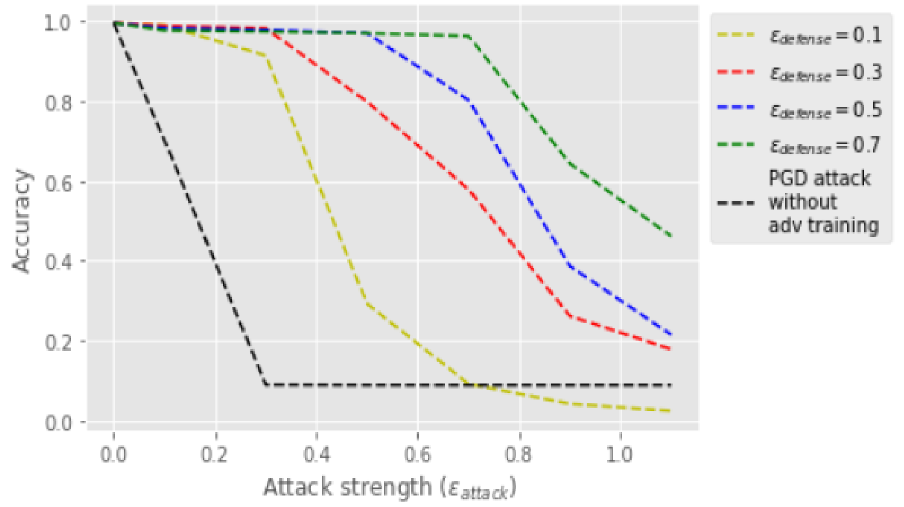
1. What is the best **ratio** between adversarial training records and clean training record?
2. How much distortion **(epsilon)** should be used to create adversarial training records?
3. Does adversarial training effect the accuracy of IDS on **clean test data**?



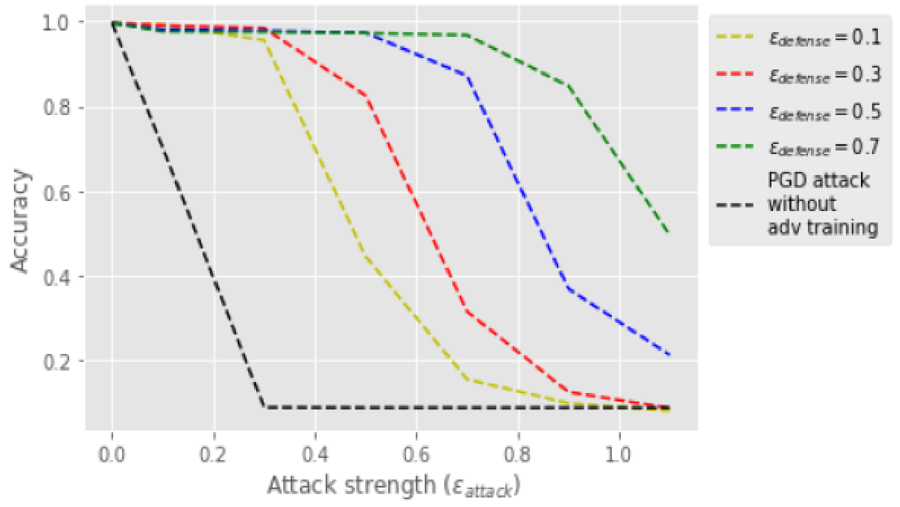
(a) Percentage of adversarial training samples in the training data = 30%



(b) Percentage of adversarial training samples in the training data = 50%



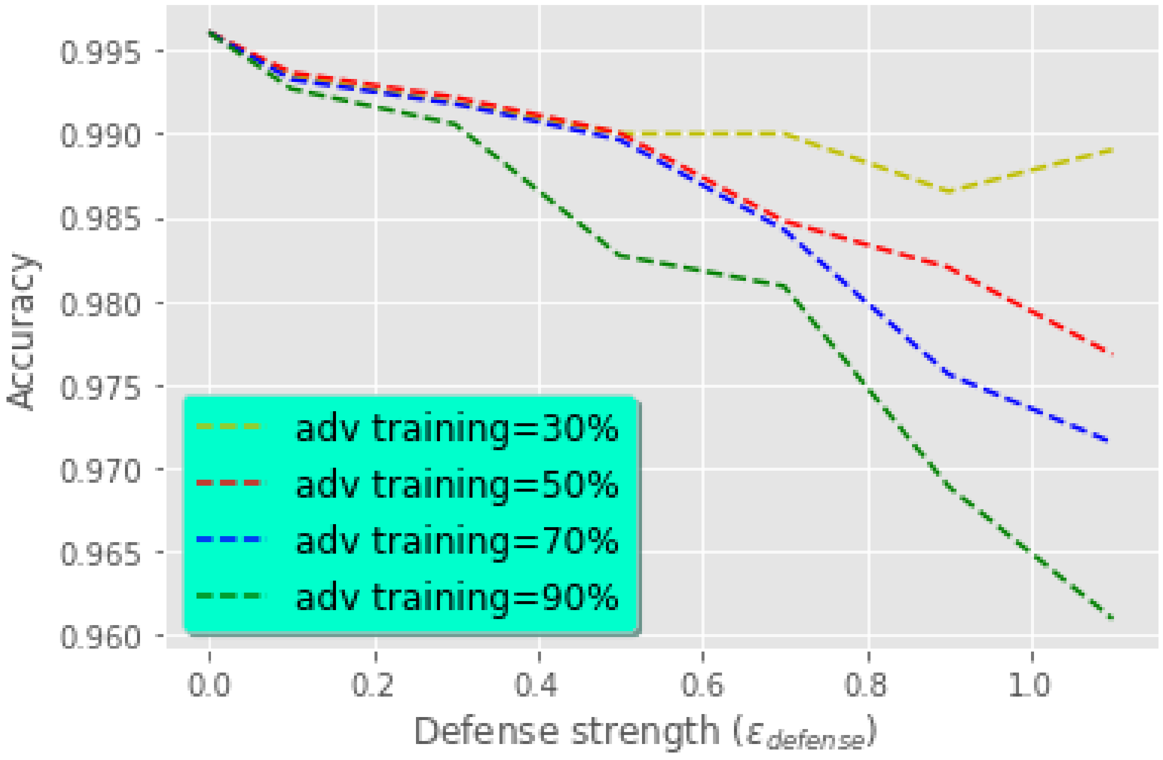
(c) Percentage of adversarial training samples in the training data = 70%



(d) Percentage of adversarial training samples in the training data = 90%

Effect of adversarial training on the robustness of intrusion detection systems

Trade-off between robustness and accuracy



Effect of adversarial training on the performance of the intrusion detection system on **clean test data**.

While results of the previous experiments indicate that **adversarial training increases the robustness** of deep learning-based intrusion detection systems, it also **slightly decreases the accuracy** of the detector when tested on clean test data.

Future work

- Assess the transferability property of adversarial attacks on intrusion detection systems.
- Propose new defense mechanisms against adversarial attacks by exploring uncertainty handling techniques.