Cyber CNI PhD Day

Hassan Chaitou

chairecyber-cni.org/

Agenda

- Supervisory context
- Introduction to the topic
- Research objective
- Ongoing works
- Conclusions and future work

Supervisory context
For doctoral students

Optimization of security risk for learning on heterogeneous data

Supervision:

- Laurent Pautet
- Thomas Robert
- Jean Leneutre

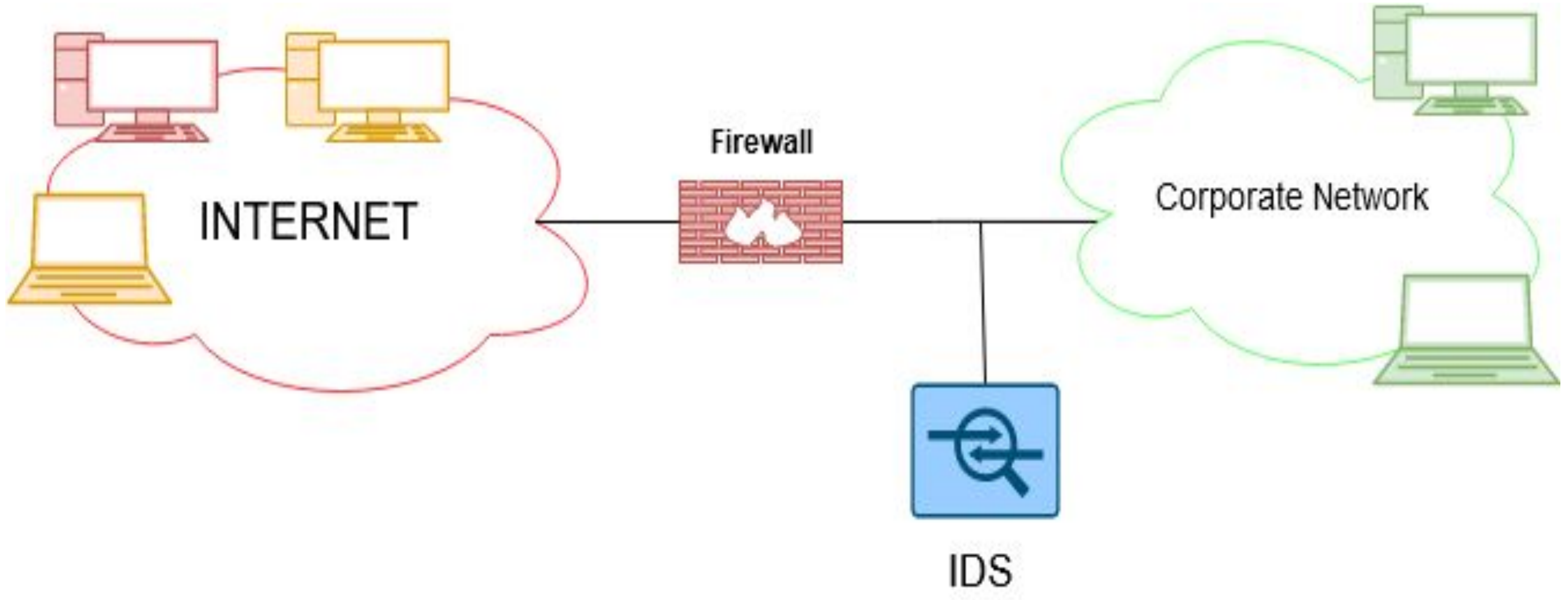
Progress of the thesis:

- Started 10/03/2020 - 09/03/2023

Introduction

An intrusion detection system (IDS) is a key component of the network security

- ❑ Misuse based IDS
- ❑ Anomaly based IDS



Machine learning techniques in IDS

Machine learning models are vulnerable to adversarial examples (Goodfellow, 2015)

Introduction

Adversarial attack

- White-box: an attacker has access to the parameter, algorithms, and structure of the target model (e.g. Fast Gradient Sign method (FGSM) , Carlini & Wagner attacks (Carlini and Wagner, 2016))
- Black-box: an attacker cannot obtain information about the target Model

Defense mechanisms against adversarial attacks

- Adversarial training (Goodfellow, 2015): the basic idea merely to create and then incorporates adversarial examples into the training process.
- Other methods: e.g. Gradient Hiding (Athalye , 2018) and Defensive Distillation (Papernot, 2015)

Assess the defense mechanism against the attack mechanism.

Research Objective

short-term objective: single IDS sensing

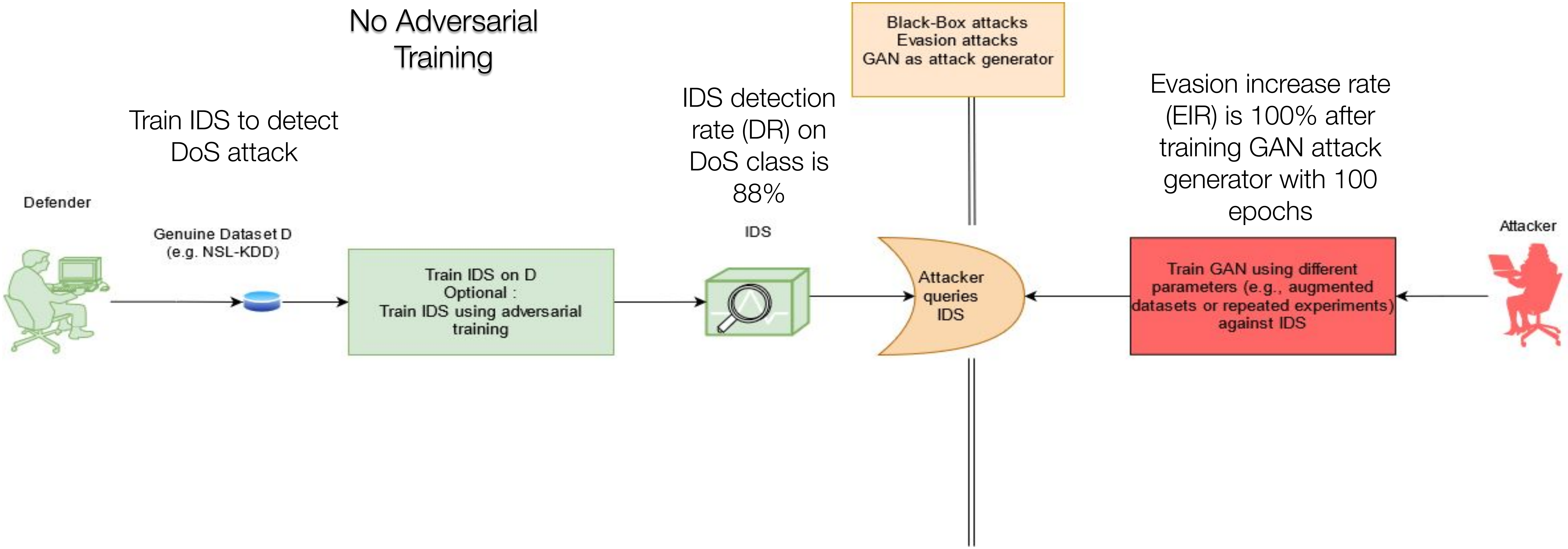
How does considering settings determined independently could impact performance either on the attacker or on the defender side?

- What are the risks associated with an attacker that trains an attack generator to perform repeated adversarial attacks against a system protected by an IDS strengthened with adversarial training?

- What is the impact of the resources invested by the attacker on the defender's performance (specifically when the attacker uses an augmented dataset in the training process)?

Long-term objective: multi-sensing defense architecture

Basic adversarial attack scenario



Evasion increase rate (EIR): the increase in the undetected adversarial malicious traffic examples by IDS compared with the original malicious traffic examples

Detection Rate (DR) refers to the proportion of correctly detected malicious traffic records by the black-box IDS to all of those attack records detected.

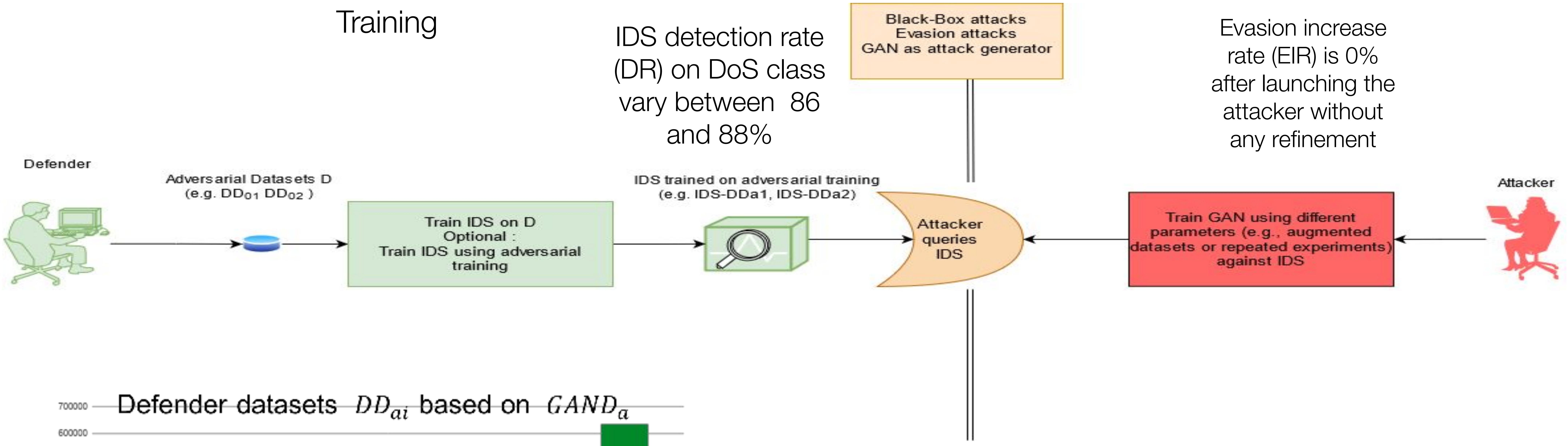
- NSL-KDD consists of five classes: Normal, Probe, DoS, U2R (user to root), and R2L (remote to local).

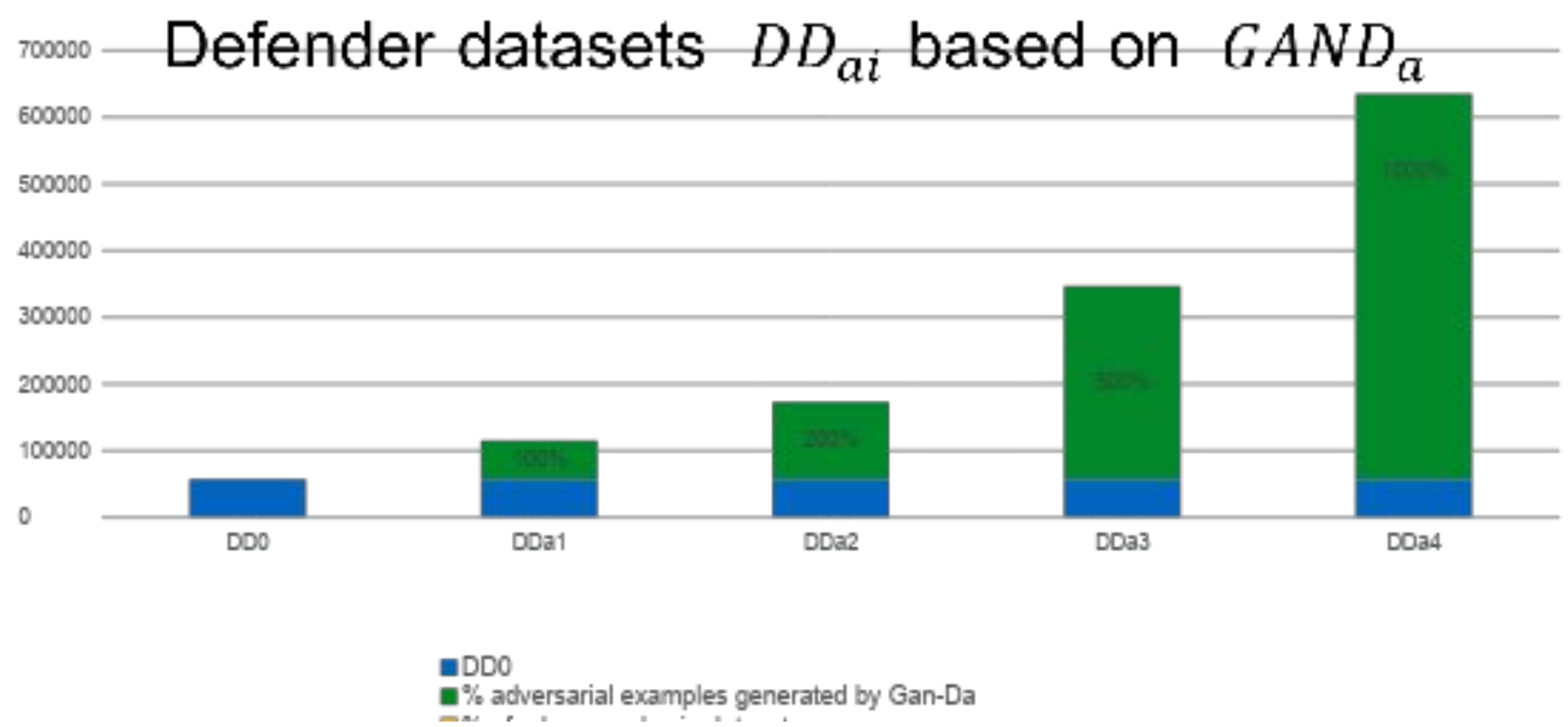
Robust IDS- adversarial training

With Adversarial Training

IDS detection rate (DR) on DoS class vary between 86 and 88%

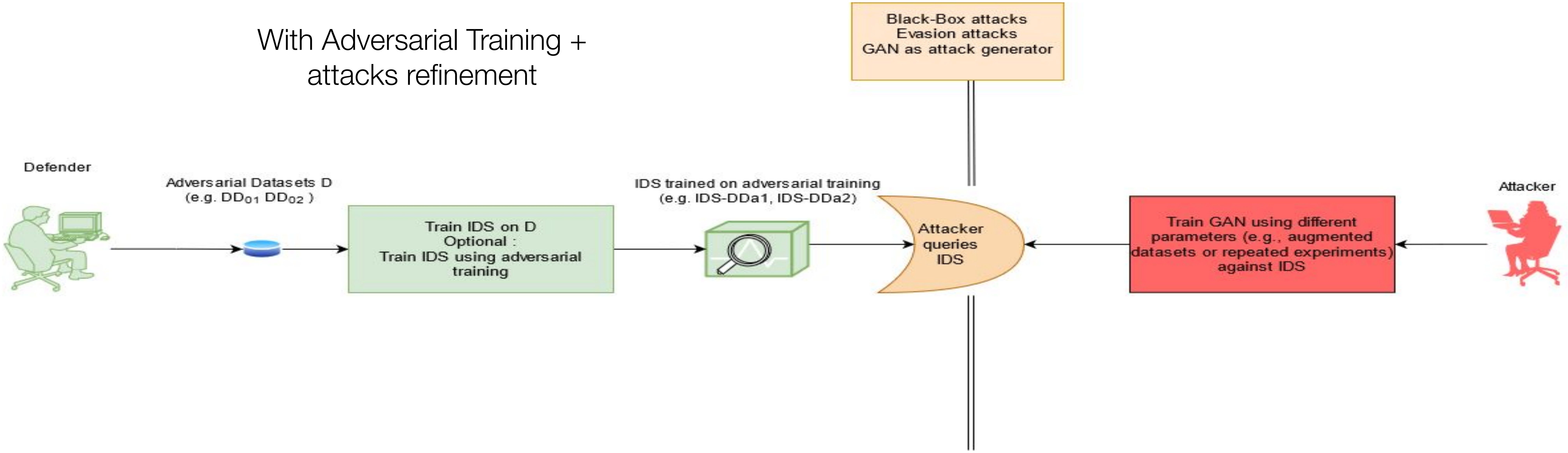
Evasion increase rate (EIR) is 0% after launching the attacker without any refinement





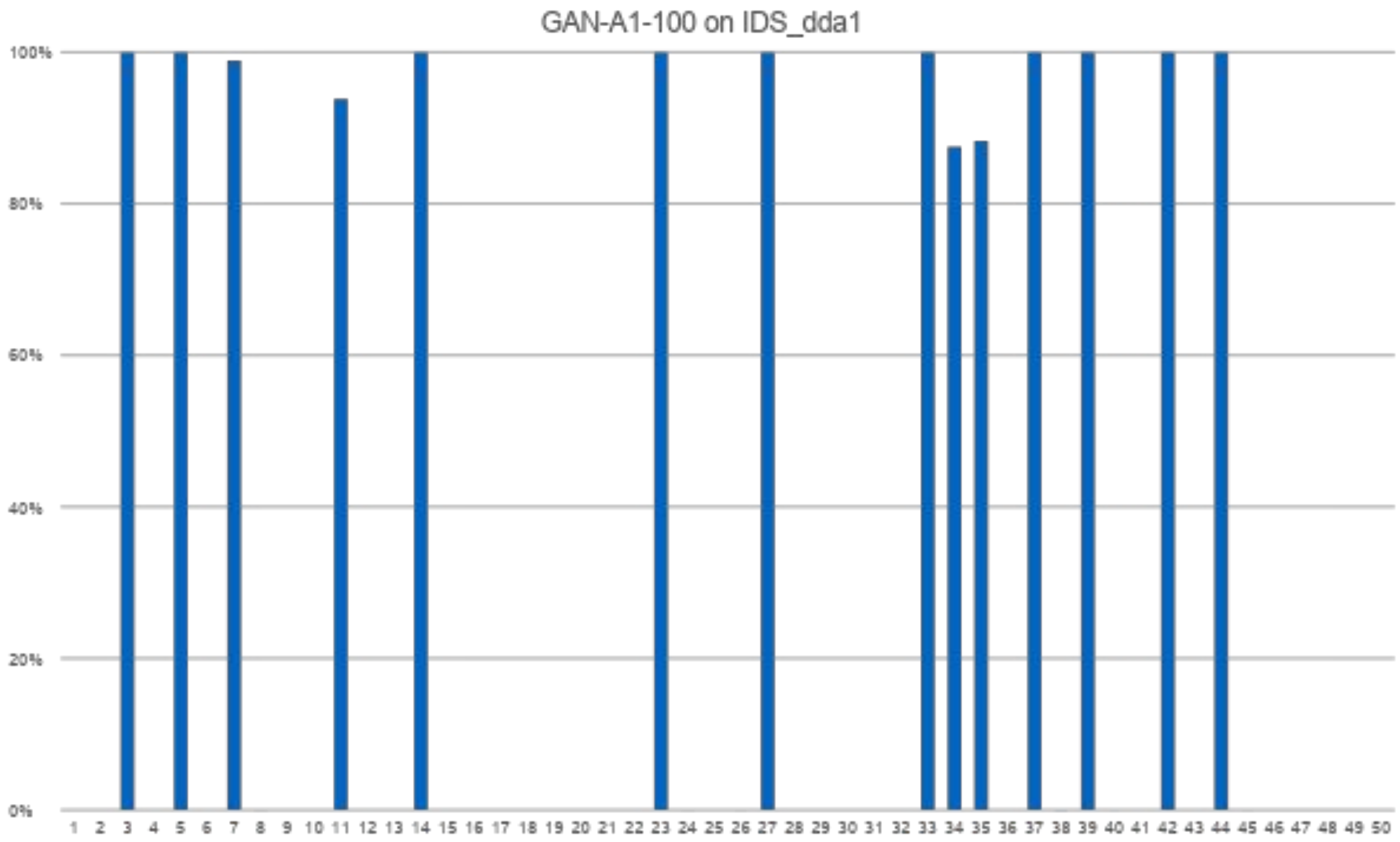
Attack refinement- increase computational resources

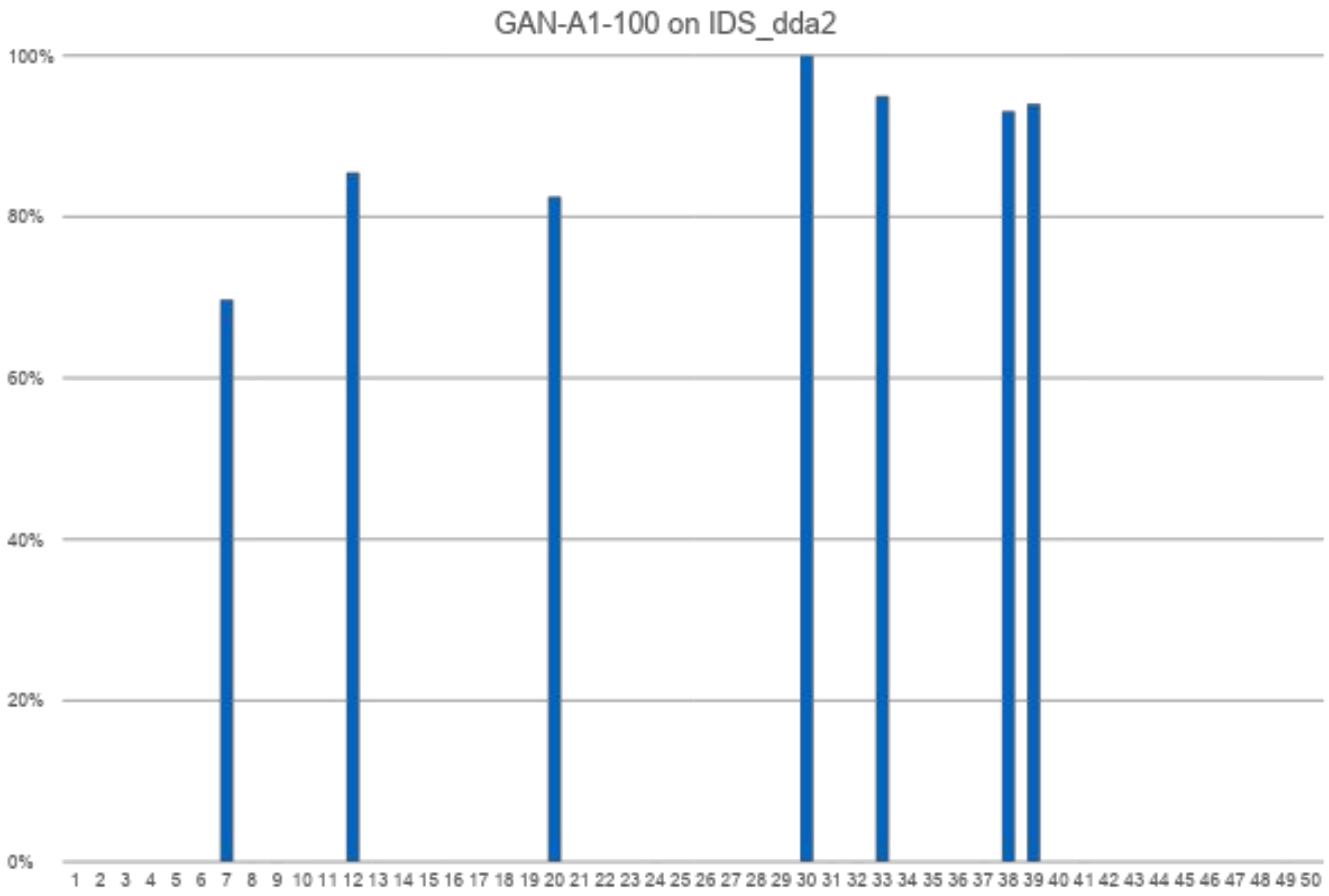
With Adversarial Training +
attacks refinement



- Modifying the computational time resources by training GAN on 100, 1000 and 5000 epochs.
- As a statistical test we repeat each experiments 50 times .

GAN-A1-100 results on IDS-DDai



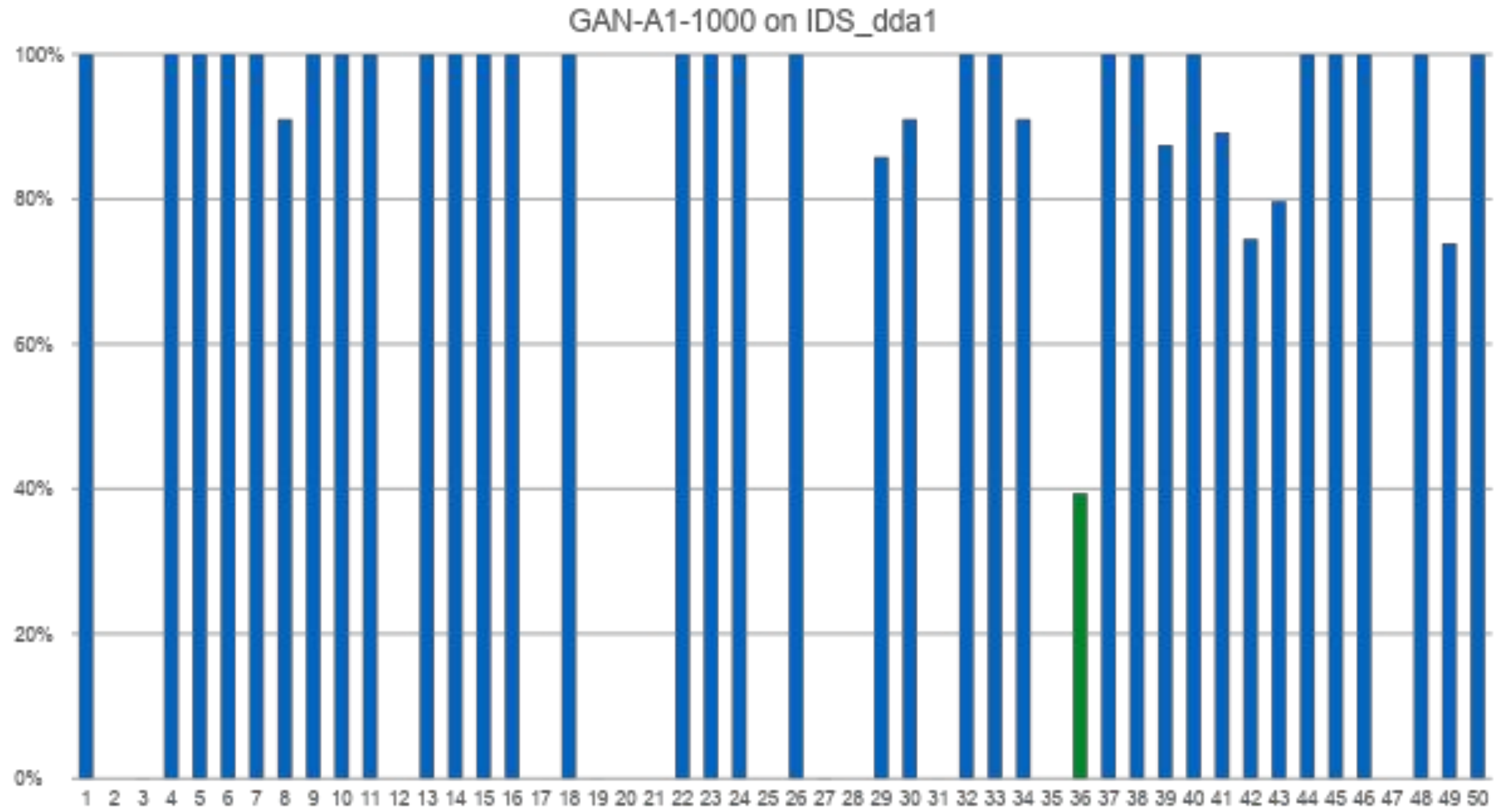


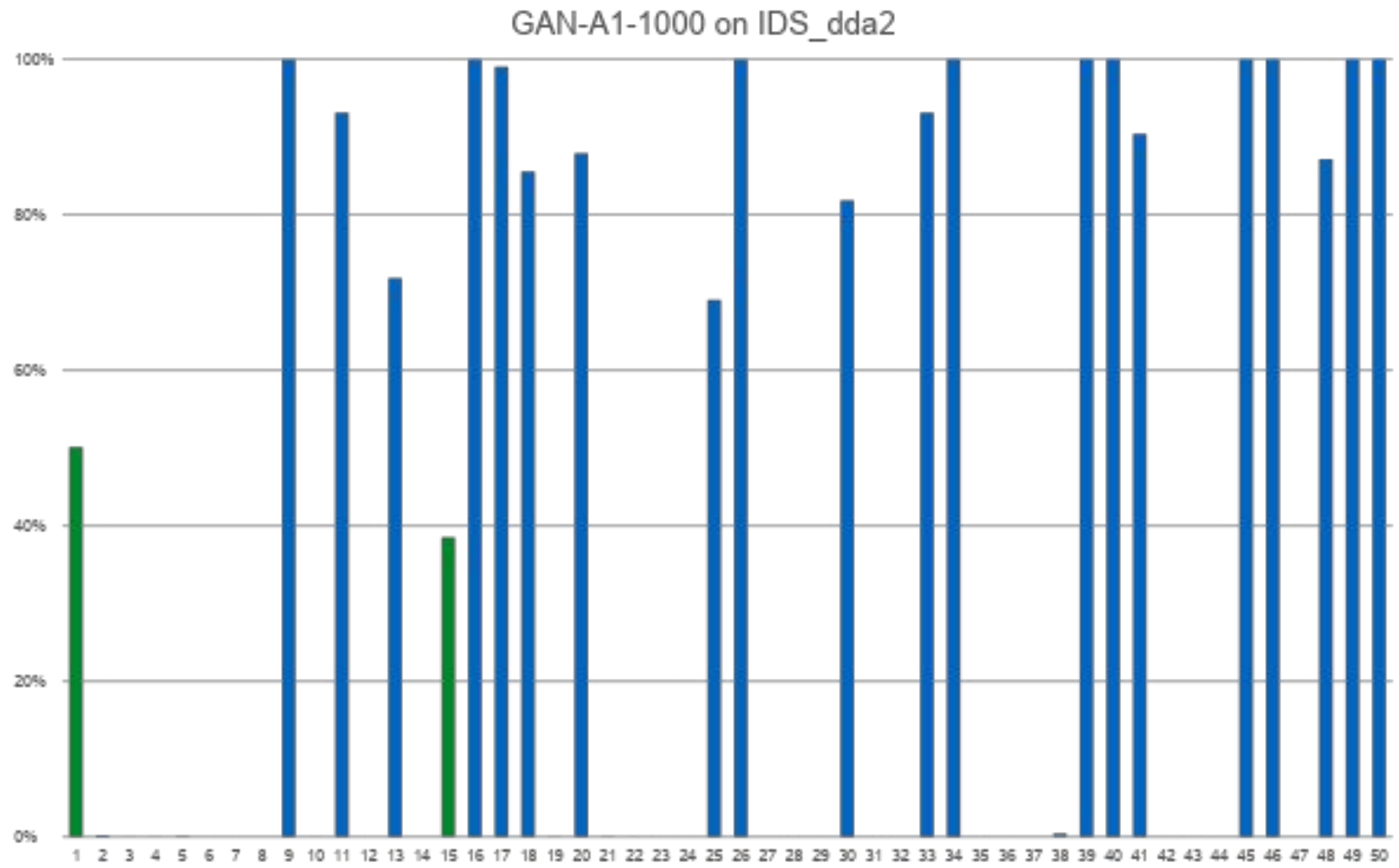
Attack success on IDS_DDa1: 14/50 with EIR > 85%

Attack success on IDS_DDa2: 7/50 with EIR > 70%

Attack success on IDS_DDa3 and on IDS_DDa4: 0/50

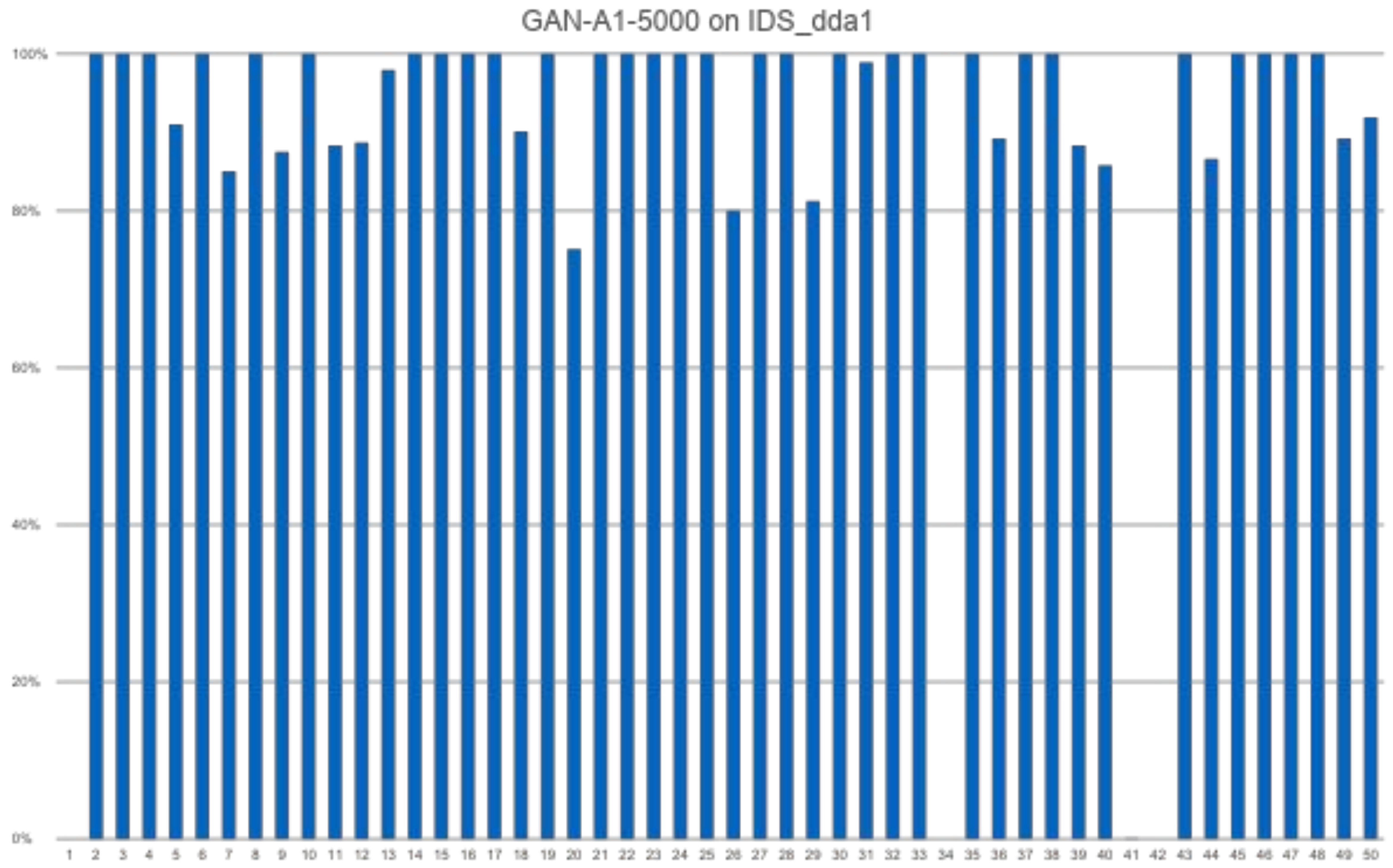
GAN-A1-1000 results on IDS-DDai

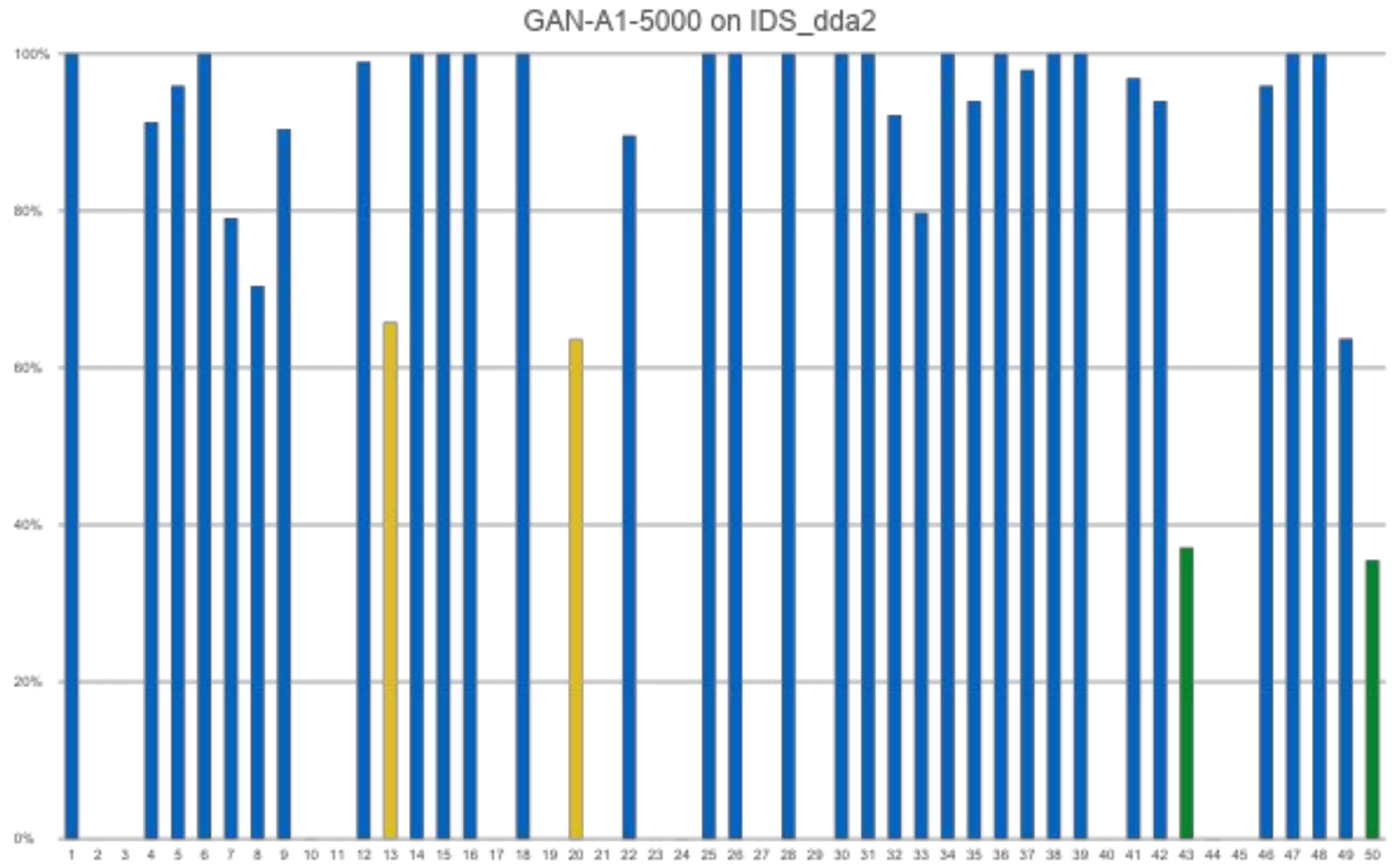




Attack success on IDS_DDa1: 36/50 with EIR > 80% and 1/50 with EIR = 39%
Attack success on IDS_DDa2: 20/50 with EIR > 70% , 1/50 with EIR = 50% and 1/50 with EIR = 39%
Attack success on IDS_DDa3: 2/50 with EIR > 85%
Attack success on IDS_DDa4: 0/50

GAN-A1-5000 results on IDS-DDai





Attack success on IDS_DDa1: 46/50 with EIR > 76%

Attack success on IDS_DDa2: 31/50 with EIR > 70% , 2/50 with EIR between [60, 69] and 2/50 with EIR between [35, 40]

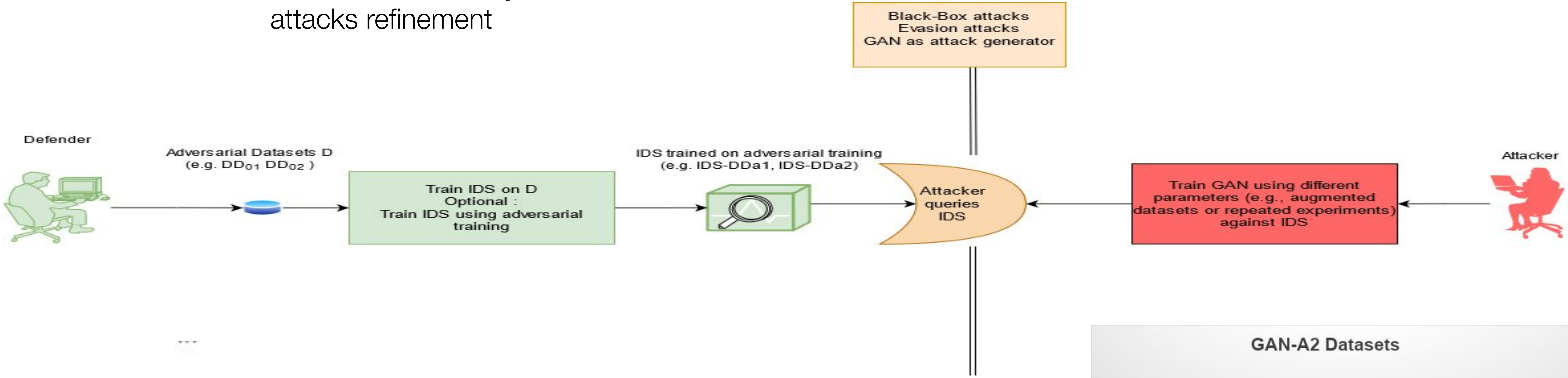
Attack success on IDS_DDa3: 7/50 with EIR > 80%

Attack success on IDS_DDa4: 2/50 with EIR = 100%

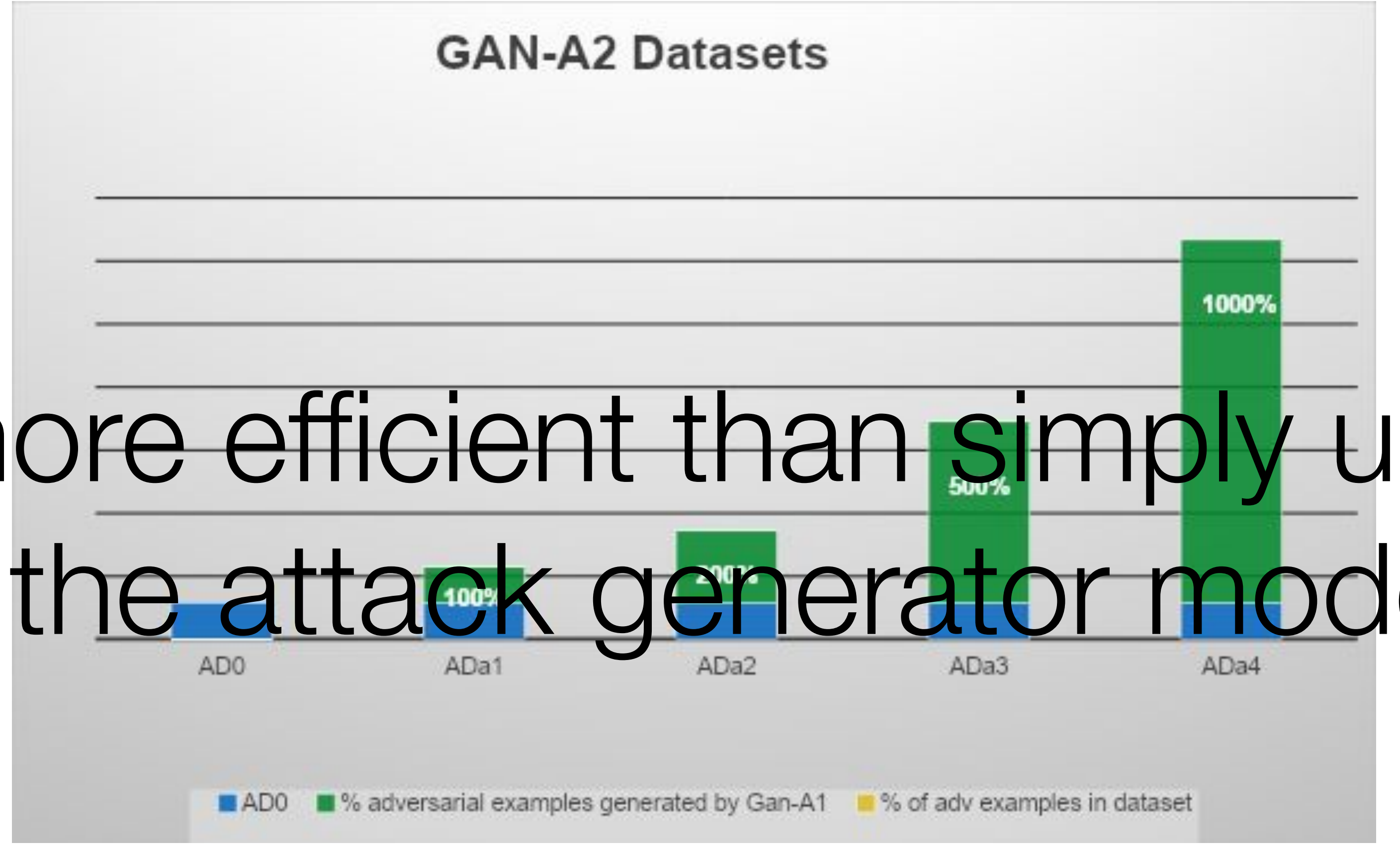
- Attack and defense scale against each other depending on resource spent
- Training process on attack side yield attack generator with almost a binary behavior (either very successful to evade detection or useless).

Attack refinement- train attack generator with augmented dataset

With Adversarial Training + attacks refinement



	$IDS-DD_{a1}$	$IDS-DD_{a2}$	$IDS-DD_{a3}$	$IDS-DD_{a4}$
$GAN-A_1-100$	11	6	0	0
$GAN-A_2-$[illegible]-100	21	8	0	0
$GAN-A_2-AD_{a2}-100$	30	10	1	0
$GAN-A_1-1000$	32	10	2	0
$GAN-A_2-AD_{a1}-1000$	35	20	4	0



Using adversarial training on attack generation means more efficient than simply using longer training processes (spend more time optimizing the attack generator model).

-) epochs.
-

Conclusions and Future work

- Attack and defense scale against each other depending on resource spent

- Explore feature space vs input space perturbations to achieve more realistic attack generators.

- **Formalize the experimental framework and try to improve its reuse.**

- Consider flow-based IDS, and distributed flow-based IDS, try to characterize their performance beyond testing on a given dataset.

References

- Explaining and Harnessing Adversarial Examples, Goodfellow et al
- Towards Evaluating the Robustness of Neural Networks, Carlini et al
- Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. Athalye et al
- Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks. Papernot et al
- Generative Adversarial Networks. Goodfellow et al