Using Machine Learning in Network Security: A New Investigation of Adversarial Evasion Attacks

Ashraf Matrawy
Carleton University
carleton.ca/ngn

Keynote at DRCN
Montreal, May 2024

Acknowledgement

I would like to acknowledge the contributions of my students to the research presented in these slides, Aboukhamis, Elshehaby, Ibitoye, and Kotha.

We work on ML in network security, security in IoT, 5G and beyond, misinformation, and usable security.

Please visit our group page for more information.
The Next Generation Networks Group `carleton.ca/ngn`

Outline

- ML is network security
- Adversarial attacks in network security
  - Our work on characterising adversarial attacks
  - Defences
- Introducing ACAT
- Gap between reality and research. The practicality question?

Core work

While our group produced significant work in this area, this presentation mostly covers our latest, in-progress work that is under review but could be found on arXiv [1], [2], [3]

Published work

- Differentially Private Self-normalizing Neural Networks for Adversarial Robustness in Federated Learning, 2022 [4].
- Temporal Partitioned Federated Learning for IoT Intrusion Detection Systems, 2024 [5].
- Could Min-Max Optimization Be A General Defense Against Adversarial Attacks?, 2024 [6].
- Evaluating Resilience of Encrypted Traffic Classification against Adversarial Evasion Attacks, 2021 [7]
- Evaluation of Adversarial Training on Different Types of Neural Networks in Deep Learning-based IDSs, 2020 [8].
- Investigating Resistance of Deep Learning-based IDS against Adversaries using min-max Optimization, 2020 [9].
- Analyzing Adversarial Attacks Against Deep Learning for Intrusion Detection in IoT Networks, 2019 [10].

ML in Network Security

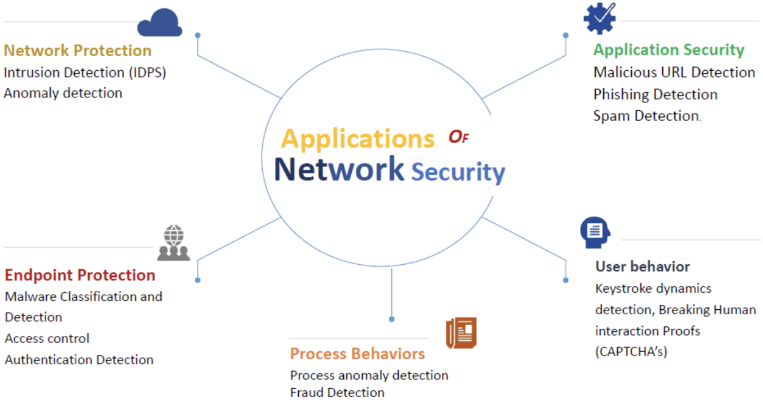


Figure: Applications of Network Security [1]

Types of Adversarial Attacks targets

Extended from the Work with Ibitoye, Aboukhamis, ElShehaby, and Shafiq [1].

- Different types of classifications.
- For the target
  - Evasion
  - Poisoning
  - Backdoor
  - Stealing
  - All of the work presented from now on is on evasion.
- **Feature vs Problem space [1]**
- Based on knowledge

Adversarial attacks - Evasion

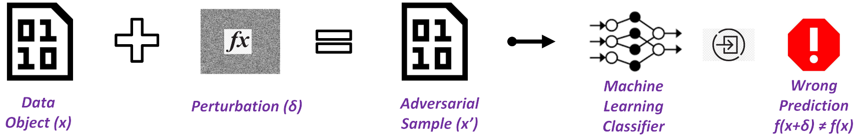


Figure: Evasion Adversarial Attack [1]

Our classification of adversarial attacks in network security

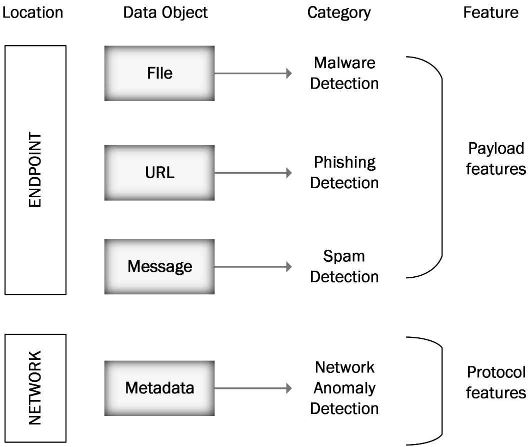


Figure: Adversarial attack classification [1]

Our work on characterising adversarial attacks

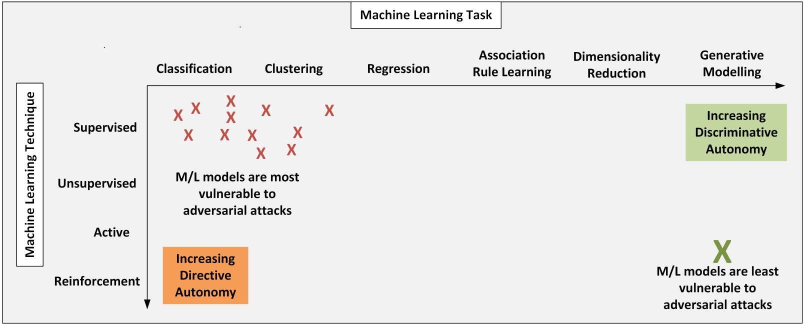

Machine Learning Task
Classification
Clustering
Regression
Association Rule Learning
Dimensionality Reduction
Generative Modelling
Machine Learning Technique
Supervised
Unsupervised
Active
Reinforcement
X X X X X X X X X X X X X
M/L models are most vulnerable to adversarial attacks
Increasing Discriminative Autonomy
Increasing Directive Autonomy
X
M/L models are least vulnerable to adversarial attacks

Figure: Adversarial Risk Grid Map [1]

Defenses against Adversarial attacks in network security

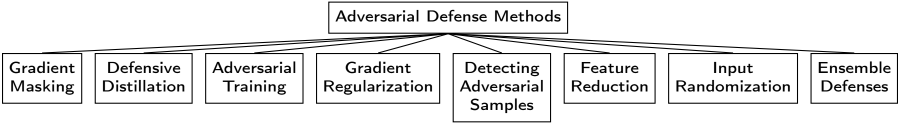


Figure: Adversarial Defense Methods [1]

Adversarial Training

- Defence using min-max [9]
- Checking the impact on different models [8]
- General defence? [6]

Adaptive Continuous Adversarial Training (ACAT)

ACAT is introduced by ElShehaby, Kotha and Matrawy [2].

- Acts as an adaptive defence that uses continuous training.
- Addresses the problem of the lack of data for adversarial training because it uses attack data for training.
- Reduces the total time of adversarial sample detection, especially in environments such as network security where the rate of attacks could be very high.
- Deals with catastrophic forgetting during periodic continuous training

Introducing ACAT

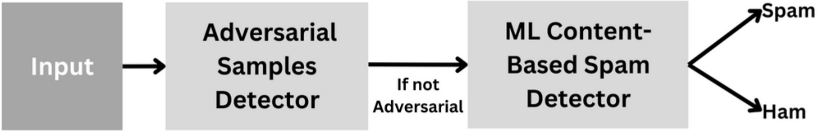


Figure: Conventional Detecting Approach [2]

Introducing ACAT

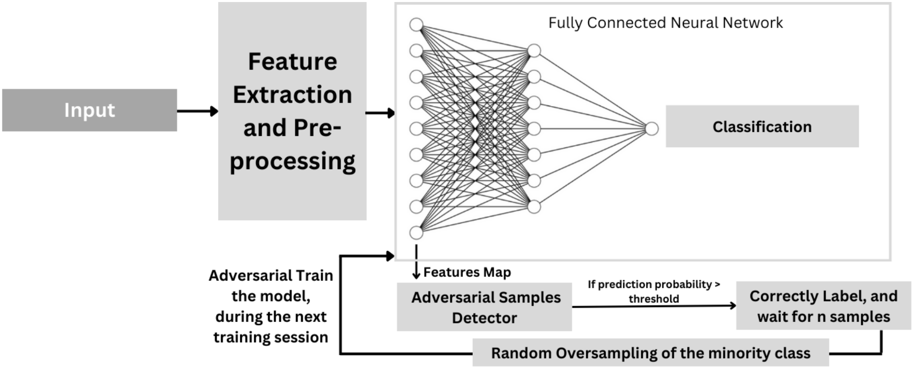

Input
Feature Extraction and Pre-processing
Fully Connected Neural Network
Classification
Features Map
Adversarial Samples Detector
If prediction probability > threshold
Correctly Label, and wait for n samples
Random Oversampling of the minority class
Adversarial Train the model, during the next training session

Figure: The Proposed Adaptive Continuous Adversarial Training (ACAT) [2]

Introducing ACAT

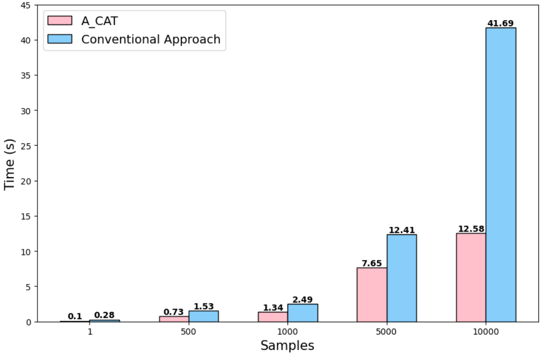


Figure: Prediction time of ACAT vs a Conventional Approach [2]

Introducing ACAT

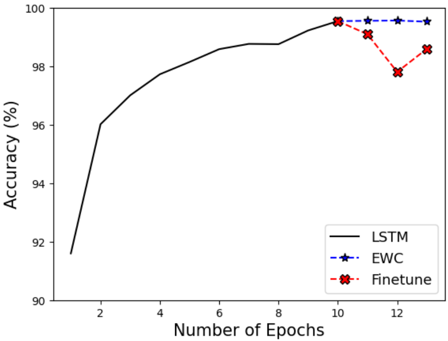


Figure: Accuracy of Fine Tuning vs EWC on the original training set (without adversarial perturbations) during Adversarial Continuous Training. The solid black line represents the 10 training epochs before deployment, while the dotted lines represent the accuracy after each adversarial training session. [2]

Introducing ACAT

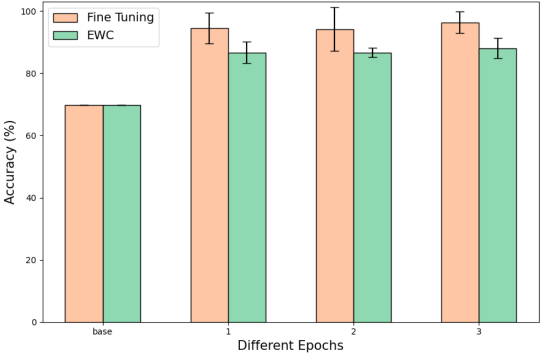
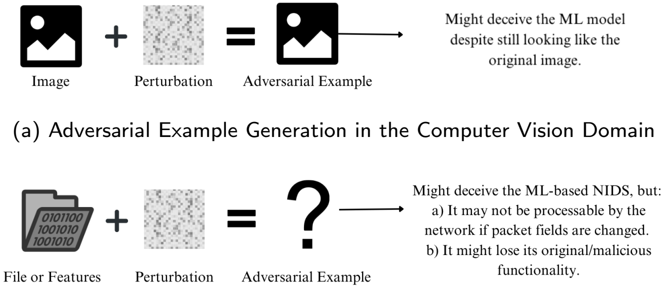


Figure: Effect of Continuous Adversarial Training: Accuracy Comparison on (D_{test}) of perturbed Enron SPAM (Fine Tuning vs EWC) [2]

Adversarial attacks in networks: Are they different?

Work with ElShehaby [3].



Might deceive the ML model despite still looking like the original image.

(a) Adversarial Example Generation in the Computer Vision Domain

Might deceive the ML-based NIDS, but:
a) It may not be processable by the network if packet fields are changed.
b) It might lose its original/malicious functionality.

(b) Adversarial Example Generation in the Network Security Domain

Figure: Adversarial Examples Generation [3]

Gap between reality and research: The practicality question?

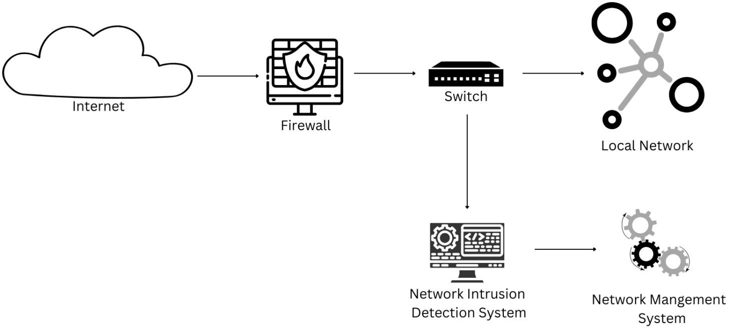
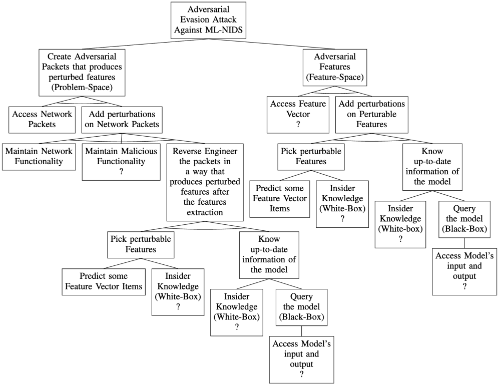


Figure: The Deployment of Network Intrusion Detection System [3]

Gap between reality and research. The practicality question?



Adversarial Evasion Attack Against ML-NIDS

Create Adversarial Packets that produces perturbed features (Problem-Space)

Adversarial Features (Feature-Space)

Access Network Packets

Add perturbations on Network Packets

Access Feature Vector ?

Add perturbations on Perturable Features

Maintain Network Functionality

Maintain Malicious Functionality ?

Reverse Engineer the packets in a way that produces perturbed features after the features extraction

Pick perturbable Features

Know up-to-date information of the model

Predict some Feature Vector Items

Insider Knowledge (White-Box) ?

Insider Knowledge (White-box) ?

Query the model (Black-Box)

Access Model's input and output ?

Pick perturbable Features

Know up-to-date information of the model

Predict some Feature Vector Items

Insider Knowledge (White-Box) ?

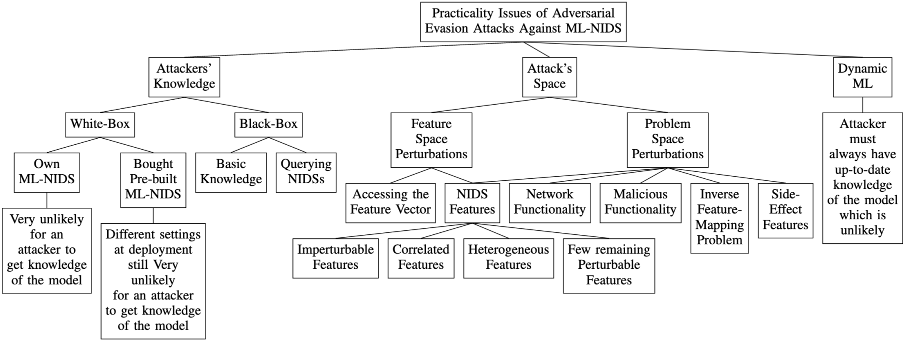
Insider Knowledge (White-Box) ?

Query the model (Black-Box)

Access Model's input and output ?

Figure: Attack Tree of Adversarial Evasion Attack Against ML-NIDS. $<$ indicates a disjunction (OR), ∢ indicates a conjunction (AND), and ? denotes a leaf node with uncertain feasibility (questionable practicality) [3]

Gap between reality and research. The practicality question?



Practicality Issues of Adversarial Evasion Attacks Against ML-NIDS

Attackers' Knowledge

Attack's Space

Dynamic ML

White-Box

Black-Box

Feature Space Perturbations

Problem Space Perturbations

Attacker must always have up-to-date knowledge of the model which is unlikely

Own ML-NIDS

Bought Pre-built ML-NIDS

Basic Knowledge

Querying NIDSs

Accessing the Feature Vector

NIDS Features

Network Functionality

Malicious Functionality

Inverse Feature-Mapping Problem

Side-Effect Features

Very unlikely for an attacker to get knowledge of the model

Different settings at deployment still Very unlikely for an attacker to get knowledge of the model

Imperturbable Features

Correlated Features

Heterogeneous Features

Few remaining Perturbable Features

Figure: Taxonomy of Practicality Isues of Adversarial Attacks Against ML-NIDS, Directed Acyclic Graph (DAG) [3]

Gap between reality and research. The practicality question?

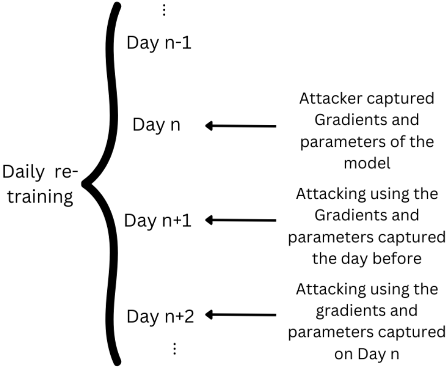

⋮
Day n-1
Day n
Attacker captured Gradients and parameters of the model
Daily re-training
Day n+1
Attacking using the Gradients and parameters captured the day before
Day n+2
Attacking using the gradients and parameters captured on Day n
⋮

Figure: Attacking Scenario with Continuous Training: the impact of adversarial attacks before (attacking in Day n) and after re-training (attacking in Day n+1 and Day n+2) [3]

Gap between reality and research. The practicality question?

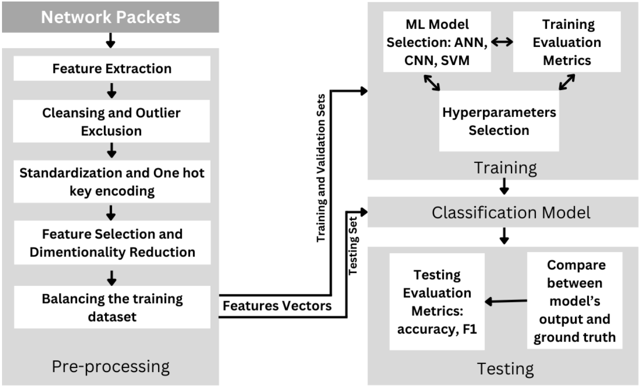
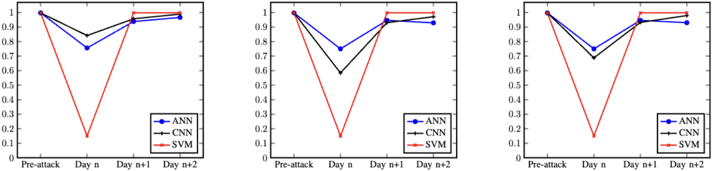

Network Packets
Feature Extraction
Cleansing and Outlier Exclusion
Standardization and One hot key encoding
Feature Selection and Dimentionality Reduction
Balancing the training dataset
Pre-processing
Features Vectors
Training and Validation Sets
Testing Set
ML Model Selection: ANN, CNN, SVM
Training Evaluation Metrics
Hyperparameters Selection
Training
Classification Model
Testing Evaluation Metrics: accuracy, F1
Compare between model's output and ground truth
Testing

Figure: Target ML-based NIDS [3]

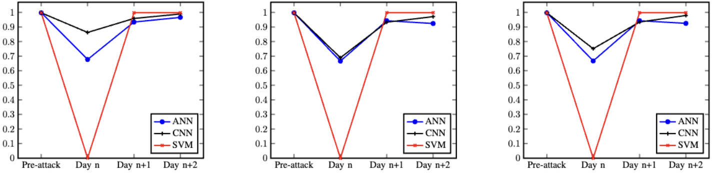
Gap between reality and research. The practicality question?



(a) FGSM (b) PGD (c) BIM

Figure: Accuracy (Y-axis) of the NIDSs before and after the attacks, where Day n represents attacking before re-training, Day n+1 represents attacking one day after re-training, and Day n+2 represents attacking two days after re-training. [3]

Gap between reality and research. The practicality question?



(a) FGSM

(b) PGD

(c) BIM

Figure: F1-measure (Y-axis) of the NIDSs before and after the attacks, where Day n represents attacking before re-training, Day n+1 represents attacking one day after re-training, and Day n+2 represents attacking two days after re-training. [3]

Gap between reality and research. The practicality question?

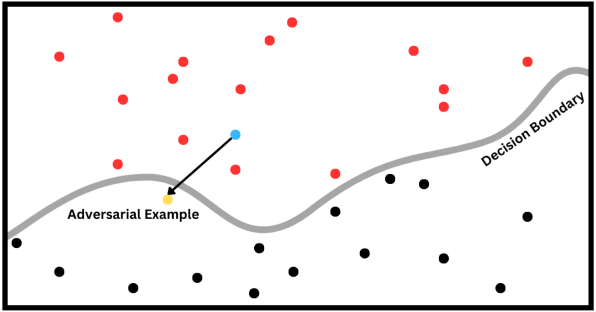


Figure: Adversarial Attacks Visualization [3]

Gap between reality and research. The practicality question?

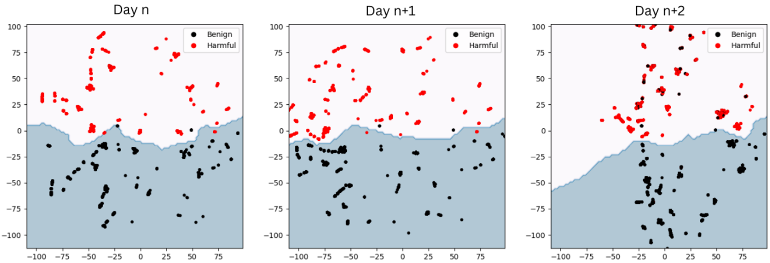


Figure: Decision Boundary Evolution using t-SNE [3]

Gap between reality and research. The practicality question?

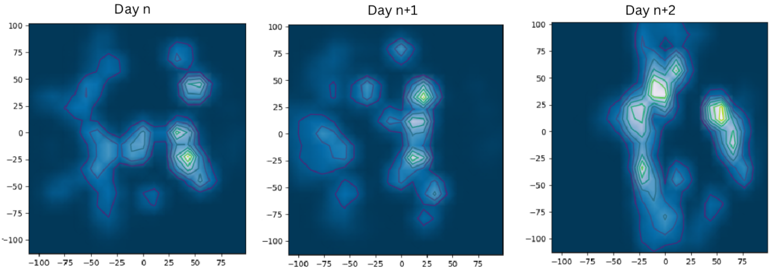


Figure: Data Distribution Evolution using t-SNE, wher color intensity encodes data density, with lighter areas representing higher concentrations of points. [3]

Conclusion

- Our work highlights several factors that could make numerous researched adversarial attacks impractical against real-world ML-based systems in network security.
- We do not claim that adversarial attacks won't harm ML-based NIDSs; rather, we find that the gap between research and real-world practicality is wide and deserves to be addressed.
- Continuous re-training, even without adversarial training, may limit the effect of such attacks.

References I

[1] O. Ibitoye, R. Abou-Khamis, M. e. Shehaby, A. Matrawy, and M. O. Shafiq, “The threat of adversarial attacks on machine learning in network security–a survey,” *arXiv preprint arXiv:1911.02621*, 2023.

[2] M. elShehaby, A. Kotha, and A. Matrawy, “Introducing adaptive continuous adversarial training (acat) to enhance ml robustness,” *arXiv preprint arXiv:2403.10461*, 2024.

[3] M. e. Shehaby and A. Matrawy, “Adversarial evasion attacks practicality in networks: Testing the impact of dynamic learning,” *arXiv preprint arXiv:2306.05494*, 2023.

[4] O. Ibitoye, M. O. Shafiq, and A. Matrawy, “Differentially private self-normalizing neural networks for adversarial robustness in federated learning,” *Computers & Security*, vol. 116, p. 102631, 2022.

References II

[5] M. AbuIssa, M. Ibnkahla, A. Matrawy, and A. Eldosouky, "Temporal partitioned federated learning for iot intrusion detection systems," in *2024 Wireless Communications and Networking (WCNC)*. IEEE, 2024.

[6] R. Abou Khamis and A. Matrawy, "Could min-max optimization be a general defense against adversarial attacks?" in *International Conference on Computing, Networking, and Communications (ICNC)*, 2024.

[7] R. Maarouf, D. Sattar, and A. Matrawy, "Evaluating resilience of encrypted traffic classification against adversarial evasion attacks," in *2021 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 2021, pp. 1–6.

Carleton University

References III

[8] R. Abou Khamis and 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)*. IEEE, 2020, pp. 1–6.

[9] R. Abou Khamis, M. O. Shafiq, and A. Matrawy, "Investigating resistance of deep learning-based ids against adversaries using min-max optimization," in *ICC 2020-2020 IEEE international conference on communications (ICC)*. IEEE, 2020, pp. 1–7.

[10] O. Ibitoye, O. Shafiq, and A. Matrawy, "Analyzing adversarial attacks against deep learning for intrusion detection in iot networks," in *2019 IEEE global communications conference (GLOBECOM)*. IEEE, 2019, pp. 1–6.

Questions?

carleton.ca/ngn