

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

**Ashraf Matrawy**  
Carleton University  
[carleton.ca/ngn](http://carleton.ca/ngn)

A presentation at the  
Interdisciplinary Research Center for Intelligent Secure Systems  
KFUPM, Dhahran, KSA  
November 19, 2024

# Acknowledgement

I would like to acknowledge the contributions of my students, Aboukhamis, Elshehaby, Ibitoye, and Kotha, to the research presented in these slides, as well as my undergraduate students Kadri, Lott, Mailloux, Morozov, and Virr.

Some of the work in this presentation was funded by Natural Sciences and Engineering Research Council of Canada (NSERC).

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](https://carleton.ca/ngn)

# Mentoring

- Graduated 7 PhD students and a number of master's students.
- Three of my past students are professors :)
- **2019** Faculty Graduate Mentoring Award, Carleton University, Nominated by my former graduate students. It was offered to seven faculty members from across all disciplines.

## Relevant experience

- Work on problems with practical applications.
- Extensive experience in program building, curriculum development, and academic administration.
- **2021 IEEE Ottawa Section Outstanding Engineering Educator Award** for recognition of **outstanding contributions to engineering research and education, and more specifically in the field of computer and network security.**
- **2022 Carleton University Research Achievement Award.**
- **2021 Carleton Faculty of Engineering and Design's Research Award,**
- Multiple Best Paper and Best Poster awards at IEEE and ACM conferences.
- Industrial experiences during sabbaticals
- Consulting
- Successful funding experience

## Outline

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

## Core work

While our group produced significant work in this area, this presentation mostly covers our latest, in-progress work [1], [2], [3], [4]

## Sample of our published work in this area

- Differentially Private Self-normalizing Neural Networks for Adversarial Robustness in Federated Learning, 2022 [5].
- Temporal Partitioned Federated Learning for IoT Intrusion Detection Systems, 2024 [6].
- Could Min-Max Optimization Be A General Defense Against Adversarial Attacks?, 2024 [7].
- Evaluating Resilience of Encrypted Traffic Classification against Adversarial Evasion Attacks, 2021 [8]
- Evaluation of Adversarial Training on Different Types of Neural Networks in Deep Learning-based IDSs, 2020 [9].
- Investigating Resistance of Deep Learning-based IDS against Adversaries using min-max Optimization, 2020 [10].
- Analyzing Adversarial Attacks Against Deep Learning for Intrusion Detection in IoT Networks, 2019 [11].

# 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
  - The work in this presentation addresses evasion adversarial attacks.
- Feature vs Problem space [1]
- Based on knowledge

# Adversarial attacks - Evasion



Figure: Evasion Adversarial Attack [1]

# Our classification of adversarial attacks in network security

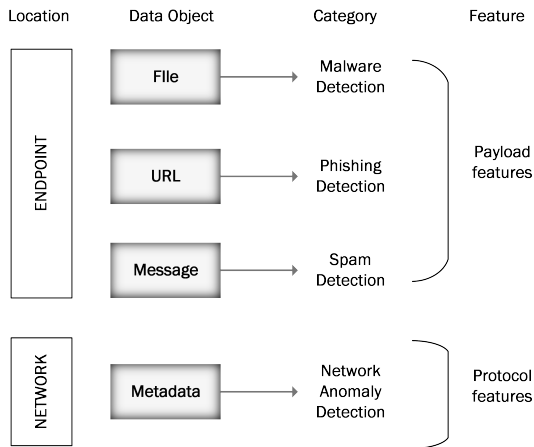


Figure: Adversarial attack classification [1]

# Our work on characterising adversarial attacks

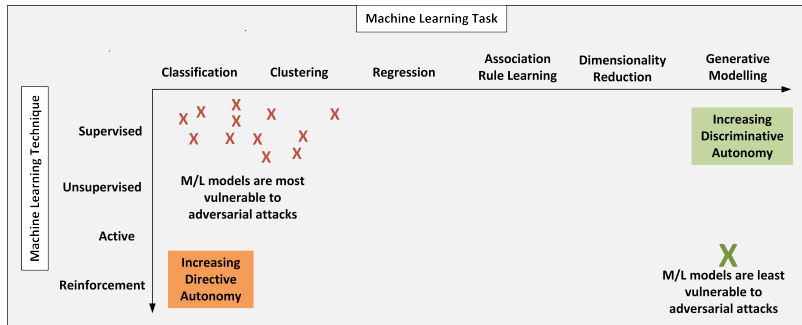


Figure: Adversarial Risk Grid Map [1]

# Defenses against Adversarial attacks in network security

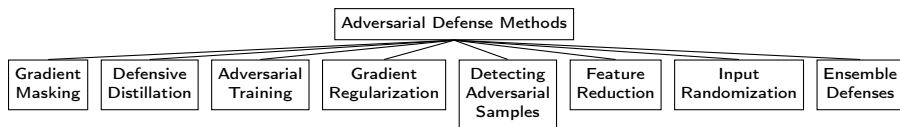


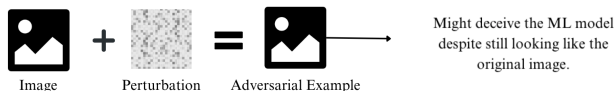
Figure: Adversarial Defense Methods [1]

# Adversarial Training

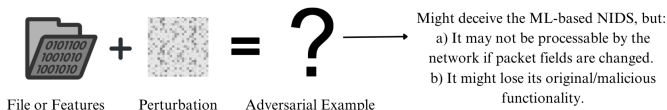
- Defence using min-max [10]
- Checking the impact on different models [9]
- General defence? [7]

# Adversarial attacks in networks: Are they different?

Work with ElShehaby [3].



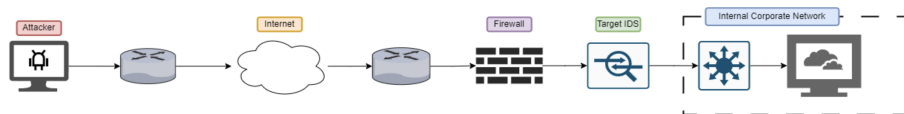
(a) Adversarial Example Generation in the Computer Vision Domain



(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 - from our paper in IEEE WF-IoT [4]



# Gap between reality and research. The practicality question?

## Original

⋮	
IP address	172.31.66.5
destination port	443
⋮	

## Manipulated

⋮	
IP address	172.31.66.0
destination port	442
⋮	

**Figure:** An example of Original and Perturbed/Manipulated IDS Features - from our paper in IEEE WF-IoT [4]

# Gap between reality and research.

## The practicality question? [4]

Attacks on ML-based NIDS must adhere to:

- Valid IP addresses required (e.g., can't use 333.333.333.333)
- Port numbers must be within valid range (0-65535)
- Protocol-specific constraints (e.g., TCP handshake)
- Multiple security layers (e.g., routers, firewalls, IDS)

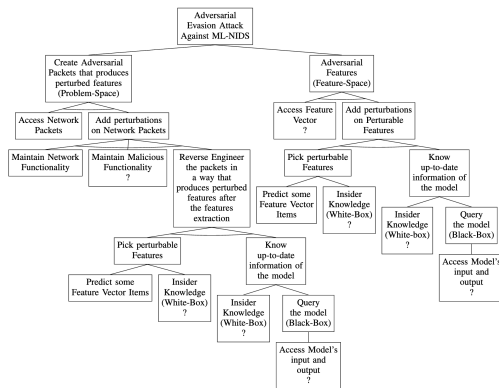
# Gap between reality and research. The practicality question? [4]

Our testing resulted in the following generated adversarial attack that are Impractical in networking context:

- IP Address:
  - Original: 172.31.66.5 (valid host address)
  - Perturbed: 172.31.66.0 (invalid - network address)
- Port:
  - Original: 443 (HTTPS, likely allowed)
  - Perturbed: 442 (uncommon, likely blocked)
- Protocol flags:
  - Invalid combinations (e.g., TCP flags in UDP packet)
  - Incorrect flag counts for TCP handshake

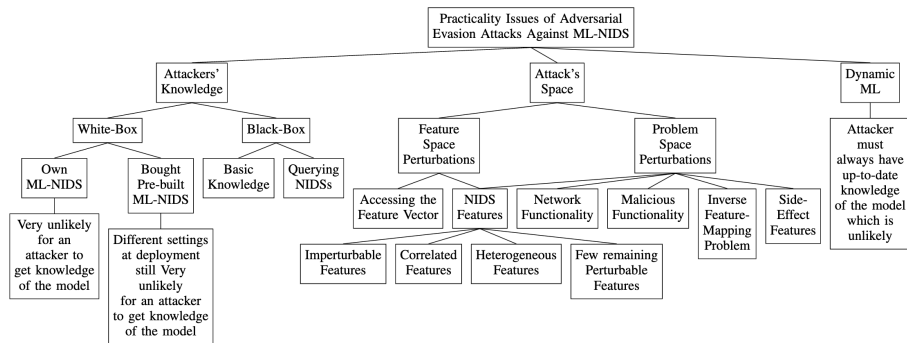
While this testing provides valuable insights, it is not comprehensive, and numerous other scenarios and edge cases should be examined to fully understand the potential impacts and effectiveness in network environments

# The practicality question? - Threat Modeling



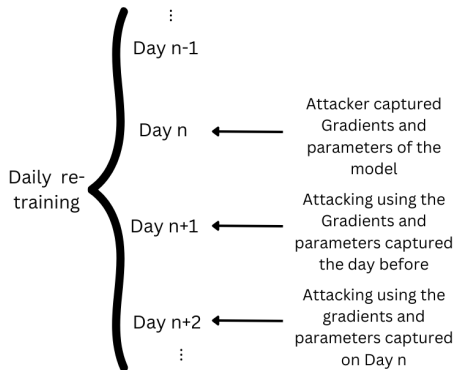
**Figure:** Attack Tree of Adversarial Evasion Attack Against ML-NIDS.  $\angle$  indicates a disjunction (OR),  $\triangleleft$  indicates a conjunction (AND), and ? denotes a leaf node with uncertain feasibility (questionable practicality) [3]

# The practicality question? - Taxonomy



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

# The practicality question? - Dynamic Learning Test



**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]

# The practicality question? - Dynamic Learning Test

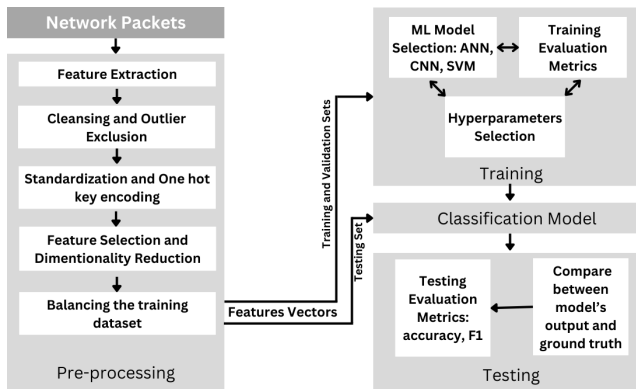
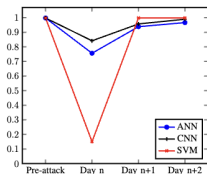
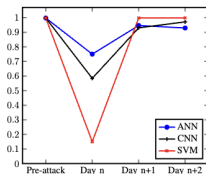


Figure: Target ML-based NIDS [3]

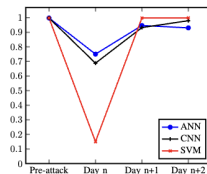
# The practicality question? - Dynamic Learning Test



(a) FGSM



(b) PGD

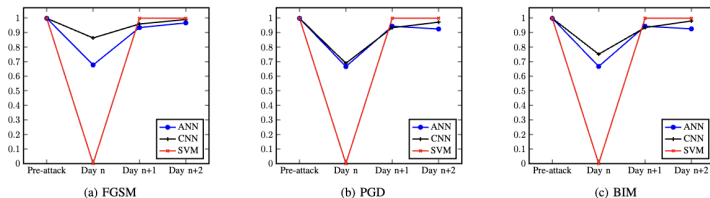


(c) BIM

**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]



# The practicality question? - Dynamic Learning Test



**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]

# The practicality question? - Dynamic Learning Test

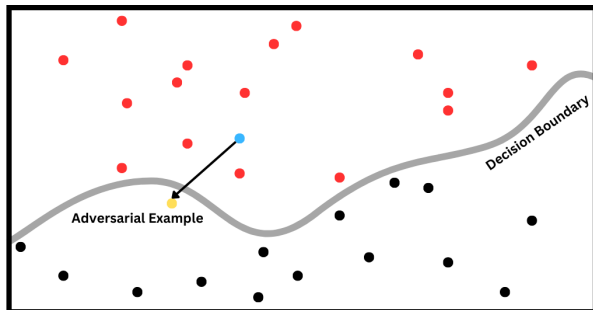


Figure: Adversarial Attacks Visualization [3]

# The practicality question? - Dynamic Learning Test

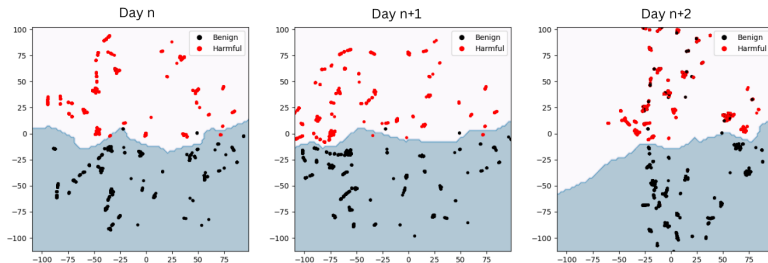
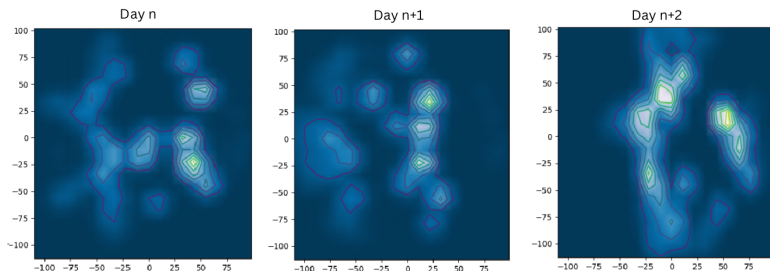


Figure: Decision Boundary Evolution using t-SNE [3]

# The practicality question? - Dynamic Learning Test



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

# Presenting a Solution to Adversarial Training

- Let's assume an adversary managed to find a practical attack, what should we do?
- One of the defences was adversarial training
- We tried other methods in the past
- **Where do you get adversarial samples to train the models?**
- **How often do you train?**

We present results in problem-space (SPAM) work, we also have new results in feature-space (NIDS) that are not presented in these slides.

# 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
- In order to evaluate ACAT, we used domain of SPAM filtering which required the following contributions that are specific to the experimental evaluation:
  - Adapting the adversarial detection approach by Ye et al. for the text-based SPAM problem.
  - Training the adversarial detector using a balanced dataset with an almost equal distribution of normal and adversarial samples, between ham and SPAM samples.

# Introducing ACAT

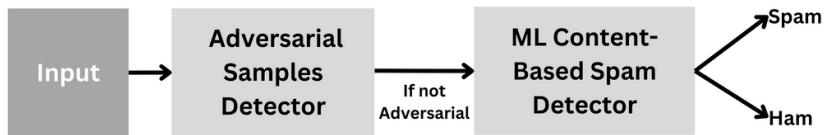


Figure: Conventional Detecting Approach [2]

# Introducing ACAT

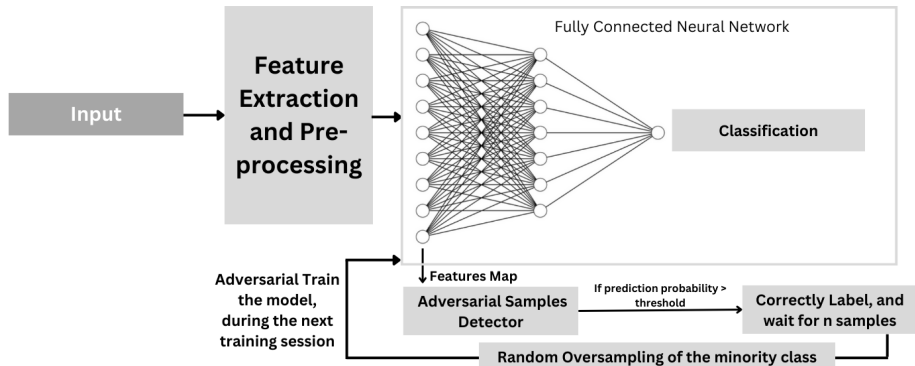


Figure: The Proposed Adaptive Continuous Adversarial Training (ACAT) [2]



# Introducing ACAT

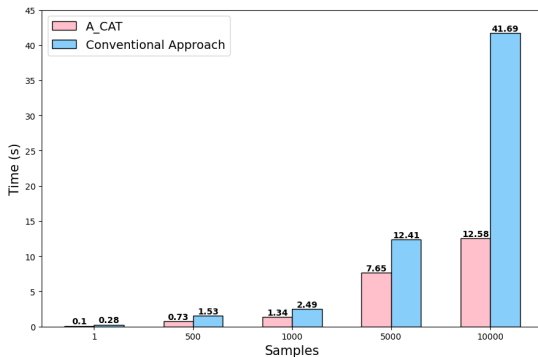
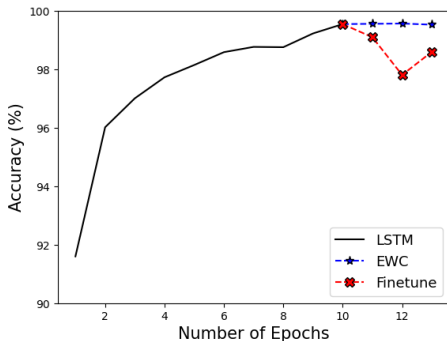


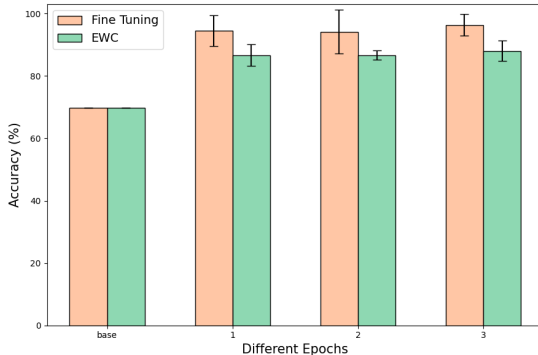
Figure: Prediction time of ACAT vs a Conventional Approach [2]

# 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 ( $D_{\text{test}}$ ) of perturbed Enron SPAM (Fine Tuning vs EWC) [2]



# 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,” *posted on arXiv preprint arXiv:2403.10461*, This work was later published in *IEEE Networking Letters*, 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.

## References II

- [4] J. Kadri, A. Lott, M. Brendan, K. Morozov, S. Virr, M. Elshehaby, and A. Matrawy, "Work in progress: Evasion adversarial attacks perturbations in network security," in *2024 IEEE 10th World Forum on Internet of Thing*. IEEE, 2024.
- [5] 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.
- [6] M. Abulssa, 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.

# References III

- [7] 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.
- [8] 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.
- [9] 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.

## References IV

- [10] 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.
- [11] 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](http://carleton.ca/ngn)