INF6422E Advanced Concepts in Computer Security

Practical Work 2 – Winter 2026

Adversarial Machine Learning and Robustness Evaluation

Instructions:

- This work is to be submitted in a group of two or three and via Moodle only.
- The submitted report must be in pdf form and also include (.py, .ipynb) with it. You're free to build it in any format you want, however (.docx, .odt, .tex, etc.).
- The report must contain a title page including the course title, the lab title, your names and student ID numbers (Matricule).
- The report must be submitted by the 18th of February 2026 before 23h59. A penalty of 10% will be applied for each day after that date.
- There are 4 labs in total, each of 20 marks/10 weightage.

Objective:

This lab explores adversarial machine learning, where attackers manipulate input data in order to fool machine learning–based security systems. Students will implement adversarial attacks and defenses, and quantitatively evaluate the robustness of deep learning models.

Adversarial threats are highly relevant in cybersecurity, as modern intrusion detection systems, malware classifiers, and biometric authentication models increasingly rely on machine learning. Understanding robustness is therefore essential for deploying trustworthy AI-powered security solutions.

Key Learning Goals:

1. Train and evaluate a baseline learning classifier.
2. Implement adversarial evasion attacks (FGSM, PGD).
3. Measure robustness degradation under attack.
4. Explore data poisoning as a training-time cybersecurity threat.
5. Apply defense mechanisms and analyze trade-offs between robustness and accuracy.

Dataset

CIFAR-10: The CIFAR-10 dataset consists of 60000 32x32 colour images in 10 classes, with 6000 images per class.
Link: https://www.cs.toronto.edu/~kriz/cifar.html

1. Baseline Machine Learning Classification Model [6 Points]

1.1 Dataset Preparation and Training Pipeline

Use CIFAR-10 and preprocess the dataset appropriately.
Split the dataset into:

- Training (70%)
- Validation (15%)
- Testing (15%)

Deliverable:

Justify the dataset split and explain why validation is necessary for secure model evaluation.

1.2 CNN Model Training

Implement a Convolutional Neural Network (CNN) using PyTorch.
The model must include at least:

- Convolution layers
- Pooling
- Fully connected classifier

- Evaluate performance using:
  - Accuracy, Precision, Recall, F1-Score

Deliverable:

Provide:

- Model architecture summary
- Training curves (loss/accuracy)
- Test performance metrics
- Confusion matrix interpretation

2. Adversarial Evasion Attacks (Test-Time Threats) [6 Points]

2.1 FGSM Attack Implementation

Implement the **Fast Gradient Sign Method (FGSM)** attack.
Generate adversarial examples using at least three epsilon values.

Deliverable:

- Visualization of adversarial images
- Accuracy drops under increasing attack strength
- Discussion: Why are evasion attacks dangerous for AI-based security systems?

2.1 PGD Attack Implementation

Implement **Projected Gradient Descent (PGD)** as a stronger iterative evasion attack.

Deliverable:

- Compare FGSM vs PGD effectiveness
- Report Robust Accuracy under PGD
- Explain trade-offs between attack strength and detectability

3. Data Poisoning Attacks (Training-Time Threats) [5 Points]

3.1 Label-Flipping Poisoning Experiment

Simulate a poisoning attack by flipping the labels of:

- 5% of the training set
- 15% of the training set

Train the model on poisoned data.

Deliverable:

- Performance comparison across poisoning levels
- Discussion: How can poisoning compromise IDS or malware classifiers?

3.2 Quantitative Poisoning Impact

Compute and report:

- Clean Accuracy
- Accuracy after poisoning

Deliverable:

- Explain how poisoning affects trust in security-critical ML systems.

4. Defenses and Robustness Trade-Offs [3 Points]

4.1 Adversarial Training Defense

Retrain the model using adversarial examples (FGSM or PGD-based training).

Deliverable:

- Robust Accuracy improvement after defense
- Comparison table: Before vs After adversarial training

References:

[1] Alotaibi, A., & Rassam, M. A. (2023). Adversarial Machine Learning Attacks against Intrusion Detection Systems: A Survey on Strategies and Defense. Future Internet, 15(2), 62. https://doi.org/10.3390/fi15020062

[2] Vassilev, A. , Oprea, A. , Fordyce, A. and Andersen, H. (2024), Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations, NIST Trustworthy and Responsible AI, National Institute of Standards and Technology, Gaithersburg, MD, [online], https://doi.org/10.6028/NIST.AI.100-2e2023, https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=957080 (Accessed February 1, 2026)

[3] Zhao, P., Zhu, W., Jiao, P., Gao, D., & Wu, O. (2025). Data poisoning in deep learning: A survey. *arXiv preprint arXiv:2503.22759*.

[4] Kure, H. I., Sarkar, P., Ndanusa, A. B., & Nwajana, A. O. (2025). Detecting and preventing data poisoning attacks on AI models. *arXiv preprint arXiv:2503.09302*.

[5] Paracha, A., Arshad, J., Farah, M.B. *et al.* Deep behavioral analysis of machine learning algorithms against data poisoning. *Int. J. Inf. Secur.* **24**, 29 (2025). https://doi.org/10.1007/s10207-024-00940-x