Unit 6: Adversarial Machine Learning and Security Challenges

Adarsh KUMAR

Universitat Politècnica de Catalunya
Department of Computer Science

Project Coordinator:
Prof. Ilker Demirkol

MERiT Project
September 3, 2025

Co-funded by
the European Union

Outline

MERiT

DIGITAL EUROPE PROGRAMME

Co-funded by the European Union

Adversarial Machine Learning: Overview

- Studies vulnerabilities of ML models to intentionally manipulated inputs.
- Adversarial examples are inputs modified to mislead ML models.
- Modifications often imperceptible but cause incorrect predictions.
- Key risk in cybersecurity ML systems.
- Attackers exploit model weaknesses to evade detection.
- Requires robust defenses and model design.

MERiT

Co-funded by the European Union

Types of Adversarial Attacks: Knowledge Perspective

- White-box: attacker has full access to model architecture, weights, gradients.
- Black-box: no internal model knowledge; attacker queries model for outputs.
- Gray-box: partial knowledge (training data but not parameters).
- Different knowledge levels dictate attack strategies.
- White-box allows more precise attacks.
- Black-box often uses query-based approximations.

MERiT

Co-funded by the European Union

Types of Adversarial Attacks: Objectives

- Evasion Attacks: modify inputs at inference to avoid detection.
- Poisoning Attacks: inject malicious data into training to corrupt model.
- Model Inversion/Extraction: infer or replicate confidential data or model.
- Evasion most practical in malware and intrusion detection.
- Poisoning threatens model integrity during retraining.
- Extraction risks intellectual property and privacy.

MERiT

Co-funded by the European Union

Techniques for Crafting Adversarial Examples

- FGSM (Fast Gradient Sign Method): uses gradient of loss to perturb input.
- PGD (Projected Gradient Descent): stronger iterative method.
- C&W and DeepFool: optimization-based, minimize perturbation magnitude while misclassifying.
- Increasing sophistication improves attack success.
- Techniques exploit model differentiability.
- Defense requires countering these attack methods.

MERiT

Co-funded by the European Union

Implications for Cybersecurity

- Malware classifiers vulnerable to modified binaries.
- Network intrusion detection may miss subtly altered traffic.
- Biometric systems can be fooled by adversarial inputs.
- Face recognition systems susceptible to adversarial images.
- Threatens reliability of AI-based security tools.
- Calls for adversarially robust ML designs.

MERiT

Co-funded by the European Union

Anomaly Detection and Statistical Outlier Analysis

- Crucial for identifying novel or stealthy attacks.
- Malicious behavior deviates statistically from normal patterns.
- Detects attacks not in labeled training data.
- Provides dynamic, adaptable defense.
- Helps detect evasion or poisoning attempts.
- Improves alert quality in cybersecurity systems.

MERiT

Co-funded by the European Union

Types of Anomaly Detection

- Point anomalies: rare individual data points (e.g., large data transfer).
- Contextual anomalies: abnormal only in specific contexts (e.g., admin login time).
- Collective anomalies: group of points showing collective abnormality (e.g., slow port scan).
- Detects diverse threat behaviors.
- Supports multi-dimensional analysis.
- Tailored detection strategies improve efficacy.

MERiT

Co-funded by the European Union

Statistical Techniques for Anomaly Detection

- Z-score Analysis: flags data beyond thresholds of standard deviation.
- Robust statistics (Median Absolute Deviation) resist skewed data effects.
- Density-based methods (Local Outlier Factor) identify sparse data points.
- Suitable for high-dimensional cybersecurity data spaces.
- Enhances resilience against noise and attacks.
- Enables effective anomaly classification.

MERiT

Co-funded by the European Union

Role in Adversarial Machine Learning

- Detect adversarial examples deviating statistically from natural data.
- Monitor unusual access or query patterns, signaling extraction attacks.
- Provide an additional defense layer in ML security.
- Compliments model-based robustness methods.
- Improves detection of subtle and novel attacks.
- Supports ongoing security situational awareness.

MERiT

Co-funded by the European Union

Statistical Divergence Measures in Adversarial Detection

- Quantify differences between probability distributions (normal vs anomalous).
- Detect distributional shifts caused by adversarial activity or data poisoning.
- Evaluate generative model and anomaly detector effectiveness.
- Supports monitoring of user behavior deviations.
- Essential for robust cybersecurity monitoring.
- Improves real-time threat detection fidelity.

MERiT

Co-funded by the European Union

Common Divergence Measures

- Kullback–Leibler (KL) Divergence: asymmetric measure, sensitive to tails.
- Jensen–Shannon (JS) Divergence: symmetric, smoother, bounded version.
- Total Variation Distance and Wasserstein Distance: additional practical metrics.
- Choice depends on detection task and data properties.
- Used to quantify anomalous distributional changes.
- Enhances adversarial robustness in ML.

MERiT

Co-funded by the European Union

Applications in Cybersecurity

- Measure shifts in user behavior indicating insider threats.
- Compare network traffic profiles to expected normal behavior.
- Assess robustness and reliability of threat classifiers.
- Detect suspicious changes signaling attacks.
- Drives adaptive defense updates.
- Improves system resilience against adversarial manipulation.

MERIT

Co-funded by the European Union

Conclusion

- Adversarial ML exposes critical security vulnerabilities.
- Understanding and detecting attacks—evasion, poisoning, model extraction—is essential.
- Statistical anomaly detection and divergence measures provide robust defense layers.
- Continuous evaluation and strengthening of ML models improves cybersecurity resilience.
- Research and defense must evolve alongside adversary sophistication.
- Building secure ML systems is key to future threat mitigation.

MERiT

Co-funded by the European Union