DOCTORAL (PHD) STUDIES

COURSE UNIT DESCRIPTION

Course unit title	Scientific areas	Faculty	Institute, department
Adversarial Machine Learning	Informatics engineering (T 007)	Faculty of Mathematics and Informatics	Institute of Data Science and Digital technologies

Study method	Number of credits	Study method	Number of credits
Lectures	0	Consultations	1
Individual works	6	Seminars	0

Summary

Course objective. The aim of the course is to enable the doctoral student to acquire systematic knowledge of adversarial machine learning, the security and privacy risks of machine learning and artificial intelligence (AI) systems. The course covers the main types of attacks against machine learning models, methods for evaluating model robustness, and approaches for improving the reliability, security, privacy, and explainability of AI-based systems. The course also aims to develop the doctoral student's ability to critically analyze scientific literature, formulate a research problem, define a threat model, evaluate attacks and defense methods, and apply adversarial machine learning concepts in different domains of informatics engineering. The course is based on guided self-study of scientific literature, individual consultations, and an individual research project related to the dissertation topic.

Prerequisites. The doctoral student is expected to have basic knowledge of machine learning, deep learning, and Python programming.

Introductory part. Definition and historical evolution of adversarial machine learning, and its place among security, privacy, and trustworthy AI. Vulnerabilities of machine learning models and data-driven decision-making systems. The role of adversarial machine learning in modern AI security, cybersecurity, and trustworthy AI research.

Adversarial machine learning:

- Basic concepts of adversarial machine learning, AI security.
- Threat models: attacker goals, available knowledge, capabilities, and constraints.
- White-box, black-box, and gray-box attack scenarios, depending on how much information the attacker has about the model.
- Targeted attacks, where the attacker aims to force a specific wrong decision, and untargeted attacks, where any wrong decision is sufficient.
- Attacks at different stages of the machine learning process: data collection, model training, validation, deployment, inference, monitoring, and model updating.
- Evasion attacks, where input data are modified at prediction time in order to mislead the model.
- Adversarial examples and their generation for different types of models and data.
- Poisoning attacks, where training data are manipulated in order to influence the trained model.
- Backdoor attacks against machine learning and deep learning models, where hidden model behavior is activated only under specific conditions.
- Privacy-related attacks, including model extraction, model inversion, and membership inference.
- Attacks and defenses in different data modalities: images, text, tabular data, graphs, time series, software, and cybersecurity data.
- Defense methods: adversarial training, robust model optimization, model ensembles, input preprocessing, anomaly detection, and uncertainty estimation.

- Evaluation of model robustness: attack success rate, robust accuracy, perturbation constraints, query budget, false positive and false negative rates, limits on input modification, and number of model queries.
- Reproducibility and reliable experimental design in adversarial machine learning research.

Cybersecurity applications:

- Application of machine learning in malware detection, intrusion detection, user authentication, behavioral biometrics, and anomaly detection.
- Specific challenges of cybersecurity data: class imbalance, noise, high dimensionality, concept drift, adversarial behavior, and changing attack patterns (concept drift).
- Static and dynamic malware analysis from the machine learning perspective.
- Feature manipulation, adversarial obfuscation, benign feature injection, and evasion of malware classifiers.
- Secure deployment and monitoring of machine learning models in cybersecurity systems.
- Responsible use of adversarial machine learning techniques in cybersecurity research.

Explainable and trustworthy AI:

- Explainable AI methods for model analysis, error analysis, and decision support.
- Local explanations, which explain individual predictions, and global explanations, which describe general model behavior.
- Application of explainable AI in cybersecurity, malware detection, authentication, and anomaly detection.
- Interpretability methods: LIME, SHAP, and saliency maps. Attacks against explanations.
- Possible manipulation of explanations and risks of misleading interpretability.
- Attacks against explanations and risks of misleading interpretability.

Practical task: to prepare an individual research work related to the doctoral student's dissertation topic. The work should include a scientific literature review, formulation of a threat model or research question, analysis of relevant adversarial machine learning methods, and an experimental, comparative, or conceptual study. The final result may be prepared as a research report.
Assessment: an individual project with a presentation, and an oral examination.

Main literature
Joseph, A. D., Nelson, B., Rubinstein, B. I. P., & Tygar, J. D. (2019). Adversarial Machine Learning. Cambridge University Press.
Vorobeychik, Y., & Kantarcioglu, M. (2018). Adversarial Machine Learning. Morgan & Claypool.
Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. http://www.deeplearningbook.org
Molnar, C. (2025). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (3nd ed.). https://christophm.github.io/interpretable-ml-book/
Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 84, 317–331.
Baniecki, H., & Biecek, P. (2024). Adversarial attacks and defenses in explainable artificial intelligence: A survey. Information Fusion, 107.
MITRE. ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems. https://atlas.mitre.org/
Arrieta, A. B., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Lecturer(s) (name, surname)	Science degree	Main publications
Viktor Medvedev	dr.	http://www.elaba.mb.vu.lt/dmsti/?aut=Viktor+Medvedev
Olga Kurasova	dr.	http://www.elaba.mb.vu.lt/dmsti/?aut=Olga+Kurasova
Rokas Gipiškis	dr.	http://www.elaba.mb.vu.lt/dmsti/?aut=Rokas+Gipiškis