CS 487/587 – Adversarial Machine Learning

Semester: Spring 2025 (January 8 – May 9, 2025)

Credits Hours: 3

Instructor: Alex Vakanski, vakanski@uidaho.edu

Office Location: TAB 311, Idaho Falls Center

Office Hours: Friday 12 p.m. – 1 p.m. PT (Zoom link on Canvas)

Course Delivery Methods:

- Virtual meetings (live meetings, students participate through Zoom)
- Classroom (live meetings, video link from Idaho Falls)
- Online (recorded Zoom videos of lectures available to students to watch after the classes)

Course Description

The course introduces students to adversarial attacks on machine learning models and defenses against the attacks. The particular focus is on adversarial attacks and adversarial examples in deep learning models, due to their prevalence in modern machine learning applications. Covered topics include evasion attacks targeting white-box and black-box machine learning models, data poisoning attacks, privacy attacks, jailbreak attacks on large language models, defense strategies against adversarial attacks, and robust machine learning models. The course also provides an overview of adversarial attacks against machine learning models used in cybersecurity applications, including malware detection and classification, network intrusion detection, spam filtering, URL detection, cyber-physical systems, and biometric systems.

Learning Outcomes

The objective is that upon the completion of the course the students should demonstrate the ability to:

1. Identify the vulnerabilities of machine learning models to various types of adversarial attacks.
2. Differentiate between adversarial evasion attacks in white-box and black-box settings and understand the principles of data poisoning attacks.
3. Explain the fundamentals of adversarial privacy attacks and outline privacy-preserving defense methods.
4. Describe jailbreak attacks on large language models and propose corresponding mitigation methods.
5. List common defense strategies against adversarial attacks and discuss approaches for improved robustness of machine learning models.
6. Identify the unique characteristics of adversarial attacks on machine learning models in the cybersecurity domain.
7. Implement adversarial attacks and defenses against conventional machine learning models and deep learning models.
8. Evaluate the effectiveness of adversarial attacks against anomaly detection systems for network intrusion detection, machine learning malware classifiers, and anti-spam filtering models.
9. Analyze the ethical and societal implications of adversarial attacks and defenses.

Prerequisites

CS 212 Practical Python, or CS 477 Python for Machine Learning, or Instructor Permission

Students are expected to have a basic understanding of linear algebra, probability and statistics, and machine learning concepts. Familiarity with neural networks and deep learning is recommended, but not required. A working knowledge of Python programming is required to complete the course assignments and the course project. Additionally, it is preferred that the students have a basic level of familiarity with at least one of the following machine learning libraries: TensorFlow, Keras, or PyTorch.

Grading
Five homework assignments (worth together 40 marks), three quizzes (30 marks), course project (10 marks), presentation (10 marks), and class attendance and participation (10 marks). The assignments and the project will involve implementation of the studied adversarial attacks and defense methods on medium-size datasets (e.g., a few thousand images or similar). To complete the assignments and the project, the students can either use the GPUs provided by Google Colab, or they can use other programming environments that provide access to GPUs.

Assessment Component	*Marks*
Assignments (x5)	40
Quizzes (x3)	30
Course Project	10
Presentation	10
Attendance and Participation	10
Total	*100*

Textbook
There is no required textbook. The reading materials for each week are listed in the Course Outline section.

Course Outline (Tentative)

Date	**Topics, Readings, Assignments**
Week 1	
Thursday, January 9	**Lecture 1: Introduction to Adversarial Machine Learning**
Week 2	
Tuesday, January 14	**Lecture 2: Deep Learning Overview**
Thursday, January 16	**Lecture 2 (cont'd): Deep Learning Overview**
Week 3	
Tuesday, January 21	**Lecture 3: Mathematics for Machine Learning** *Reading:* 1. Goodfellow (2014) Explaining and Harnessing Adversarial Examples (pdf)
Thursday, January 23	**Lecture 3 (cont'd): Mathematics for Machine Learning** **Due:** Assignment 1
Week 4	

Tuesday, January 28	**Lecture 4: Evasion Attacks Against White-box Models** 1. Carlini (2017) Towards Evaluating the Robustness of Neural Networks (pdf) 2. Papernot (2016) The Limitations of Deep Learning in Adversarial Settings (pdf) 3. Xiao (2018) Spatially Transformed Adversarial Examples (pdf) 4. Su (2019) One Pixel Attack for Fooling Deep Neural Networks (pdf)
Thursday, January 30	**Lecture 4 (cont'd): Evasion Attacks Against White-box Models** **Quiz 1**
Week 5	
Tuesday, February 4	**Lecture 5: Evasion Attacks Against Black-box Models** 1. Brendel (2017) Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models (pdf) 2. Papernot (2016) Transferability in Machine Learning: From Phenomena to Black-Box Attacks using Adversarial Samples (pdf) 3. Chen (2019) HopSkipJumpAttack: A Query-efficient Decision-based Adversarial Attack (pdf) 4. Guo (2019) Simple Black-box Adversarial Attacks (pdf)
Thursday, February 6	**Lecture 5 (cont'd): Evasion Attacks Against Black-box Models**
Week 6	
Tuesday, February 11	**Lecture 6: Adversarial Attacks Against Large Language Models** 1. Zou (2023) Universal and Transferable Adversarial Attacks on Aligned Language Models (pdf) 2. Paulas (2024) AdvPrompter: Fast Adaptive Adversarial Prompting for LLMs (pdf) 3. Wei (2023) Jailbroken: How Does LLM Safety Training Fail? (pdf) 4. Grehsake (2023) Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection (pdf) 5. Wan (2023) Poisoning Language Models During Instruction Tuning (pdf)
Thursday, February 13	**Lecture 6 (cont'd): Adversarial Attacks Against Large Language Models** **Due:** Assignment 2
Week 7	
Tuesday, February 18	**Guest Lecture by Dr. Arman Zharmagambetov** Paulas, Zharmagambetov (2024) AdvPrompter: Fast Adaptive Adversarial Prompting for LLMs (pdf) **Lecture 6 (cont'd): Adversarial Attacks Against Large Language Models**
Thursday, February 20	**Lecture 7: Defenses Against Evasion Attacks** 1. Tramer (2018) Ensemble Adversarial Training: Attacks and Defenses (pdf) 2. Papernot (2016) Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks (pdf)

	3. Xu (2017) Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks (pdf) Madry (2017) Towards Deep Learning Models Resistant to Adversarial Attacks (pdf)
Week 8	
Tuesday, February 25	**Lecture 7: Defenses Against Evasion Attacks** 1. Carlini (2022) (Certified!!) Adversarial Robustness for Free! (pdf) 2. Zhang (2019) Theoretically Principled Trade-off between Robustness and Accuracy (pdf 3. Carmon (2019) Unlabeled Data Improves Adversarial Robustness (pdf) 4. Cohen (2019) Certified Adversarial Robustness via Randomized Smoothing (pdf)
Thursday, February 27	**Lecture 7 (cont'd): Defenses Against Evasion Attacks** **Due:** Assignment 3
Week 9	
Tuesday, March 4	**Quiz 2**
Week 10	
Tuesday, March 18	**Guest Lecture by Andy Zou** Zou (2023) Universal and Transferable Adversarial Attacks on Aligned Language Models (pdf) **Lecture 8: Poisoning Attacks** 1. Liu (2018) Trojaning Attack on Neural Networks (pdf) 2. Gao (2020) Backdoor Attacks and Countermeasures on Deep Learning: A Comprehensive Review (pdf) 3. Bhagoji (2019) Analyzing Federated Learning through an Adversarial Lens (pdf) 4. Zhou (2021) Deep Model Poisoning Attack on Federated Learning (pdf)
Thursday, March 20	**Lecture 9: Defenses against Poisoning Attacks** 1. Wang (2019) Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks (pdf) 2. Huang (2019) NeuronInspect: Detecting Backdoors in Neural Networks via Output Explanations (pdf) 3. Steihhardt (2017) Certified Defenses for Data Poisoning Attacks (pdf)
Week 11	
Tuesday, March 25	**Lecture 10: Adversarial Machine Learning in Cybersecurity – Part I, Network Intrusion Detection** 1. Severi (2021) Explanation-Guided Backdoor Poisoning Attacks Against Malware Classifiers (pdf) 2. Rosenberg (2021) Adversarial Machine Learning Attacks and Defense Methods in the Cyber Security Domain (pdf) 3. Kuppa (2019) Black Box Attacks on Deep Anomaly Detectors (pdf)

	4. Demetrio (2019) Explaining Vulnerabilities of Deep Learning to Adversarial Malware Binaries (pdf)
Thursday, March 27	**Lecture 10 (cont'd): Adversarial Machine Learning in Cybersecurity – Part II, Malware Detection** **Due:** Assignment 4
Week 12	
Tuesday, April 1	**Lecture 10 (cont'd): Adversarial Machine Learning in Cybersecurity – Part III, Spam Filtering and URL Detection** 1. Erba (2019) Constrained Concealment Attacks against Reconstruction-based Anomaly Detectors in Industrial Control Systems (pdf) 2. Shirazi (2019) Adversarial Sampling Attacks Against Phishing Detection (pdf) 3. Anderson (2016) DeepDGA: Adversarially-Tuned Domain Generation and Detection (pdf) 4. Wang (2018) With Great Training Comes Great Vulnerability: Practical Attacks against Transfer Learning (pdf)
Thursday, April 3	**Lecture 10 (cont'd): Adversarial Machine Learning in Cybersecurity – Part IV, Cyber-physical and Biometric Systems**
Week 13	
Tuesday, April 8	**Lecture 11: Privacy Attacks Against Machine Learning Models** 1. Shokri (2018) Membership Inference Attacks Against Machine Learning Models (pdf) 2. Rigaki (2021) A Survey of Privacy Attacks in Machine Learning (pdf) 3. Hitaj (2017) Deep Models Under the GAN: Information Leakage from Collaborative Deep Learning (pdf) 4. Fredrikson (2015) Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures (pdf)
Thursday, April 10	**Lecture 11 (cont'd): Privacy Attacks Against Machine Learning Models**
Week 14	
Tuesday, April 15	**Lecture 12: Defenses Against Privacy Attacks** 1. Abadi (2016) Deep Learning with Differential Privacy (pdf) 2. Liu (2020) When Machine Learning Meets Privacy: A Survey and Outlook (pdf) 3. Nasr (2018) Machine Learning with Membership Privacy using Adversarial Regularization (pdf) 4. Papernot (2018) Scalable Private Learning with PATE (pdf) **Due:** Assignment 5
Thursday, April 17	**Lecture 12 (cont'd): Defenses Against Privacy Attacks** **Due:** Project proposal
Week 15	
Tuesday, April 22	**Lecture 13: Explainability in Machine Learning**

	1. Templeton et al. (2024) Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet (pdf) 2. Cunningham et al. (2023) Sparse Autoencoders Find Highly Interpretable Features in Language Models (pdf) 3. Sundararajan et al. (2017) Axiomatic Attribution for Deep Networks (pdf) 4. Montavon et al. (2017) Explaining Nonlinear Classification Decisions with Deep Taylor Decomposition (pdf)
Thursday, April 24	**Lecture 13 (cont'd): Explainability in Machine Learning**
Week 16	
Tuesday, April 29	**Lecture 14: Bias and Fairness in Machine Learning** 1. Navigli et al. (2023) Biases in Large Language Models: Origins, Inventory, and Discussion (pdf) 2. Gallegos et al. (2023) Bias and Fairness in Large Language Models: A Survey (pdf) 3. Mehrabi et al. (2021) A Survey on Bias and Fairness in Machine Learning (pdf) 4. Yang et al. (2017) An Adversarial Training Framework for Mitigating Algorithmic Biases in Clinical Machine Learning (pdf)
Thursday, May 1	**Lecture 14 (cont'd): Bias and Fairness in Machine Learning** **Due:** Project report
Week 17	
Tuesday, May 6	**Quiz 3**

Academic Integrity

Students are expected to adhere to the highest academic standards of honesty and integrity. At UI, we assume students will do their own work. Plagiarism—passing off someone else's work as your own, without citing the source—would not be tolerated. This includes direct copying, rephrasing, and summarizing, as well as taking someone else's idea and putting it in different words. The best avenue for avoiding plagiarism issues is to fully cite all sources used for preparing assignments, texts, and exams.

Learning Civility

In any environment in which people gather to learn, it is essential that all members feel as free and safe as possible in their participation. To this end, it is expected that everyone in this course will be treated with mutual respect and civility, with an understanding that all of us (students, instructors, professors, guests, and teaching assistants) will be respectful and civil to one another in discussion, in action, in teaching, and in learning.

Should you feel our classroom interactions do not reflect an environment of civility and respect, you are encouraged to meet with your instructor during office hours to discuss your concern. Additional resources for expression of concern or requesting support include the Dean of Students office and staff (208-885--6757), the UofI Counseling & Testing Center's confidential services (208-885-6716), the UofI Office of Equity and Diversity (208-885-2468), or the Office of Civil Rights and Investigations (208-885-4285).

Center for Disability Access & Resources (CDAR)
University of Idaho is committed to ensuring an accessible learning environment where course or instructional content are usable by all students and faculty. If you believe that you require disability-related academic adjustments for this class (including pregnancy-related disabilities), please contact Center for Disability Access and Resources (CDAR) to discuss eligibility. A current accommodation letter from CDAR is required before any modifications, above and beyond what is otherwise available for all other students in this class will be provided. Please be advised that disability-related academic adjustments are not retroactive. CDAR is located at the Bruce Pitman Building, Suite 127. Phone is 208-885-6307 and e-mail is cdar@uidaho.edu. For a complete listing of services and current business hours visit https://www.uidaho.edu/current-students/cdar.

Inclusivity Statement
As a professor/course instructor at the University of Idaho, I acknowledge the importance of diversity and inclusion and how these attributes contribute to the promotion of a positive educational experience. It is my intent to facilitate a healthy, productive, and safe learning environment where diverse thoughts, perspectives, and experiences are welcomed, and individuals' identities *(including, but not limited to: race, sex, class, sexual orientation, gender identity, ability, religious beliefs, etc.)* are valued and honored. I recognize that as an educator, it is my responsibility to take the initiative to continually learn about diverse perspectives and identities; therefore, if at any point during the course, you feel uncomfortable or concerned, I am more than willing to discuss suggestions, feedback, and anything else that might improve the general effectiveness of this course.

Healthy Vandals Policies
Please visit the University of Idaho COVID-19 webpage often for the most up-to-date information about the UofI's response to Covid-19.

Vandal Food Pantry
The Vandal Food Pantry is a free resource stocked weekly with food, grocery bags, and various hygiene items. Its eight locations across campus are accessible during building hours and open to all. Please take what you need.

Green Dot Safety Program
What's Your Green Dot? It's up to all of us to make a safer campus. Vandal Green Dot is a program that helps students learn about the power of the bystander, how to recognize potentially risky situations, and realistic ways to intervene. Together we can bring down the number of people being hurt by interpersonal violence on our campus. No one has to do everything, but everyone has to do something! Learn more and get involved by visiting UI's Green Dot Safety Program or emailing greendot@uidaho.edu.

Help and Resources
Student Resources
The University of Idaho provides student support to ensure a successful learning experience.

- Student Resources Webpage
- SI-PASS (Peer Assisted Study Sessions) SI-PASS provides regularly scheduled, peer-led study sessions for difficult courses.

Library Help
The UofI Library website has many databases that will help you find relevant and reliable books, articles, images, and more. Don't hesitate to contact a librarian for research assistance.

- UofI Library Website
- Help - Reference Services
- Help for Distance Ed Students

Technology Help
The UofI Student Technology Center provides many technology-related services to students.

- PHONE: 208-885-HELP (208-885-4357)
- Technology Help Email: support@uidaho.edu
- Technology Help Website

Writing Support
The UofI Writing Center provides one-on-one assistance to student writers and other members of the campus community.

- PHONE: 208-885-6644
- Writing Center Email: writingcenter@uidaho.edu
- Writing Center Website

UofI Moscow Land Acknowledgement
UofI Moscow is located on the homelands of the Nimiipuu (Nez Perce), Palus (Palouse) and Schitsu'umsh (Coeur d'Alene) tribes. We extend gratitude to the indigenous people that call this place home, since time immemorial. UofI recognizes that it is our academic responsibility to build relationships with the indigenous people to ensure integrity of tribal voices.