राष्ट्रीय प्रौद्योगिकी संस्थान पटना / NATIONAL INSTITUE OF TECHNOLOGY PATNA
संगणक विज्ञान एंव अभियांत्रिकी विभाग / DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
अशोक राजपथ, पटना-८००००५, बिहार / ASHOK RAJPATH, PATNA-800005, BIHAR
Phone No.: 0612-2372715, 2370419, 2370843, 2371929 Ext- 200, 202 Fax-0612-2670631 Website: www.nitp.ac.in

No:- Date:

CSX4171: Adversarial Machine Learning **L-T-P-Cr: 3-0-0-3**

Pre-requisites: None

Objectives:

- To study the approach and techniques for the continued development and deployment of learning in security-sensitive and adversarial environments.
- To understand both positive and negative results for the feasibility of classifier evasion.
- To understand secure learning as a game between an attacker and a defender with two real-world case studies of practical learning systems.

Course Outcomes*:* By the end of this course, students will be able to:

Sl. No.	Course Outcome (CO)	Program Outcome (PO)
1	Explain the fundamental concepts of adversarial machine learning and their relationship to statistical machine learning.	PO1, PO2
2	Analyze the security vulnerabilities of machine learning models using a security framework and identify different types of adversarial attacks (e.g., causative attacks, exploratory attacks).	PO2, PO5
3	Implement and evaluate the effectiveness of adversarial attacks against machine learning models in case studies involving spam filtering and anomaly detection.	PO3, PO4, PO5
4	Design and assess privacy-preserving mechanisms such as differential privacy to protect sensitive data during machine learning.	PO3, PO6, PO7
5	Develop and evaluate defenses against adversarial attacks, considering techniques like retraining with data replacements and corruption-resilient detectors.	PO3, PO4, PO11
6	Discuss open challenges and emerging defensive technologies in adversarial machine learning.	PO2, PO12

UNIT I

Overview of adversarial machine learning, Statistical machine learning. A framework for secure learning, security analysis, framework, exploratory attacks, causative attacks, repeated learning games, Privacy-Preserving learning.

UNIT II

Attacking a hypersphere learner, causative attacks on hypersphere detectors, hypersphere attack description, optimal unconstrained attacks, imposing time constraints on the attack, attacks against retraining with data replacements, constrained attackers.

UNIT III

Availability attack Case Study: SpamBayes, Spam Filter, threat model, causative attacks, RONI Defense, Experiments with SpamBayes. Integrity Attack Case Study: PCA Detector, traffic anomalies, corrupting the PCA subspace, Corruption-Resilient Detectors, empirical evaluation.

UNIT IV

Privacy-Preserving Mechanism for SVM learning, Privacy Breach case studies, problem setting: privacy-preserving learning, Support Vector Machines: A brief primer, differential privacy.

UNIT V

Near optimal evasion of classifiers, convex classes. Adversarial machine learning challenges, open problems, defensive technologies.

Text/References Books:

- Adversarial Machine Learning by Anthony D. Joseph, Blaine Nelson, Benjamin I. P. Rubinstein, J. D. Tygar
- Adversarial Machine Learning by Yevgeniy Vorobeychik, Murat Kantarcioglu