Topic 2: Adversarial Machine Learning and Defense Strategies.

Adversarial attacks can disrupt any AI/ML based system functionalities; while handling such attacks are challenging, but also provide significant research opportunities. This talk will cover emerging adversarial machine learning (AML) attacks on systems and the state-of-the-art defense techniques. First, I will discuss how and where adversarial attacks can happen in an AI/ML model and framework. I will then present classification of adversarial attacks and their severity and applicability in real-world problems and what steps can be taken to mitigate their effects. The role of GAN in in adversarial attacks and as a defense strategy. I will discuss a dual-filtering strategy could mitigate adaptive or advanced adversarial manipulations for wide-range of ML attacks with higher accuracy. The developed dual-filter software can be used as a wrapper to any existing ML-based decision support system to prevent a wide variety of adversarial evasion attacks. The DF framework utilizes two set of filters based on positive (input filters) and negative (output filters) verification strategies that can communicate with each other for higher robustness.

References:

- Dual-filtering (DF) schemes for learning systems to prevent adversarial attacks. D Dasgupta, KD Gupta - Complex & Intelligent Systems, pp 1-22, January 2022
- Who is responsible for Adversarial Defense. K Dattagupta and D. Dasgupta. Workshop on Challenges in Deploying and monitoring Machine Learning Systems, ICML 2021.
- Applicability issues of evasion-based adversarial attacks and mitigation techniques. KD Gupta, D Dasgupta, Z Akhtar - 2020 IEEE Symposium Series on Computational …, 2020