Dependable and Secure AI-ML

1. Credit requirement:(L-T-P: 3-0-0, Credit: 3)
2. Please select the committee for Approval: PGPEC
3. Name of the Dept: CoAI
4. Please Specify the Level of the Subject: PG level
5. Whether the subject will be offered as compulsory or elective: Elective
6. Prerequisite(s) for the subject, if any (Please give the subject numbers and names): AI61005 ARTIFICIAL INTELLIGENCE: FOUNDATIONS AND APPLICATIONS, Machine learning
7. **Course Objective:** As ML and AI applications become prevalent in various aspects of everyday life, their dependability and security take on increasing importance. The advent of Deep Neural Network (DNN) based AI with its close to human-level precision have paved the way for being used in several safety-critical applications. Often a single failure can lead to catastrophic results in such applications, making the trust and dependability of such systems imperative. But is AI fundamentally robust, specifically against malicious adversaries? Unfortunately, existing research shows that classical machine learning can be targeted by adversaries to breach their decision boundaries. Such adversarial inputs have been shown to be catastrophic for applications like automated driving, where such AI systems are seamlessly getting deployed. For instance, a crafted visibly imperceptible change to a 'STOP' signal can be wrongly inferred by a ML engine as 'GO', leading to obvious dire consequences. Therefore, improving the robustness of these models is of significant importance to avoid any disastrous event. Moreover, traditional redundancy-based fault mitigation techniques cannot always be employed in all sorts of AI applications due to their high overheads and undesirable resource consumption. On the other hand, huge data volume and processing requirement of AI-ML algorithms are often outsourced to cloud platforms which adds another concern of security. That raises pertinent question on how to perform AI processing on cloud maintaining data as well as model security. In this course, we discuss such few aspects of trustworthy AI mainly focusing on reliability, security and safety of emerging AI-ML applications. The course will cite examples from emerging safety-critical artificial intelligence applications such as self-driving car and health applications.
8. **Study Materials:**

   Some of the important references are:

   - FAULT TOLERANCE IN ARTIFICIAL NEURAL NETWORKS, George Ravuama Bolt, D.Phil. Thesis University of York, 1992
   - Book: Privacy-Preserving Machine Learning, J. Morris Chang, Di Zhuang, and G. Dumindu Samaraweera, Manning publication.
   - Raphael Bost, Raluca Ada Popa, Stephen Tu, Shafi Goldwasser: Machine Learning Classification over Encrypted Data. NDSS 2015
   - Michele Minelli. Fully homomorphic encryption for machine learning. Cryptography and Security [cs.CR]. Université Paris sciences et lettres, 2018. English. ffNNT : 2018PSLEE056ff. fftel-01918263v2f

   Paper References:

   1. A. Avizienis, J. -. Laprie, B. Randell and C. Landwehr, "Basic concepts and taxonomy of dependable and secure computing," in IEEE Transactions on Dependable and Secure Computing, vol. 1, no. 1, pp. 11-33, Jan.-March 2004, doi: 10.1109/TDSC.2004.2.

   2. Bagchi, Saurabh, et al. "Vision Paper: Grand Challenges in Resilience: Autonomous System Resilience through Design and Runtime Measures." IEEE Open Journal of the Computer Society (2020).

   3. Game-Theoretic Methods for Robustness, Security, and Resilience of CPS Control Systems (https://www.semanticscholar.org/paper/Game-Theoretic-Methods-for-Robustness

4. Measuring the Reliability of Reinforcement Learning Algorithms

(https://arxiv.org/abs/1912.05663)

5. MTFuzz: Fuzzing with a Multi-Task Neural Network

(https://arxiv.org/abs/2005.12392)

6. Model Assertions for Monitoring and improving ML models

(https://arxiv.org/pdf/2003.01668.pdf)

7. Dependable Deep Learning: Towards Cost-Efficient Resilience of Deep Neural Network Accelerators against Soft Errors and Permanent Faults.

(https://ieeexplore.ieee.org/stamp/stamp.jsp?tp= arnumber=9159734)

8. Improving the Dependability of Machine Learning Applications

(https://academiccommons.columbia.edu/doi/10.7916/D8P2761H)

9. A Review: Generative Adversarial Networks

https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8833686

10. Formal Scenario-Based Testing of Autonomous Vehicles: From Simulation to the Real World

(https://arxiv.org/pdf/2003.07739.pdf)

11. Trojan attack on Neural Networks

(https://docs.lib.purdue.edu/cgi/viewcontent.cgi? article=2782 context=cstech)

12. Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks

(https://people.cs.uchicago.edu/ ravenben/publications/pdf/backdoor-sp19.pdf)

9. **Syllabus**

(a) Introduction to dependable AI: Resilience, robustness, safety and security [2]

(b) Reliable Neural Networks: Fault Models, Assessing Fault Tolerance, Redundancy, Reliability during the Learning Phase. [4]

(c) Methodology for Fault Tolerance: Visualisation Levels for Neural Networks, Fault Locations, Fault Manifestations, Fault Coverage. [4]

(d) Game-Theoretic Methods for Robustness, Security, and Resilience [2]

(e) Fuzzing for vulnerability detection [2]

(f) Integrity checks and monitoring [2]

(g) Low-cost fault-mitigation techniques, improving the dependability through software testing [3]

(h) Measuring the Reliability of Reinforcement Learning Algorithms, Generative adversarial networks [4]

(i) Case study: 1. Data-driven verification for automotive systems
2. Robustness in Non-stationary Health Records [2]

(j) Advanced Topic (if time permits): Provable safety and provable Defense, Formal Scenario-Based Testing of Autonomous Vehicles:From Simulation to the Real World

(k) Secure AI: Privacy concerns in ML and DL, Adversarial models: Honest-but-curious adversary model, semi-honest entity, active adversary model. [2]

(l) Attacks against ML/DL: [4]
Evasion/Adversarial attack, Poisoning, Inference, Trojans, Backdoor attacks Case Study: Facial Recognition Systems

(m) Differential Privacy basics: [2] Properties of Differential Privacy: privacy preservation, Sensitivity, randomization, composition, and stability;
Differential Privacy in Supervised Learning, Differential Privacy in Unsupervised Learning

(n) Federated machine learning: [2]

Federated Learning basics: Model Training in Federated Learning and optimisation, Privacy-Preservation in centralised FL framework, Attack Models on FL, Privacy-preservation solutions

Case Study : FL in finance

(o) Homomorphic encryption and machine learning : [2]

Basics of homomorphic encryption, Secure hyperplane decision, Naïve Bayes, and decision trees-polynomial approximations , Division-Free Integer Algorithms for Classification, Homomorphic evaluation of deep neural networks, Case study on medical data

10. Names of the faculty members of the Department/Centers/School who have the necessary expertise and will be the willing to teach the subject (Minimum two faculty members should be willing to teach the subject)

11. Do the contents of the subject have an overlap with any other subject offered in the Institute?

- Approximate percentage of overlap:
- Reasons for offering the new subject in spite of the overlap:
  No existing course with similar focus exists.