Cybersecurity in AI and ML (5 Days)

By Dr. Vishwanath Rao

Course Objectives

By the end of the course, participants will:

1. **Understand the Intersection of Cybersecurity and AI/ML**: Grasp the fundamental cybersecurity concepts and how AI/ML technologies are leveraged to enhance and compromise security.
2. **Identify Security Threats in AI/ML Systems**: Recognize the unique vulnerabilities of AI/ML systems, including adversarial attacks, model inversion, data poisoning, and evasion.
3. **Defend AI/ML Systems**: Learn how to secure AI/ML systems by implementing robust models, validating data pipelines, and mitigating adversarial attacks.
4. **Utilize AI for Cyber Defense**: Explore how AI and ML can be deployed to strengthen cybersecurity measures, including threat detection, anomaly detection, and automated incident response.
5. **Address Ethical and Regulatory Issues**: Understand the ethical concerns surrounding AI/ML in cybersecurity, such as bias, privacy, and the dual-use dilemma, while navigating legal and regulatory frameworks.
6. **Hands-on Implementation**: Gain practical experience through labs and case studies, applying AI/ML techniques to real-world cybersecurity challenges.

Target Audience

This course is designed for:

- **Cybersecurity Professionals**: Security analysts, engineers, and architects who want to integrate AI/ML techniques into their security strategies.
- **AI/ML Practitioners**: Data scientists, ML engineers, and AI researchers looking to understand the security implications of their models.
- **IT Professionals and System Administrators**: Those responsible for managing and securing AI/ML systems.
- **Security Consultants and Auditors**: Consultants who audit or advise organizations on secure AI/ML implementations.
- **Software Developers**: Developers creating AI-powered applications or systems with a security focus.
- **Students and Academics**: Those pursuing careers or research in cybersecurity, AI, or data science.

Course Prerequisites

Participants should have:

1. **Basic Cybersecurity Knowledge**: Familiarity with foundational cybersecurity concepts, including network security, encryption, and common cyber threats (e.g., malware, phishing).

2. **Understanding of AI/ML Concepts**: A basic understanding of machine learning concepts, including types of learning (supervised, unsupervised, reinforcement), as well as exposure to common ML algorithms (e.g., decision trees, neural networks).
3. **Programming Skills**: Proficiency in a programming language commonly used in AI/ML, such as Python, with experience in libraries like TensorFlow, Keras, or PyTorch being advantageous.
4. **Familiarity with Data Science Concepts**: Knowledge of data preprocessing, feature engineering, and model evaluation metrics.

Course Contents

AI, ML and DL

- Overview of AI: Key concepts, history, and modern applications
- Machine Learning (ML) Basics:
- Supervised
- Unsupervised
- Reinforcement Learning

ML Algorithms:

- Regression
- Classification
- Clustering
- Neural Networks
- Deep Learning

Data in ML:

- Data collection
- Preprocessing
- feature engineering

Generative AI

- What is Generative AI?:
- Overview of generative models and their capabilities (e.g., GANs, VAEs, Transformers)
- Types of Generative Models:
- Generative Adversarial Networks (GANs)
- Variational Autoencoders (VAEs)
- Transformer-based Models (e.g., GPT, BERT)
- Applications of Generative AI: Text, image, video, and audio generation, content creation, and data augmentation

MLOps

- MLOps – Manual Lifecycle
- MLOps – Automated Pipeline

Introduction to AI and ML Security:

- Overview of AI and ML technologies
- Importance of security in AI and ML systems
- Distinction between traditional software security and AI/ML security

Threat Landscape:

- Common security threats and attacks targeting AI and ML systems
- Adversarial attacks: poisoning, evasion, data integrity attacks
- Privacy risks: data leakage, membership inference attacks
- Model stealing and intellectual property theft
- Intersection of AI, ML, and Cybersecurity
- How AI/ML Enhances Cybersecurity:
- AI-driven threat detection
- Anomaly detection with ML
- Predictive analytics in cybersecurity
- Behavior-based detection of malware and attacks
- Current Applications: Intrusion Detection Systems (IDS), Antivirus systems, Network Traffic Analysis, Fraud Detection
- Security Threats and Vulnerabilities in AI/ML Systems

Secure Development Tools and Techniques:

- Introduction to security testing tools (e.g., static analysis, dynamic
- analysis)
- Automated security testing and continuous integration (CI) pipelines
- Detailed discussion on ART – Adversarial Robustness Toolbox
- Other tools will be address later in the course

Adversarial Machine Learning:

- Adversarial Attacks
- Poisoning Attack
- Evasion Attack
- Data Disclosure
- Data Integrity Attack
- Data Leakage
- Membership inference Attack

Different Evasion Attack

Whitebox attacks

- Auto Projected Gradient Descent (Auto-PGD)
- Shadow Attack
- Wasserstein Attack
- PE Malware Attacks
- Brendel & Bethge Attack
- Targeted Universal Adversarial Perturbations
- High Confidence Low Uncertainty (HCLU) Attack
- Iterative Frame Saliency
- Dpatch
- Carlini & Wagner (C&W) L_2 and L_inf attack
- Basic Iterative Method (BIM)
- Universal Perturbation
- DeepFool
- Virtual Adversarial Method
- Fast Gradient Method

Blackbox attacks

- Square Attack
- HopSkipJump Attack
- Threshold Attack
- Pixel Attack
- Simple Black-box Adversarial (SimBA)
- Spatial Transformation
- Query-efficient Black-box
- Zeroth Order Optimisation (ZOO)
- Decision-based/Boundary Attack
- Geometric Decision-based Attack (GeoDA)

Poison Attacks

- Poisoning Attack on Support Vector Machines (SVM)
- Backdoor Attack
- Clean-Label Backdoor Attack
- Adversarial Embedding Backdoor Attack
- Hidden Trigger Backdoor Attack
- Bullseye Polytope
- Backdoor Attack on Deep Generative Models (DGM)
- BadDet Attacks

Extraction Attacks

- Functionally Equivalent Extraction
- Copycat CNN
- KnockoffNets

Inference Attacks

- Attribute Inference
  - Attribute Inference Black-Box
  - Attribute Inference White-Box Lifestyle Decision Tree
  - Attribute Inference White-Box Decision Tree
- Membership Inference
  - Membership Inference Black-Box
  - Membership Inference Black-Box Rule-Based
  - Label-Only Boundary Distance Attack
  - Label-Only Gap Attack
- Model Inversion
  - MIFace
- Reconstruction
  - Database Reconstruction

Model Stealing and Intellectual Property Theft

- Model Stealing
- Model Disclosure
- Intellectual Property Theft
- Model Robustness

Adversarial Machine Learning- Defense

- Adversarial Examples
- Crafting Attacks
- Defending against clever attacks
- Defensive Distillation
- Detection and Response

Secure model deployment architectures

- Containerization and isolation for model serving
- Authentication and access control in AI/ML systems

Privacy preserving

- Differential Privacy

- Secure Aggregation
- Homomorphic Encryption
- Zero-Knowledge Proofs

Data Security and Governance:

- Data Exfil
- Ethical implications of AI and ML security
- Bias and fairness in AI/ML systems
- Responsible AI principles and guidelines

Input Validation and Output Encoding:

- Importance of input validation and output encoding
- Techniques for sanitizing and validating input data
- Preventing common injection attacks (e.g., SQL injection, XSS)

Secure Communication:

- Securing network communication (e.g., HTTPS/TLS)
- Implementing secure APIs and web services
- Transport layer security best practices

Data Protection:

- Encryption fundamentals
- Protecting sensitive data at rest and in transit
- Key management and secure storage practices

Authentication and Authorization:

- Understanding authentication and authorization mechanisms
- Secure implementation of authentication (e.g., password hashing,
- multi-factor authentication)
- Role-based access control (RBAC) and least privilege principles

Intrusion Detection Systems (IDS):

- Introduction to IDS and their role in threat detection
- Types of IDS (e.g., network-based, host-based) and their deployment models

Intrusion Prevention Systems (IPS):

- Overview of IPS and their capabilities
- Techniques for preventing and mitigating intrusions in real-time

- Security Information and Event Management (SIEM):
- Understanding SIEM and its role in centralized log management and analysis
- Using SIEM for threat detection, incident response, and compliance reporting

OWASP Top 10 AI Security Risks:

- Identification of the top security risks specific to AI systems
- Examples of vulnerabilities and threats in AI applications
- OWASP ML Top 10
- Introduction to OWASP LLM Top 10

Security Awareness and Training:

- Importance of security awareness for developers
- Techniques for promoting a security culture within development teams
- Continuous learning and staying updated on security trends

Introduction to Risk Management Framework (RMF):

Overview of NIST RMF and its importance in cybersecurity

RMF Process Overview:

Detailed explanation of the six steps in the RMF process:

- Prepare
- Categorize
- Select
- Implement
- Assess
- Authorize
- Monitor

Securing Cloud Applications:

- Application security considerations in the cloud
- Secure development practices for cloud-native applications

Introduction to Threat Detection and Prevention:

- Overview of cybersecurity threats and their impact
- Importance of proactive threat detection and prevention measures
- Understanding threat intelligence sources and feeds
- Techniques for collecting, analyzing, and applying threat intelligence

Vulnerability Management:

- Identifying and assessing vulnerabilities in systems and applications
- Strategies for prioritizing and remedying vulnerabilities

Data Security in AI and ML:

- Overview of data security challenges
- Importance of data protection for maintaining confidentiality, integrity and availability
- Data Privacy Regulations and Compliance
- Data Classification and Sensitivity
- Data Collection and Acquisition
- Data Storage and Encryption
- Data Transmission and Network Security
- Data Masking and Anonymization
- Data Governance and Compliance Monitoring

Incident Response and Data Breach Management

- Data breach notification requirements and mitigation strategies

Introduction to GDPR and CCPA:

- Overview of the GDPR and its objectives
- Scope of the regulation and its applicability to organizations handling personal data of EU residents
- Data protection principles, including lawfulness, fairness, and transparency
- Purpose limitation, data minimization, and storage limitation principles
- Rights of data subjects, including the right to access, rectification, erasure and data portability
- Understanding the right to be forgotten and its implications

Data Processing Requirements:

- Requirements for lawful processing of personal data
- Conditions for obtaining valid consent
- Data Breach Notification
- Requirements for data breach notification under GDPR
- Timelines and procedures for reporting data breaches to supervisory authorities and affected individuals
- Data Protection Officer (DPO) Role
- Responsibilities and qualifications of the DPO under GDPR
- Role of the DPO in ensuring compliance with GDPR requirements
- Cross-Border Data Transfers
- Legal mechanisms for transferring personal data outside the EU

- Understanding the requirements for adequacy decisions, standard contractual clauses, and binding corporate rules
- GDPR Enforcement and Penalties:

Security Tools for Cybersecurity in AI and ML

- Model Scan
- Nb Defense
- Foolbox
- Audit.ai
- ML Privacy Meter
- Text Attack
- ART – Adversarial Robustness Toolbox
- AIF 360
- Guard- Rails
- AUGLY
- Giskard Test Tool
- OpenAttack
- Garak – LLM
- ANON – LLM
- Safe-Tensors
- PyRit

LLMs and Security:

Overview of Large Language Models (LLMs) and their applications
Introduction to security risks associated with LLMs

- Injection Attacks
- Broken Authentication
- Sensitive Data Exposure
- Risks of exposing sensitive information through LLM outputs
- XML External Entities (XXE)
- Security Misconfiguration
- Cross-Site Scripting (XSS)
- Insecure Deserialization
- Insufficient Logging and Monitoring
- Future Trends and Emerging Technologies

Introduction to RAG Applications and AI Agents:

- Overview of RAG applications and AI agents and their role in risk

- management and assurance processes.
- Introduction to the security challenges associated with these systems.

Data Privacy and Confidentiality Risks

- Data Integrity Risks
- Authentication and Authorization Risks
- Third-Party and Supply Chain Risks
- Model Security Risks
- Compliance and Regulatory Risks
- Social Engineering Attacks

Future Trends and Emerging Technologies:

- Emerging trends in RAG applications and AI security.
- Implications of new technologies (e.g., blockchain, federated learning) on
- security risks and mitigation strategies