SECURITY & PRIVACY OF AI

Subject Code: CS35322CS	**Subject Type:** DE (Elective-3)	**Evaluation Scheme:** MSE (30), TA (20), ESE (50)	**L-T-P:** 3-0-0 (3)

Pre-Requisites: Machine Learning basics, Deep Learning fundamentals, Cryptography and Network Security (basic), Probability & Statistics

Course Objective: To understand security threats and privacy risks in AI/ML systems. To study adversarial machine learning attacks and defenses. To explore privacy-preserving AI techniques. To design secure and trustworthy AI systems.

Course Outcomes: At the end of the course, the student will be able to

CO	Course Outcomes
CO1	Understand security and privacy fundamentals in AI systems
CO2	Analyze adversarial attacks and threat models
CO3	Apply defense mechanisms to secure AI models
CO4	Implement privacy-preserving ML techniques
CO5	Design and evaluate secure AI applications

Syllabus:

Unit I: Foundations of AI Security and Privacy: Introduction to AI Security and Privacy: Evolution of AI systems and emerging security concerns, Importance of security and privacy in AI applications, Differences between traditional cybersecurity and AI security, Real-world examples of AI failures and attacks; Threat Models in AI: White-box, Black-box, Grey-box attack models, Attacker capabilities and knowledge assumptions, Training-time vs inference-time attacks, Active vs passive adversaries; Security Goals in AI Systems: Confidentiality, Integrity, Availability (CIA triad), Authenticity and accountability in AI, Model robustness and reliability; Privacy Risks in AI Systems: Data leakage and information exposure, Risks in centralized vs distributed learning, Privacy concerns in datasets and model outputs; Trustworthy AI Principles: Robustness and reliability, Fairness and bias mitigation, Explainability and interpretability, Transparency and accountability; Overview of Adversarial Machine Learning.

Unit II: Adversarial Machine Learning: Evasion Attacks (Test-Time Attacks): Concept of adversarial examples, Gradient-based attacks (FGSM, PGD), Optimization-based attacks, Physical-world adversarial attacks; Poisoning Attacks (Training-Time Attacks): Data poisoning techniques, Label flipping and clean-label attacks, Impact on model performance; Backdoor Attacks: Trigger-based attacks, Trojan attacks in neural networks, Detection challenges; Real-World Adversarial Examples: Adversarial images and patches, Attacks in NLP and speech systems, Attacks in autonomous systems; Attack Models and Scenarios: Active vs passive attacks, Training vs inference attacks, Targeted vs untargeted attacks.

Unit III: Defense Mechanisms in AI: Adversarial Training: Basic concept and formulation, Robust training techniques, Trade-off between accuracy and robustness; Out-of-Distribution (OOD) Detection: Definition and importance, Detection methods, Relationship with adversarial examples; Certified Robustness Methods: Formal robustness guarantees, Randomized smoothing, Verification techniques; Defense Against Poisoning & Backdoor Attack: Data sanitization techniques, Anomaly detection methods, Robust learning algorithms; Secure Model Evaluation: Robustness evaluation metrics, Benchmark datasets and testing strategies, Adversarial testing frameworks; Robust ML System Design: Secure ML pipelines, Model hardening techniques, Deployment considerations.

Unit IV: Privacy-Preserving Machine Learning: Privacy Risks in ML Models: Data exposure during training and inference, Leakage through model outputs, Privacy vs utility trade-off; Membership Inference Attacks (MIA): Attack methodology, Overfitting and privacy leakage, Defense strategies; Model Inversion & Property Inference: Inferring sensitive attributes, Class reconstruction attacks, Risks in healthcare and biometrics; Model Extraction Attacks: Stealing model parameters, Query-based attacks, Surrogate model construction; Differential Privacy (DP): Definition and principles, epsilon privacy parameter, DP in machine learning (DP-SGD); Secure Multi-Party Computation (SMC): Secure collaborative computation, Secret sharing techniques, Applications in ML; Homomorphic Encryption (HE): Computation on encrypted data, Types of HE, Applications in secure AI; Federated Learning Privacy: Distributed training framework, Secure aggregation, Privacy risks and mitigation.

Unit V: Security & Privacy in Modern AI Systems: Security in Deep Learning Models: Security in CNNs (vision systems), Security in LLMs (prompt injection, data leakage), Security in AI agents and autonomous systems; Privacy in AI Applications: Privacy in NLP systems, Privacy in computer vision (face recognition risks), Data anonymization techniques; Secure Deployment of AI Systems: Secure model serving, Monitoring and incident response, AI system lifecycle security; Trusted Execution Environments (TEE): Secure enclaves (Intel SGX, ARM TrustZone), Confidential computing, Secure inference; AI Governance, Ethics, and Regulations: GDPR and data protection laws, Ethical AI principles, Responsible AI frameworks; Case Studies and Applications: Healthcare AI (privacy-sensitive data), Financial fraud detection systems, Recommendation systems, Smart cities and IoT security.

Learning Resources:

Text Books:

1. Jason Edwards, Adversarial Machine Learning: Mechanisms, Vulnerabilities, and Strategies for Trustworthy AI, Wiley, 2026.
2. Anthony D. Joseph et al., Adversarial Machine Learning, Cambridge University Press, 2019.

Reference Books:1. Ehsan Nowroozi et al. (Eds.), Adversarial Example Detection and Mitigation Using Machine Learning, Springer, 2025.