NIRMA UNIVERSITY

Institute:	Institute of Technology
Name of Programme:	BTech CSE
Course Code:	4CS105DE25
Course Title:	Securing AI Models
Course Type:	Disciplinary Minor-Elective
Year of Introduction:	2025-26

L	T	Practical Component				C
		LPW	PW	W	S	
3	0	2	-	-	-	4

Course Learning Outcomes (CLO):

At the end of the course, the students will be able to –

1. explain the security fundamentals related to AI models and their standards (BL2)
2. apply ethical considerations and responsibilities associated with AI development and security (BL3)
3. make use of the best practices for handling sensitive data in AI applications while ensuring compliance with relevant laws and standards (BL3)
4. analyse the security measures for AI models, including projects for deployment purpose. (BL4)

Unit	Contents	Teaching Hours (Total 45)
Unit-I	**Introduction to AI Security:** What is AI Security? Key Concepts in AI Security, Common AI security challenges and threats, Understanding the importance of securing AI models, Ethics and responsibility of securing AI, Overview of security frameworks and guidelines specific to AI, such as NIST SP 800-193 and OWASP AI Security. **Model Robustness:** Understanding AI models: types and lifecycle, Risks and Threats to AI model, Techniques for model robustness, Model explainability and transparency, adversarial attacks, Effects of adversarial attacks and preventions, Defenses against adversarial attacks.	10
Unit-II	**Data and Dataset:** Data Exploration and visualization, Data splitting and cross-validation, Data Augmentation, Imbalanced data, Working with time series data, Dataset preparation, data labeling and annotation, Data Versioning and management, Data ethics and bias, Data quality and governance. **Privacy and Data Protection:** Data security and privacy, Privacy in AI and data protection regulations, Data anonymization and de-identification, Federated learning and differential privacy, GDPR and AI model compliance.	08
Unit-III	**Secure Model Deployment:** Model deployment and inference security, Containerization and isolation techniques, Secure APIs and model servers, Monitoring and anomaly detection in deployed models.	08

	AI Infrastructure Security: Cloud security for AI, Securing AI Deployment Environments, Network Security for AI Systems, Incident response, and disaster recovery	
Unit-IV	**Ethical and Responsible AI:** Threat Intelligence for AI Security, Threat Hunting in AI Systems, Bias and fairness in AI, Explainability and interpretability, Ethical Considerations in AI Security, Privacy-Preserving AI, Responsible AI Governance, Regulatory Overview, Case Studies in Responsible AI. **Best Practices in AI Model Security:** Implementing robust data security protocols, Secure and private AI design principles, Regular monitoring and audits, employing multi-factor authentication, prioritizing user education, and Developing incident response plans.	10
Unit-V	**Case Studies and Practical Applications:** Real-world examples of AI security breaches, Hands-on exercises in identifying vulnerabilities, Secure development and best practices, and Model-specific security challenges. **Emerging Threats and Future Trends:** Emerging AI security threats, Deepfake technology and countermeasures, AI in cybersecurity, Preparing for the future of AI security.	09

Self-Study:

The self-study contents will be declared at the commencement of the semester. Around 10% of the questions will be asked from self-study content.

Suggested Readings/ References:

1. Emmanuel Ameisen, Building Machine Learning Powered Applications: Going from Idea to Product, O'Reilly.
2. Bernard Marr, Artificial Intelligence in Practice: How 50 Successful Companies Used AI and Machine Learning to Solve Problems, Wiley
3. Laurent Gil and Allan Liska, Security with AI and Machine Learning, O'Reilly Media, Inc.
4. Anthony D. Joseph, Blaine Nelson, Benjamin I. P. Rubinstein, and J. D. Tygar, Adversarial Machine Learning, Cambridge University Press

Suggested List of Experiments:

Sr. No.	Title	**Hours**
1	Environment setup, including Python, relevant libraries, and virtualization or containerization tools	02
2	Containerization and sandboxing for model deployments Configure firewall rules and network security for AI model servers.	02
3	Implement user authentication and authorization mechanisms. Secure API key management	04
4	Model various encryption techniques. Implement data encryption at rest and in transit.	04
5	Implement HTTPS and secure API protocols. Implement rate limiting and access control for model APIs.	02
6	Implement anomaly detection for API requests. Set up a comprehensive logging and log monitoring systems.	04

7	Watermark AI models to track theft. Create surrogate models through distillation.	02
8	Explore legal protections and intellectual property rights.	04
9	Final project: Develop and deploy a secure AI model using the skills learned in previous experiments.	06