CO1	Demonstrate a thorough understanding of key principles related to dependable machine learning, including explainability, interpretability, safety, and robustness
CO2	Assess and critique the reliability of machine learning models, including their performance, safety measures, and resilience to adversarial attacks
CO3	Develop machine learning models that provide explanations for their predictions and decisions, ensuring transparency and trustworthiness
CO4	Effectively communicate the reliability, limitations, and safety measures of machine learning systems to both technical and non-technical stakeholders.

S. NO	Contents	Contact Hours
UNIT 1	Introduction: Overview, Motivation, Challenges – medical and surveillance. Explainable AI: Accuracy-explainability, Tradeoff, Interpretability Problem, Predictability, Transparency, Traceability, Causality, Reasoning, Attention and Saliency	**10**
UNIT 2	Interpretable AI: Prediction Consistency, Application Level Evaluation, Human Level Evaluation, Function, Level Evaluation. Adversarial Robustness: Adversarial Attacks and Defenses	**10**
UNIT 3	Trustworthy AI: Integrity, Reproducibility, Accountability, Bias-free AI: Accessibility, Fair, Data Agnostics Design, Disentanglement. Privacy-Preserving AI: Federated Learning, Differential Privacy and Encrypted Computation	**12**

UNIT 4	Verified AI: Environment and Specification Modeling, Design with Formal Inductive Synthesis, and Evaluation. Platforms for AI Safety	**10**
	TOTAL	**42**