Dependable Machine Learning	L	T	P	Machine Learning
	3	0	2	

Course Objective: To provide characteristic details of AI and machine learning systems to make them dependable, such as explainability, interpretability, safety etc.

S. NO	Course Outcomes (CO)
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**

REFERENCES		
S.No.	**Name of Books/Authors/Publishers**	**Year of Publication / Reprint**
1	Fairness and Machine Learning: Limitations and Opportunities by Solon Barocas, Moritz Hardt, Arvind Narayanan, MIT Press	2023
2	Ethics of Artificial Intelligence by S. Matthew Liao, OUP USA	2020
3	Interpretable Machine Learning, by Christoph Molnar, https://www.lulu.com/	2020
4	The Ethics of Artificial Intelligence by Luciano Floridi, Oxford Univ Press	2023
5	https://fairmlclass.github.io/	2022

B.Tech. Information Technology				
Course code: Course Title	**Course Structure**			**Pre-Requisite**
Embedded Systems	**L**	**T**	**P**	Knowledge of Computer Architecture, Microprocessors
	3	**1/0**	**0/2**	