INSTITUTE OF AERONAUTICAL ENGINEERING
(Autonomous)
Dundigal - 500 043, Hyderabad, Telangana

COURSE CONTENT

Explainable AI (XAI)								
VIII Semester: CSE								
Course Code	**Category**	**Hours / Week**			**Credits**	**Maximum Marks**		
		L	**T**	**P**	**C**	**CIA**	**SEE**	**Total**
ACSE41	**Elective**	**3**	**-**	**-**	**3**	**40**	**60**	**100**
Contact Classes: 48	**Tutorial Classes: Nil**	**Practical Classes: Nil**				**Total Classes: 48**		
Prerequisite: Information Security and Management								

I. COURSE OVERVIEW:
This course provides a comprehensive introduction to Explainable Artificial Intelligence (XAI), focusing on the need for transparency, interpretability, and trust in machine learning and deep learning models. It covers fundamental concepts, interpretable models, model-agnostic and model-specific explainability techniques, evaluation methods, ethical considerations, and real-world applications. The course enables students to understand, analyze, and apply explainable AI techniques to build trustworthy and accountable AI systems.

II. COURSES OBJECTIVES:
The students will try to learn

I. The fundamentals and importance of Explainable Artificial Intelligence (XAI) in modern AI systems.
II. The concepts of interpretability, transparency, and trust in machine learning and deep learning models.
III. The working principles of interpretable and intrinsic machine learning models.
IV. Various model-agnostic and model-specific explainability techniques such as LIME, SHAP, and feature importance methods.

III. COURSE OUTCOMES:
At the end of the course students should be able to:

CO1 Understand the need for Explainable Artificial Intelligence and differentiate between black-box and interpretable models.
CO2 Explain the fundamental concepts, types, and principles of explainability in machine learning and deep learning models
CO3 Apply interpretable and intrinsic models such as linear models and decision trees for transparent decision making.
CO4 Analyze and implement model-agnostic explainability techniques such as LIME and SHAP for interpreting complex models.
CO5 Evaluate explainability methods using appropriate metrics and address ethical, fairness, and trust issues in AI systems.
CO6 Demonstrate the use of Explainable AI techniques in real-world applications such as healthcare, finance, computer vision, and natural language processing

IV. COURSE CONTENT:
MODULE - I: INTRODUCTION TO EXPLAINABLE AI (XAI) (09) (9)
Introduction to Explainable AI, need for explainability, black-box vs interpretable models, transparency and trust in AI systems, interpretability vs explainability, types of explanations, global vs local explainability, intrinsic vs post-hoc explanations.

MODULE – II: INTERPRETABLE AND INTRINSIC MODELS (10)
Interpretable machine learning models, linear regression and logistic regression, decision trees, rule-based systems, sparse models, attention mechanisms, interpretable neural networks, layer-wise relevance propagation (LRP).

MODULE – III: EXPLAINABILITY TECHNIQUES IN XAI (10)
Model-agnostic explainability methods, feature importance, LIME (Local Interpretable Model-Agnostic Explanations), SHAP (Shapley Additive explanations).

Partial Dependence Plots (PDP), Individual Conditional Expectation (ICE) plots, gradient-based explainability methods, saliency maps, integrated gradients, Grad-CAM, visual explainability for deep learning models

MODULE – IV: EVALUATION, ETHICS AND TRUST IN XAI (10)
Evaluation of explainability methods, quantitative and qualitative metrics, human-centric evaluation, fairness and bias in AI models, ethical issues in explainable AI, accountability, transparency, regulatory and legal aspects of XAI.

MODULE – V: APPLICATIONS OF EXPLAINABLE AI ((9)
Explainable AI in healthcare, explainability in finance and credit risk analysis, explainable AI for autonomous systems, applications in computer vision, applications in natural language processing, real-world case studies and industry use cases of XAI.

IV. TEXT BOOKS:

1. Christoph Molnar, *"Interpretable Machine Learning: A Guide for Making Black Box Models Explainable"*, 2nd Edition, Leanpub, 2022.
2. Uday Kamath, John Liu, James Whitaker, *"Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning"*, Springer, 2021.
3. Parikshit Narendra Mahalle, Yashwant Sudhakar Ingle, *"Explainable Artificial Intelligence: Foundations, Techniques, and Applications"*, Wiley, 2020

V. REFERENCE BOOKS:

1. Pradeepta Mishra, *"Explainable AI Recipes: Implement Solutions to Model Explainability and Interpretability with Python"*, Apress, 2023.
2. Denis Rothman, *"Hands-On Explainable AI (XAI) with Python"*, Packt Publishing, 2022.

VI. ELECTRONICS RESOURCES:

1. https://www.ibm.com/topics/explainable-ai
2. https://christophm.github.io/interpretable-ml-book/

VII. MATERIALS ONLINE

1. Course Outline Description
2. Lecture notes
3. PowerPoint presentation
4. Definitions and Terminology
5. Tutorial Question Bank
6. Case Studies
7. Real life Examples
8. Complex Engineering Problems
9. Tech Talk Topics
10. Concept Video Topics
11. Open-ended Exercises
12. Assignments
13. Model Question Paper – I
14. Model Question Paper – II
15. GATE Question Bank
16. Previous Question Papers and Solutions