Uka Tarsadia University

B. Tech.

CSE / CSE (CC) / CE (SE)

Semester VII

Program Elective - VI

XAI

AI6013

EFFECTIVE FROM July-2024

Syllabus version: 1.00

Subject Code	Subject Title
AI6013	XAI

Teaching Scheme				Examination Scheme				
Hours		Credits		Theory Marks		Practical Marks		Total Marks
Theory	Practical	Theory	Practical	Internal	External	Internal	External	
4	0	4	0	40	60	0	0	100

Objectives of the course:

- To familiarize concepts related to Explainable Artificial Intelligence (XAI) and interpretable methods, Students will gain a basic proficiency in interpreting and explaining the decisions of ML and AI systems.
- To understand the performance of a machine learning model and its ability to produce explainable and interpret able predictions.

Course outcomes:

Upon completion of the course, the student shall be able to,

CO1: Understand what XAI is, its scope, and impact on various domains.
CO2: Understand the methods and terminologies involved in XAI.
CO3: Differentiate the methods used in XAI and apply suitable XAI Models or approaches for given application.
CO4: Design and develop XAI use cases for real time applications.
CO5: Design of test procedures to assess the efficiency of the developed model.
CO6: Identify and evaluate the most used XAI techniques and algorithms, critically evaluate and contextualize the performance and reliability of Explanations, and identify their limitations and biases.

Sr. No.	Topics	Hours
	Unit – I	
1	**Introduction to Explainability:** Defining explainability, Overview of explanations, Known issues in explainability. **Introduction to Explainable Artificial Intelligence:** Fundamentals of XAI, Categorization of XAI, Taxonomy of XAI methods for Machine Learning , Machine Learning interpretability, Causal model induction, Causality learning, XAI techniques and limitations.	8
	Unit – II	

2	**Interpretability:** Difference between interpretability and explainability, Interpretability methods to explain Black-Box model, Scope of interpretability, Apply interpretability on Regression, Logistic regression, Generalized additive models, Decision tree, Neural network interpretation, Evaluation of interpretability.	10
	Unit – III	
3	**Deep Explanation:** Attention mechanisms, Modular networks, Feature identification, Learn to explain, Feature visualization, Deep visualization, gradcam and activation maps, Sensitivity analysis. **XAI Models:** Ante-hoc Explainability (AHE) models, Post-hoc Explainability (PHE) models, Interactive Machine Learning (IML), Black box Explanation through Transparent Approximation (BETA) models, Hybrid models.	12
	Unit – IV	
4	**XAI Methods:** XAI Techniques, Local Interpretable Model-Agnostic Explanations (LIME), Understanding mathematical representation of LIME, Shapley Additive exPlanations (SHAP), Diverse Counterfactual Explanations (DiCE), Layer wise Relevance Propagation (LRP). **Trust and acceptance:** Metrics to evaluate XAI, Trustworthy explainability acceptance, Power Quality Disturbance (PQD) classification, Methods for measuring human intelligence, Evaluating AI system.	10
	Unit – V	
5	**Unstructured Data:** Pre-model explainability on unstructured data supervised wrapper for clustering models. **Text and Image:** Lime for text data, Shape for text classifiers, Layer-wise relevance propagation, XRAI, Grad-Cam time series explainers LLM, Foundation Models, Feature selection for explainability.	12
	Unit – VI	
6	**Building Trustworthy Model with Explainable AI:** Medical diagnosis, Making AI decisions trustworthy for physicians and patients, Sales predictions on the house sale, Recent trends.	8

Text books:

1. Molnar and Christoph, "Interpretable machine learning, A Guide for Making Black Box Models Explainable", 2019.
2. Uday Kamath and John Liu, "Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning", Springer, ISBN 9783030833558.

Reference books:

1. Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen and Klaus-Robert Müller, "Explainable AI: Interpreting, Explaining and Visualizing Deep Learning", Springer Cham, 978-3-030-28954-6, 2019.
2. Michael Munn and David Pitman, "Explainable AI for Practitioners", O'Reilly Media, Inc., ISBN: 9781098119133, 2022.
3. Kamal Malik, Moolchand Sharma, Suman Deswal, Umesh Gupta, Deevyankar Agarwal, and Yahya Obaid Bakheet Al Shamsi, "Explainable Artificial Intelligence for Autonomous Vehicles - Concepts, Challenges, and Applications" CRC Press, ISBN 9781032655017.

Course objectives and Course outcomes mapping:

- To familiarize concepts related to Explainable Artificial Intelligence (XAI) and interpretable methods, Students will gain a basic proficiency in interpreting and explaining the decisions of ML and AI systems.CO1, CO2.
- To understand the performance of a machine learning model and its ability to produce explainable and interpret able predictions. CO3, CO4, CO5.

Course units and Course outcomes mapping:

Unit No.	Unit Name	Course Outcomes					
		CO1	CO2	CO3	CO4	CO5	CO6
1	Introduction to Explainability and Explainable Artificial Intelligence	✓					
2	Interpretability		✓				
3	Deep Explanation and XAI Models			✓			
4	XAI Methods, and Trust and acceptance				✓		
5	Unstructured Data, and Text and Image					✓	
6	Building Trustworthy Model with Explainable AI						✓

Programme outcomes:

PO 1: Engineering knowledge: An ability to apply knowledge of mathematics, science, and engineering.

PO 2: Problem analysis: An ability to identify, formulates, and solves engineering problems.

PO 3: Design/development of solutions: An ability to design a system, component, or process to meet desired needs within realistic constraints.

PO 4: Conduct investigations of complex problems: An ability to use the techniques, skills, and modern engineering tools necessary for solving engineering problems.

PO 5: Modern tool usage: The broad education and understanding of new engineering techniques necessary to solve engineering problems.

PO 6: The engineer and society: Achieve professional success with an understanding and appreciation of ethical behavior, social responsibility, and diversity, both as individuals and in team environments.

PO 7: Environment and sustainability: Articulate a comprehensive world view that integrates diverse approaches to sustainability.

PO 8: Ethics: Identify and demonstrate knowledge of ethical values in non-classroom activities, such as service learning, internships, and field work.

PO 9: Individual and team work: An ability to function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.

PO 10: Communication: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give/receive clear instructions.

PO 11: Project management and finance: An ability to demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

PO 12: Life-long learning: recognition of the need for, and an ability to engage in life-long learning.

Programme outcomes and Course outcomes mapping:

Programme Outcomes	Course Outcomes					
	CO1	CO2	CO3	CO4	CO5	CO6
PO1				✓		
PO2						

PO3					✓	
PO4						
PO5			✓			
PO6						✓
PO7						
PO8						
PO9						
PO10				✓		
PO11						
PO12						