CS5495: EXPLAINABLE AI

New Syllabus Proposal

Effective Term
Semester A 2025/26

Part I Course Overview

Course Title
Explainable AI

Subject Code
CS - Computer Science
Course Number
5495

Academic Unit
Computer Science (CS)

College/School
College of Computing (CC)

Course Duration
One Semester

Credit Units
3

Level
P5, P6 - Postgraduate Degree

Medium of Instruction
English

Medium of Assessment
English

Prerequisites
CS3334 Data Structures

Precursors
Nil

Equivalent Courses
Nil

Exclusive Courses
Nil

Part II Course Details

Abstract

The goal of this course is to introduce students to explainable AI (XAI) methods, which aim to explain the predictions of black-box AI models. Such explanations are important for establishing good communication, trust, clarity, and understanding of AI models, which can increase their adoption in critical systems and for solving complex problems. This course is intended to give a broad overview of different XAI methods from a practical standpoint, with a focus on applying XAI and interpreting and analyzing the AI systems. At the end of the course, students will have both working knowledge of and practical experience implementing and applying XAI on different AI models and different domains.

Course Intended Learning Outcomes (CILOs)

	CILOs	Weighting (if app.)	DEC-A1	DEC-A2	DEC-A3
1	Identify and explain common explainable AI methods.		x		
2	Implement explainable AI methods.			x	
3	Apply explainable AI methods to analyse real-world AI models.			x	
4	Evaluate the effectiveness of different explainable AI methods and discuss their advantages and disadvantages.			x	

A1: Attitude
Develop an attitude of discovery/innovation/creativity, as demonstrated by students possessing a strong sense of curiosity, asking questions actively, challenging assumptions or engaging in inquiry together with teachers.

A2: Ability
Develop the ability/skill needed to discover/innovate/create, as demonstrated by students possessing critical thinking skills to assess ideas, acquiring research skills, synthesizing knowledge across disciplines or applying academic knowledge to real-life problems.

A3: Accomplishments
Demonstrate accomplishment of discovery/innovation/creativity through producing /constructing creative works/new artefacts, effective solutions to real-life problems or new processes.

Learning and Teaching Activities (LTAs)

	LTAs	Brief Description	CILO No.	Hours/week (if applicable)
1	Lecture	Students will engage with selected XAI methods, and the intuition and principles behind them. The XAI methods will be illustrated with both toy and real-world examples to motivate the students' understanding. Implementation issues will be discussed, as well as available software toolboxes.	1, 4	2 hours

2	Tutorial	In each week's tutorial session, students will use XAI methods on small examples to gain better understanding of the lecture material.	1	1 hour
3	Assignments	Students will implement and apply XAI methods to small AI models and interpret the results. Students can then observe the effectiveness of the methods, and evaluate their differences.	2, 3, 4	1 every 4 weeks
4	Course Project	Students will implement and apply XAI methods to analyze real-world AI models. Students will report their results in a course report and during a poster/presentation session held at the end of the semester.	2, 3, 4	

Assessment Tasks / Activities (ATs)

	ATs	CILO No.	Weighting (%)	Remarks ("-" for nil entry)	Allow Use of GenAI?
1	In-class exercises	1	10	-	Yes
2	Assignments	2, 3, 4	30	-	Yes
3	Course Project	2, 3, 4	30	Students can only use GenAI for editing the English of the report, debugging code, or brainstorming. GenAI cannot be used for other aspects, e.g., writing code or analyzing experiment results. For a student to pass the course, at least 30% of the maximum mark for the examination AND course project must be obtained.	Yes

Continuous Assessment (%)

70

Examination (%)

30

Examination Duration (Hours)

2

Minimum Examination Passing Requirement (%)

30

Additional Information for ATs

For a student to pass the course, at least 30% of the maximum mark for the examination AND course project must be obtained.

Assessment Rubrics (AR)

Assessment Task

1. In-class exercises

Criterion

1.1 CAPACITY for LEARNING about explainable AI methods

Excellent

(A+, A, A-) High

Good

(B+, B, B-) Significant

Fair

(C+, C, C-) Moderate

Marginal

(D) Basic

Failure

(F) Not even reaching marginal levels

Assessment Task

2. Assignments

Criterion

2.1 ABILITY to IMPLEMENT and APPLY explainable AI to small models and INTERPRET the results
2.2 ABILITY to COMPARE the accuracy and efficiency of explainable AI methods.

Excellent

(A+, A, A-) High

Good

(B+, B, B-) Significant

Fair

(C+, C, C-) Moderate

Marginal

(D) Basic

Failure

(F) Not even reaching marginal levels

Assessment Task

3. Course Project and Presentation

Criterion

3.1 ABILITY to IMPLEMENT and APPLY explainable AI to real-world AI models and INTERPRET the results
3.2 ABILITY to EVALUATE, COMPARE, and CONTRAST different explainable AI methods.

Excellent

(A+, A, A-) High

Good

(B+, B, B-) Significant

Fair

(C+, C, C-) Moderate

Marginal

(D) Basic

Failure

(F) Not even reaching marginal levels

Assessment Task

4. Examination

Criterion

4.1 ABILITY to EXPLAIN explainable AI methods and INTERPRET their results.
4.2 ABILITY to EVALUATE, COMPARE, and CONTRAST different explainable AI methods.

Excellent

(A+, A, A-) High

Good

(B+, B, B-) Significant

Fair

(C+, C, C-) Moderate

Marginal

(D) Basic

Failure

(F) Not even reaching marginal levels

Part III Other Information

Keyword Syllabus

a. Interpretability and explainability
   i. Importance and scope
   ii. Properties and evaluation methods
b. Interpretable models
   i. Linear regression
   ii. Logistic regression
   iii. Generalized Linear Models (GLMs)
   iv. Decision Tree & Random Forest
c. Global XAI Methods
d. Local XAI Methods
   i. Local surrogate models
   ii. Counterfactual explanations
   iii. Shapley values and SHAP
   iv. Gradient-based methods
   v. Perturbation methods
e. XAI for deep neural networks (DNNs)
   i. Feature visualization
   ii. Feature attribution and saliency maps
   iii. Concept detection
   iv. Adversarial examples
   v. Data instance attribution
   vi. Mechanistic interpretability
   vii. Interpretable DNNs

Reading List

Compulsory Readings

	Title
1	Christoph Molnar, Interpretable Machine Learning: A Guide For Making Black Box Models Explainable (2nd ed.), 2022. https://christophm.github.io/interpretable-ml-book/
2	Academic publications will be provided for reading.

Additional Readings

	Title
1	Academic publications will be provided for reading.