


Course Code: CSE3023 Core Elective – 7 th Sem, CSE	Course Name: XAI: Explainable AI	 BML MUNJAL UNIVERSITY™ <small>FROM HERE TO THE WORLD</small>
Credits: 3 (L-D-P-2-0-2)	Contact Hours: 4 sessions per week <i>[Each session is of 55 minutes]</i>	
Course Faculty: Dr. Manisha Saini	Course Coordinator: Dr. Manisha Saini Email: manisha.saini@bmu.edu.in	

Aim of the course: The goal of the course is that the student develop knowledge and skills in a variety of topics in explainable AI (XAI) including: the need for and importance of explaining different AI methods, the taxonomy of XAI, and classical and well known XAI methods. The student will develop the knowledge in both theoretical and practical terms.

Course Overview and Context:

The Explainable AI (XAI) course focuses on the principles and practices essential for understanding and interpreting AI systems. It covers foundational topics in machine learning, deep learning, generative AI, and large language models, emphasizing techniques like SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), saliency maps, and attention mechanisms. Students will learn to apply these tools to make complex AI models more transparent and interpretable, addressing both ethical and social implications.

This course integrates theoretical knowledge with practical applications, providing insights into the challenges and future directions of XAI, including its application to generative models and large language models. Students will gain skills in interpreting AI decisions, understanding various XAI techniques, and communicating findings to stakeholders. Through real-world case studies and practical exercises, participants will develop the expertise needed to enhance the transparency and accountability of AI systems.

Course Outcomes:

CO1	Understand what Explainable AI is, its scope, and impact on various domains.
CO2	Identify and evaluate the most used XAI techniques and algorithms.
CO3	Develop practical skills in Python for implementing and interpreting XAI methods and Interpret the results

Course Content:

Unit:1 INTRODUCTION TO XAI

Understanding the need for explainability in AI, Importance of interpretability and explainability in AI, The trade-off between complexity and transparency. Ethical considerations in XAI, Categories of XAI methods -ante-hoc and post-hoc, Taxonomy of XAI techniques for Machine Learning and deep learning.

Unit:2 INTERPRETABILITY METHODS & MODEL-AGNOSTIC XAI

Common interpretability techniques, Local Interpretability Techniques - Generating SHAP (SHapley Additive exPlanations) values for model features, Visualizing feature importance using heatmaps and bar charts. Global interpretability methods (e.g., partial dependence plots, feature interaction analysis). Model-specific interpretability techniques, Model-Agnostic - Implementing LIME (Local Interpretable Model-Agnostic Explanations), SHAP (SHapley Additive exPlanations), Contrastive explanations, Explaining ensemble models.

Unit:3 INTERACTIVE MACHINE LEARNING TECHNIQUES & DEEP EXPLANATION TECHNIQUES

Interactive Machine Learning (IML) techniques -Building user-friendly explanation interfaces, Human-in-the-loop XAI. Neural Network Interpretability - Visualizing saliency maps using gradient-based methods, Interpreting CNNs using Class Activation Mapping (CAM), Feature visualization techniques for neural networks. Deep Explanation techniques - Attention mechanisms for interpreting neural networks, Activation maps and gradient-based approaches, Saliency maps and occlusion analysis, Explainable AI Grad-CAM.

Unit:4 POST HOC EXPLANATION APPROACHES & ETHICAL CONSIDERATIONS IN XAI

Implementing model distillation to transfer knowledge, Using SHAP to explain ensemble models, Generating counterfactual explanations for individual predictions- Exploring bias and fairness issues in XAI. Fairness and bias in explainable AI.

Unit:5 USER-CENTRIC XAI AND EVALUATION

Designing user-centric explanations for different stakeholders, Conducting user studies to evaluate explanation effectiveness. Metrics for evaluating the quality of explanations. Real-World Applications and Case Studies

Unit:6 Generative Adversarial Networks (GANs) and Generative AI (GenXAI) in Explainable AI

Real-world applications and case studies, including Generative Adversarial Networks (GANs) in XAI, Explainable Generative AI (GenXAI), providing a comprehensive overview of XAI techniques tailored to generative models and large language models (LLMs).

Course Competencies and Instruction Schedule:

	Competency	CO	No of sessions
1	Understanding the need for explainability in AI, Importance of interpretability and explainability in AI, The trade-off between complexity and transparency. Ethical considerations in XAI, Categories of XAI methods -ante-hoc and post-hoc, Taxonomy of XAI techniques for Machine Learning and deep learning.	CO1	5 sessions

2	Common interpretability techniques, Local Interpretability Techniques - Generating SHAP (SHapley Additive exPlanations) values for model features, Visualizing feature importance using heatmaps and bar charts. Global interpretability methods (e.g., partial dependence plots, feature interaction analysis).	CO1	4 sessions
3	Model-specific interpretability techniques, Model-Agnostic - Implementing LIME (Local Interpretable Model-Agnostic Explanations), SHAP (SHapley Additive exPlanations), Contrastive explanations, Explaining ensemble models.	CO2	3 sessions
4.	Interactive Machine Learning (IML) techniques -Building user-friendly explanation interfaces, Human-in-the-loop XAI. Neural Network Interpretability - Visualizing saliency maps using gradient-based methods, Interpreting CNNs using Class Activation Mapping (CAM), Feature visualization techniques for neural networks.	CO2	4 sessions
5	Deep Explanation techniques - Attention mechanisms for interpreting neural networks, Activation maps and gradient-based approaches, Saliency maps and occlusion analysis.	CO2	2 sessions
6	Implementing model distillation to transfer knowledge, Using SHAP to explain ensemble models, Generating counterfactual explanations for individual predictions- Exploring bias and fairness issues in XAI. Fairness and bias in explainable AI.	CO2	3 sessions
7	Designing user-centric explanations for different stakeholders, Conducting user studies to evaluate explanation effectiveness.	CO3	3 sessions
8	Metrics for evaluating the quality of explanations. Real-World Applications and Case Studies	CO3	3 sessions
9	Real-world applications and case studies, including Generative Adversarial Networks (GANs) in XAI	CO2	2 sessions
10	Explainable Generative AI (GenXAI), providing a comprehensive overview of XAI techniques tailored to generative models and large language models (LLMs).	CO2	3 sessions

CO/PO Mapping:

CO- PO and PSO Mapping

CO/PO Mapping	P01	PO2	PO 3	P O 4	PO 5	P O 6	PO 7	PO 8	PO 9	PO1 0	PO1 1	PO1 2	PSO 1	PSO 2	PSO 3	PSO 4
CO1			3	3	3	2				3		3	3	3		
CO2	2	3		3	3	2					1		3	3		
CO3		2	3	3	2			3	3	1		3	3	3	1	1

Experiential Learning Component:

Project as mentioned in the above table, is the experiential learning component for this course. Students will be given challenging real-life problems. They will be asked to build solutions by applying suitable learning algorithms. The students are also expected to implement and show results of the proposed solution and perform a comparative analysis with different available algorithms. A separate assessment will be conducted for evaluating the solution provided by each student. The students are also expected to implement and show the results of the proposed solution or attempt to reimplement and improve on a research paper on a topic of their choice. Approximately 60-70% is experiential learning.

Assessment Pattern: The final grade will be based on the marks/ grades obtained in the mid-semester and end-semester evaluation and other assessments defined in the assessment table. The relative grading method described in the university's academic regulations will be followed to grade the students. The student must secure a minimum of 40% of marks after completing all the assessments in the following table to become eligible for grading.

Assessment:

Component	Duration	Weightage (%)	Evaluation Week	Remarks
Programming Assignments	2 Weeks	20 %	Continuous	Participation +Viva
Quiz	20 mins	20 %	After Mid semester	Assess understanding of theoretical concepts
Mid-Semester Project Evaluation	As per the University norms	20%	Continuous	Project proposal, Literature survey, Methodology, and Preliminary Results
End Term Project Evaluation	As per the University norms	40%	During the last two weeks of the course	Project-Based

Recourse Examination Policy: In case a student fails the course, a one-time recourse is permitted as per the academic regulations of the University. Recourse is allowed **only for the End Semester examination** with 20% weightage.

Student Responsibilities:

Attend lectures and do the given Assignments as per instructions.

Attendance Policy: Students are expected to attend classes regularly. Failure to follow the classes regularly and adhere to the expected attendance percentage will result in losing quiz and other marks and a reduction of the grade as per the University's grading policy.

Learning Resources:**Text Book:**

- 1) Denis Rothman. Hands On Explainable AI (XAI) with Python. Packt 2020
- 2) Molnar, Christoph. "Interpretable machine learning". Lulu. com, 2020.
- 3) Biecek, Przemyslaw, and Tomasz Burzykowski. "Local interpretable model-agnostic explanations (LIME)." Explanatory Model Analysis Explore, Explain and Examine Predictive Models 1 (2021): 107-124.
- 4) Kleppmann, Martin. Designing data-intensive applications: The big ideas behind reliable, scalable, and maintainable systems. " O'Reilly Media, Inc.", 2017

Reference Book/Other Resources:

- 1) Molnar, Christoph. Interpretable Machine Learning. Leanpub 2019 Online version publicly available at: <https://christophm.github.io/interpretable-ml-book/scope-of-interpretability.html>
- 2) Schneider, Johannes. "Explainable Generative AI (GenXAI): A Survey, Conceptualization, and Research Agenda." arXiv preprint arXiv:2404.09554 (2024).[Research Paper]

Note: Instructors will regularly post the necessary learning resources such as lecture resources to the online course management portal i.e. Maitri/Google-classroom.

Student Responsibilities:

- Attend lectures regularly and get access to the course materials shared by the instructors.
- Check announcements at LMS/Google-classroom and emails on a regular basis.
- Submit assignments on time.
- Regularly, check your marks on the LMS and make sure they are up to date.
- You should participate in class and do whatever it takes for you to grasp this material. Never hesitate to ask questions.
- Please communicate any concerns by talking to the instructor or writing email.

Attendance Policy: Students are expected to attend the classes regularly and be present in class in time. Late coming in class is not permitted. Failure to attend the classes regularly and adhere to the expected attendance percentage will result in a reduction of the grade as per the University's grading policy.

Make-up policy: No make-up exam will be conducted for unexcused absences. The faculty needs to be informed in advance in case the student is not going to appear for any evaluation component, and it is at the discretion of the faculty to sanction makeup for an evaluation component.

Behavior Expectations: No mobile phones and other destructive gadgets are permitted in the class.

Academic Dishonesty/Cheating/Plagiarism: Plagiarism and dishonesty in any form in any evaluation component will lead to appropriate disciplinary action.