Assignment: Explainable AI

Objective

The aim of this assignment is to develop advanced skills in critically assessing, improving, and applying explainable AI methods. Students may choose **one of two tasks**, reflecting different research-oriented approaches:

1. Identifying and addressing **limitations of an XAI method**
2. Replicating and critically analysing an **XAI research paper**

Task 1: Fix a Limitation of an XAI Method

You will identify, demonstrate, and propose improvements for a limitation of a chosen XAI method.

Steps:

1. Choose a specific XAI method (e.g., LIME, SHAP, Integrated Gradients, Anchors, Counterfactual Explanations, etc.).
2. Identify a limitation of the chosen method (e.g., instability, lack of faithfulness, computational cost, lack of causality, poor handling of time series/structured data).
3. Select a dataset of your choice and demonstrate the limitation through experiments (this can be both qualitative or quantitative depending on the limitation you identify).
4. Propose a solution:
    - If feasible, implement and evaluate your proposed fix.
    - If implementation is too complex, provide a conceptual solution supported by convincing arguments and references, showing feasibility.
5. Critically assess whether your solution improves the utility of the XAI method identified and outline open challenges.

Deliverables:

- Report (max. 10 pages, excluding references and appendix) containing:
    - Introduction of the method and limitation
    - Experimental demonstration
    - Proposed solution (conceptual/implemented)
    - Evaluation and critical reflection
- Implementation package, including:
    - Source code (well-documented, reproducible)
    - Dataset(s) used (or clear instructions to access them if restricted)
    - Scripts/notebooks for running experiments

Task 2: Replication of an XAI Paper
You will replicate and critically analyze a research paper in XAI.

Steps:

1. Identify a paper in the XAI field that you would like to replicate. The paper must be approved by the lecturers (Branka Hadji Mishvea, Lucia Gomez and Faizan Ahmed) before proceeding.
2. Implement the methods and reproduce results (figures, tables, metrics). Partial replication is acceptable if constraints exist (e.g., dataset access, computation).
3. Compare your findings to the original results. Discuss reproducibility challenges, limitations of the approach, and robustness of the method.
4. Extend the paper's either by:
   - Conceptual extension: Provide a well-argued discussion of how the methodology could be extended or improved (e.g., addressing a limitation, adapting to a new type of data, or incorporating additional techniques).
   - Practical extension (if feasible): Implement and test the approach on a new dataset, in a different domain, or with a variant of the original algorithm. Provide results and compare them to the original findings.

Deliverables:

- Report (max. 10 pages, excluding references and appendix) containing:
  - Summary of the original paper (problem, method, contributions)
  - Replication methodology and results
  - Comparative analysis (original vs. replication)
  - Critical reflection on reproducibility and robustness
- Implementation package, including:
  - Source code (well-documented, reproducible)
  - Dataset(s) used (or instructions to access them if not openly available)
  - Scripts/notebooks to reproduce results and extensions

General Requirements

- **Programming & Tools**: Use Python (preferred). Other languages allowed.
- **Reproducibility**: Submit code and documentation (using our Git & Quantlet).
- **Academic Integrity**: Proper referencing, originality, and adherence to scientific standards are mandatory.

Qualitative Performance Descriptors

Pass / Fail Decision

Fail: The submission does not meet the minimum standards in one or more of the core criteria (relevance, technical work, critical thinking, or presentation). Typical fail reasons: lack of reproducibility, missing deliverables, plagiarism, superficial treatment of the task, or failure to demonstrate understanding of XAI.

Pass: The submission meets the minimum requirements across all criteria and demonstrates an acceptable level of relevance, technical correctness, reflection, and presentation.

1. **Relevance & Rigor** *(formerly 30%)*

Fail: Inappropriate or unclear choice of method/paper; research question not meaningful or aligned with XAI.

Pass – Poor: Method/paper is acceptable but only superficially justified; limited understanding of its scope.

Pass – Good: Clear and appropriate choice; demonstrates sound understanding of context and relevance; research question well framed.

Pass – Excellent: Choice is highly relevant, original, and insightful; clear articulation of research gap; shows comprehensive knowledge of XAI landscape.

2. **Technical Work** *(formerly 30%)*

Fail: Experiments, code, or replication attempt missing, non-functional, or plagiarized; results cannot be reproduced.

Pass – Poor: Code is minimally functional but incomplete; experiments shallow or error-prone; partial reproducibility.

Pass – Good: Solid implementation with working experiments; results largely reproducible; technical competence clearly demonstrated.

Pass – Excellent: High-quality, well-documented, reproducible implementation; creative or challenging extensions; demonstrates mastery of tools and methods.

3. **Critical Thinking** *(formerly 25%)*

Fail: No meaningful reflection; limitations ignored; purely descriptive without analysis.

Pass – Poor: Some reflection present but shallow; limitations mentioned but not explored; limited originality.

Pass – Good: Clear identification of limitations; balanced and well-argued critical reflection; shows understanding of trade-offs and challenges.

Pass – Excellent: Insightful and original critique; strong connection between evidence and argument; proposes thoughtful and feasible improvements/extensions.

4. **Presentation & Reproducibility** *(formerly 15%)*

Fail: Report is missing, incoherent, plagiarized, or lacks required deliverables (code, data, documentation).

Pass – Poor: Report delivered but poorly structured; language and formatting hinder clarity; limited reproducibility.

Pass – Good: Report is clear, well-structured, and academically sound; reproducibility ensured with code and data; presentation coherent.

Pass – Excellent: Report is professional, concise, and engaging; excellent writing style; reproducibility fully ensured; presentation highly effective.

Deadlines

Due dates	**Assignment**	**Content**
09-10-2025	Project topic hand-out	Choose a project and inform lecturers
02-02-2026	Hand-in of report & supporting materials	
12-02-2026	Presentation of work	20-minute presentation +10 minutes Q&A

The lecturer team is available for intermediate coaching or alignment sessions. These can be scheduled individually directly via email.