Final project
Explainable AI

Summary: Write a report about other types of explanation methods (from a list of options) and write code for small demonstration on a domain/task (e.g. healthcare).

1) Report (80% points)
The report should consider different aspects of the explanation methods:

- Choose at least 3 papers on the chosen type of explanation.
- Describe the *motivation and context* for using such type of explanations
- Propose a *research question* that will look into a specific point of interest.
- *Describe* at least two explanation methods of this type.
- Describe how these methods can be *evaluated*.
- Discuss *pitfalls*.
- Propose a use case in a specific domain/task (see below for more details).

2) Code demonstration (20% points)
Show how the explanation methods work in practice. You are free to use existing libraries (python).

- Choose a dataset related to the use case.
- Show the use case results.

Total word count: 1400 words (+/- 10%), excluding figures, tables and code.

Notes:

- *At least one explanation method should not have been covered in the lectures.*
- *To describe an XAI method, use a level of detail similar to that of the lectures (or higher, if you prefer).*

3) Deadline: Submit your report on **04/04/2025** on Brightspace (Activities > Assignments).

Please see below a list of possible topics and the grading scheme.

General topics and examples

1. **Explanations for Large language models**
   a. Compare attention-based explanations (e.g., attention maps) vs. post-hoc methods (e.g., SHAP, LIME).
   b. Evaluate how well explanations align with human intuition.
   c. Demonstrate explanations for biased outputs.

2. **Interactive explanations**
   a. Explore how users can interact with explanations (e.g., counterfactual explanations, explorable explanations).
   b. Investigate how different user interfaces affect trust in AI.
   c. Implement a simple interactive visualization of explanations.

3. **Explanations for Time series**
   a. Compare attribution methods (e.g., Integrated Gradients, LRP) for time series forecasting.
   b. Analyze which features drive predictions in models like LSTMs, Transformers.
   c. Implement explanations for a financial dataset.

4. **Evaluation of explanations**
   a. Discuss different evaluation metrics: faithfulness, stability, completeness.

b. Compare human-centered vs. quantitative evaluation methods.
c. Implement an experiment testing explanation quality.

5. XAI for uncertainty and anomaly detection
a. Explaining Uncertainty in Bayesian Neural Networks – Why does an AI model express low or high confidence?
b. Interpretable Anomaly Detection – How explainability helps detect fraudulent transactions or cyberattacks.

6. Robustness, fairness, and bias in XAI
a. Robustness of Explanations – Do explanations change under small perturbations in input data?
b. Bias Detection through XAI – How explainability methods can reveal bias in AI models.
c. Fairness vs. Transparency Trade-off – Does making a model interpretable reduce accuracy or fairness?

7. **Model-specific explainability** (e.g XAI for CNNs, in XLR, GNNs, etc).

8. Different topic in consultation with the course coordinator.

Grading Rubric for XAI research paper

Mandatory Requirements (Fail if missing, 0 points)

	FAIL	**PASS**
References	Less than 3 references.	At least 3 references included.
Sections	2 or more sections missing (Introduction, Methods, Results, Discussion)	Introduction, Research Question, Methods, Results, and Discussion all present.
XAI Methods	Less than 2 methods OR both from lectures.	2 XAI methods included, one novel (not covered in lectures).
Code Availability	Code missing, inaccessible, or not cited.	Code uploaded to GitHub/GitLab, accessible, and cited in the paper.
Word count	Text (excluding references, figures, code) doesn't align with the 1400 (+/- 10%) word count	Text (excluding references, figures, code) aligns with the 1400 (+/- 10%) word count

Sections	**What's evaluated?**	**Grading Criteria**			
		Not Satisfactory	**Sufficient**	**Good**	**Excellent**
Introduction (0-15 points)	**Understanding & Integration**	0 No clear integration of literature.	2-4 Literature mentioned but messy, unclear arguments.	5-6 Good synthesis of ideas, logical connections.	7.5 Excellent integration, strong argument flow
	Critical Analysis	0 No analysis of literature gaps/limitations.	2-4 Some critical points raised but shallow.	5-6 Discusses strengths/ weaknesses clearly.	7.5 In-depth, insightful critique of literature, finishing up with the research gap that tights to your RQ
Research Question & Motivation (0-5 points)		0 No clear research question	1-3 Research question present but lacks specificity.	4 Clearly defined, domain-specific RQ, but not well-justified	5 Clearly defined, domain-specific RQ, and well-justified.
Methods (0-20 points)	**XAI Method Descriptions**	0 Less than 2 methods or descriptions are unclear.	2-4 Two methods described, but explanations lack clarity or depth.	5-6 Good, detailed explanation of two methods, including	7.5 Excellent, detailed explanation of two methods, including technical details.

				technical details. Little inaccuracy.	
	Comparison & Analysis	0 No discussion of advantages/limitations.	2-4 Some advantages / limitations discussed but superficial.	5-6 Good theoretical comparison, strengths & weaknesses outlined.	7.5 Strong theoretical comparison, strengths & weaknesses clearly outlined.
	Evaluation methodologies	0 No evaluation process mentioned.	1-3 Evaluation mentioned but lacks detail.	4 Clear, well-explained evaluation of XAI methods.	5 Detailed explanation of evaluation of XAI methods with respect to the two methods.
Code Documentation (5 points)	**Code Explanation**	0 No description of the code.	1-3 Attempted to describe code, but explanations are incomplete and unclear.	4 Clear explanation of methodology implementation, and rationale.	5 Excellent explanation of methodology, implementation, and rationale.
Results & Discussion (10 points)	**Results and Analysis**	0 No evaluation results discussed.	1-3 Some results mentioned, lacks interpretation.	4 Good evaluation and discussion of results.	5 Excellent evaluation, insightful discussion of results.
	Pitfalls & Biases	0 No discussion of pitfalls/biases.	1-3 Some pitfalls mentioned but lacks depth/accuracy	4 Identifies key pitfalls.	5 Identifies key pitfalls, and provides ethical considerations.
Use Case Discussion (20 points)	**Domain-Specific Application**	0 No real-world application provided.	6 Domain mentioned but lacks clear explanation.	8 Good argumentation and real-world application but could be improved	10 Easy to follow and relevant arguments, excellent real-world application and consideration of target population.
	Explanation Relevance	0 No discussion on how methods apply to the domain.	6 Some examples provided but unclear.	8 Relevant examples demonstrating XAI methods in the domain.	10 Excellent and clear/relevant examples demonstrating XAI methods in the domain.
Structure, Clarity, and Writing (10 points)	**Logical Flow & Structure**	0 Disorganized, very difficult to follow.	1 At times disorganized, difficult to follow.	2 Some structure but weak transitions.	3 Well-organized, clear transitions between sections.
	Writing Quality	0 Poor grammar, spelling, or citation formatting.	1 Some errors in writing or formatting.	2 Well-written, minimal errors, correct citations.	2 Well-written, minimal errors, correct citations.
Code Quality & Reproducibility (15 points)	**Readability & Correctness**	0 Poorly commented, unclear function naming.	5 Some comments, but readability issues.	8 Well-documented, clear comments & structure. A few mistakes	10 Excellent, clear comments and structure, no mistakes
	Reproducibility	0 No instructions for running the code.	1-3 Instructions provided but unclear/incomplete.	4 Clear README with installation steps & usage guide. Struggled to run.	5 Clear README with installation steps & usage guide. Easy to run.

Evaluation

Research paper		/80
Code & description		**/20**

Total Score		**/100**