CS 671 Interpretable and Explainable AI
Course for Graduate students, Spring 2026
Instructor: Lesia Semenova

Quick details

Monday 12.10 – 3.10pm (we will have a break during this time). Busch Campus, SEC-210

Overview and schedule.
Please see schedule and intro here: https://lesiasemenova.github.io/ixai_spring2026/

Seminar Paper Preference Form. Please fill out this table by February 2:

https://docs.google.com/spreadsheets/d/15PCiTQdHR_-3eL2EI3q6GcpGyKIsUsf19tl143u8XaM/edit?usp=sharing

Please enter your name, email, and NetID and indicate your preferences for presenting seminar papers.

- Use **1 (highest interest)** through **4 (lowest interest)** to rank topics you would like to present.
- You may assign the **same ranking to multiple topics** (ties are allowed).
- You may rank **as many or as few topics as you like**.
- If you know you **cannot attend a particular week**, please enter **NA** for that week.
- You **must select at least 4 weeks** (rank 1 through 4).

I will use these preferences to assign paper presentations and will make every effort to accommodate availability and interests.

Office hours. After class and during class break. I'm usually happy to stay for ~20 minutes after class for quick questions or to discuss project ideas, paper selection, or course logistics. For longer discussions, please follow up by email.

Prerequisites. This is a graduate-level course. Students should be comfortable with basic linear algebra, probability, and core machine learning concepts. You should also be able to implement and debug ML experiments in Python (e.g., NumPy, scikit-learn, and PyTorch), including working with real datasets and writing reasonably organized, reproducible code.

Contact.
For official or sensitive communication: lesia.semenova@rutgers.edu
For projects coordination and team discussions (informal communication) we will use slack.

Grading and Requirements

The class includes two types of assignments: paper presentation and research project.

Paper presentation: you are expected to present one or two papers, either individually or in a team of two, during one of the classes. Each paper discussion will last 35 minutes and include:

1. The presentation slides – 20 minutes
2. The Q&A session and paper discussion – 15 minutes

All slides that you produce for the class, including paper presentations and project presentations, will be submitted as your homework and milestone assignments.

The paper presentation should include:

1. **Paper motivation** – why the paper is important and what problem it aims to address. If the paper includes a running example or use case, it may be helpful to explain the work using it.
2. **Paper contributions** – the core result that the paper presents (e.g., a theorem, an empirical observation, or a method).
3. **Supporting evidence for the key result** – if the contribution is a proof, what are the key steps? If it is empirical, what experiments were performed and what questions do they answer? If applicable, what social arguments or human studies were conducted?
4. **Implications of the work** – where can this work be useful, and how can it be applied?
5. **Discussion points** – please include at least 3–4 discussion points, including but not limited to:
   a. Something particularly interesting or surprising you learned from the paper
   b. Possible downsides, limitations, edge cases, or unforeseen issues
   c. Potential future extensions or additional applications

Class project. This is an almost semester-long project that aims to answer an interesting research question related to interpretability or explainability. Project directions may vary and can include, but are not limited to: designing a new method; evaluating existing methods by proposing a benchmark; applying interpretability algorithms to new problems or domains; proving interesting properties of models; studying trustworthy properties under model multiplicity; connecting interpretability or explainability with other trustworthy AI/ML research; analyzing new datasets; or developing new tools for data visualization or feature importance. On February 9, after the overview of project topics, we will discuss potential project ideas.

There will be three milestones for the project:

1. **Project proposal (5–10 min)** – describe your project idea, motivation, planned tools or methods, what you expect to learn, the research question you are addressing, and how your project sits within the existing literature.
2. **Midpoint check-in (5–10 min)** – present project updates, including initial results, lessons learned, successes and challenges. Indicate whether you changed direction and what you plan to do next. The slides should include a link to your GitHub repository.

3. **Final presentation (10–20 min)** – present your project, including clear motivation, main results, methods, observations, experiments, and lessons learned. The slides should include a link to your GitHub repository.

What makes a strong project?
Strong projects ask a clear research question, engage meaningfully with prior work, and demonstrate careful reasoning and experimentation. Novelty is not required; depth, clarity, and critical thinking are valued. Negative results and well-executed evaluations are welcome.

You may work individually or in a team of two or three. Teams of four or more are discouraged. If your work results in an interesting finding, I will help with framing the results for a potential publication.

Here is how you will be evaluated:
Paper presentations – 30%
Project – 50%
Engagement during the class – 20%

Engagement includes asking questions, contributing to discussions, and providing thoughtful feedback during paper presentations and project updates.

Logistics

Syllabus Changes. The schedule, readings, and assignments may be adjusted during the semester to better match class interests, pacing, or guest availability.

Attendance. Regular attendance is expected, as the course is discussion-based.

Presentations. For paper and project presentations, all team members are expected to present.

Collaboration. Collaboration is encouraged within project teams. Any external assistance, tools, or prior work that meaningfully influenced your project should be clearly acknowledged.

Absence. If you need to miss a class or scheduled presentation due to illness or an emergency, please email me as soon as possible. For paper presentations, students are encouraged to arrange a swap with another presenter when feasible and let me know once the swap is agreed upon. If arranging a swap is not possible, I will work with you to find an alternative.

Computing Resources: Google Colab, iLab at Rutgers.

Academic Integrity. Students are expected to follow Rutgers' Academic Integrity Policy. All submitted work must be your own (or your team's) and properly acknowledge sources. Violations will be handled according to university policy.

Accessibility and Accommodations. Students with disabilities or other circumstances requiring accommodations should contact the Office of Disability Services and notify me as early as possible so appropriate arrangements can be made.

Student Well-Being. Graduate school can be demanding. If you are experiencing stress or personal difficulties that affect your participation, please reach out early. Rutgers offers counseling and mental health resources for students.

Mandatory Reporting and Support Resources. As an instructor, I am a mandatory reporter under Rutgers policy. This means that if you disclose to me information about sexual misconduct, sexual assault, dating or domestic violence, or stalking, I am required to share that information with the appropriate university office so that support resources can be offered.

If you would like to speak with someone confidentially, Rutgers provides confidential resources such as Counseling Services, CAPS, and other designated confidential advocates. If you have questions about reporting options or available resources, I can help direct you to the appropriate offices.

Recording and Privacy. Students may not record or redistribute class sessions, presentations, or discussion materials without permission.

LLM-use

LLMs may be used in this course as **assistive tools**, but not as substitutes for understanding, experimentation, or original reasoning. LLMs are known to "hallucinate" or generate technically incorrect information, including fabricated citations, flawed mathematical proofs, and non-functional code. You are responsible for the correctness, clarity, and substance of all submitted and presented work.

Allowed uses include:

- clarifying concepts or terminology
- debugging code you are running yourself
- improving writing clarity and organization
- helping with quick experiment implementation (which you then run and verify)
- checking for obvious errors in math or notation
- exploring alternative formulations
- brainstorming project ideas (final choice, framing, and execution must be yours)

Not allowed:

- using LLMs to generate core content that you do not understand or could not reproduce independently
- submitting images, plots, diagrams, or figures generated directly by an LLM (or other generative tools**)** instead of results produced by your own code or analysis
- submitting results, explanations, or analyses that you cannot explain or defend if asked

If you are unsure whether a particular use of an LLM is appropriate, ask.

The standard we will use is simple: **you should understand and be able to reproduce everything you submit.**