CS532-T Topics in AI: Interpretability and Explainability

Instructor: Xin Tang (xin.tang@msl.ubc.ca)
Teaching Assistant: Alex He-Mo (alexhemo@student.ubc.ca)
Meeting Time: Wednesday & Friday 11: 00 - 12: 30
Office Hour: Wednesday (Alex) & Friday (Xin) 12: 30 - 13: 30
Location: CEME 1212

Note: The course design was inspired by Harvard CS282BR. We thank Drs. Hima Lakkaraju, Jiaqi Ma, and Suraj Srinivas for sharing their materials.

1.Overview

As machine-learning models are increasingly deployed to support decision-makers in high-stakes domains such as biomedicine, healthcare, finance, and law, it becomes critical that those decision-makers understand—or at least trust—the functionality of the models that guide them. This graduate-level course therefore immerses students in the fast-moving field of **explainable and interpretable machine learning (XAI)**.

Over the term we will trace the intellectual foundations of XAI by examining seminal position papers, while simultaneously probing what *interpretability* and *explainability* means to different stakeholder groups - from biologists and clinicians seeking actionable explanations to ML engineers debugging their systems. We will then traverse the major families of interpretable models and explanation techniques, including prototype-based reasoning, sparse linear surrogates, rule-learning approaches, saliency maps, generalised additive models, and counter-factual analyses. Along the way we will also investigate how interpretability interacts with allied desiderata such as fairness, robustness, and privacy, and we will grapple with the distinctive opportunities and challenges involved in demystifying foundation models like large language models and diffusion models.

The course blends lectures by the instructor, student paper presentations, and guest talks by researchers whose work has defined the area. A semester-long **team project** gives students the chance either to pursue novel research ideas *or* to carry out a rigorous benchmarking/reproducibility study, thereby contributing new knowledge and practical guidance to the community.

2. Prerequisites

Students are expected to be comfortable with linear algebra, probability, algorithms, and machine learning at roughly the level of CS340 and, ideally, CS440. Proficiency in Python (including numpy/pandas/sklearn/torch/tensorflow) and basic software-engineering practices is assumed.

3 Format, Assignments, and Assessment

The course will be graded based on the following components:

Component	Weight
Class participation	**10 %**
Paper presentations	**30 %**
Semester-long project (research *or* benchmarking/reproducibility)	**60 %**

Late-Day Policy: Each student/team has **four** late days usable on any deliverable except the final report/presentation. Additional late submissions incur a 10 percentage-points penalty per calendar day.

Remote Policy: Students are required to attend class in person unless they provide either 1) a signed note from a licensed medical doctor or 2) prior approval from the departmental office.

Communication Policy: Students are encouraged to post their questions on the Canvas Discussion tab so that everyone can benefit from the answers. Other students may have similar questions, and this helps ensure the information is shared with the whole class.

GenAI Policy: LLM such as ChatGPT is allowed. However, students are fully responsible for the accuracy, originality, and integrity of any work they submit or present, regardless of whether AI tools were used in its creation.

3.1 Paper Presentations & Class Discussion (30 %)

Because interpretability in machine learning is a swiftly evolving field, a substantial portion of our time together will be devoted to close reading and discussion of original research. In the opening weeks the instructor and teaching staff will survey foundational work; later, guest lecturers who authored seminal papers will share their perspectives. Students will then sign up—first come, first served—to present papers and guide class dialogue. We encourage presenters to work in teams of two or three, and each team will analyse two papers. A typical presentation will last roughly twenty-five to thirty minutes, followed by five minutes of questions.

When preparing, focus on four questions:

(a) **Context**: how does the paper fit into the broader literature?

(b) **Contribution**: what are its principal contributions?

(c) **Evidence**: how do the theoretical and empirical analyses substantiate those claims?

(d) **Limitations**: where are its weaknesses and how might they be repaired?

All students should at least skim the assigned reading in advance so that discussion can be lively and substantive. Participation, meaning both attendance and thoughtful engagement, will therefore be noted.

3.2 Team Project and Checkpoints (60 %)

Each student will join a team of 4–6 members; any team smaller or larger than this range must be discussed with, and approved by, the course instructor. Each team should pursue one of two project tracks for the remaining sixty percent of the course grade.

Option 1: Research track. Teams frame an original question whose answer would constitute a genuine contribution to the body of knowledge on interpretability or explainability. The desired end-state is a study strong enough—whether or not it ultimately "works"—to be polished into a conference submission.

Option 2: Benchmarking / Reproducibility / Review track. Teams select an influential stream of work and subject it to rigorous scrutiny: re-implementing baseline code, curating data, designing diagnostic experiments, analysing failure modes, or synthesising findings into a coherent perspective. A successful effort will leave the community with a reliable benchmark, a transparent reproduction, or a definitive critical survey.

Regardless of track, assessment will emphasise clarity of thought, methodological rigour, and intellectual honesty rather than purely positive empirical results. Project work unfolds through three milestones and a capstone deliverable:

Checkpoint 1 (Proposal, due near the end of Week 3, 10%). Submit a two-page prospectus that states the precise research question or benchmarking goal, situates it in related literature, explains why existing methods fall short, sketches a plan of attack, and lists concrete criteria by which success or insight will be judged. The staff will return written feedback within a week; incorporate that guidance promptly, because subsequent milestones assume the revised plan is in force.

Checkpoint 2 (Baseline implementation and critique, due in Week 7, 10%). Choose one representative paper from the syllabus (or another approved source), reproduce its principal experiment in code, and add well-commented notebooks plus a concise written critique (roughly three pages). Describe what reproduced exactly, where replication diverged from the original, and which design choices appear fragile. For the review track this component may instead take the form of a curated corpus of results with unified metrics.

Checkpoint 3 (Mid-term progress report, due in Week 11, 10%). Deliver a three-to-four-page update that contains (i) a refined formal problem statement; (ii) a detailed account of your current methodology, including algorithms, data processing, and evaluation protocols; (iii) preliminary findings—positive or negative—together with diagnostic visualisations or tables; and (iv) a candid assessment of remaining risks and a timetable for completion. Lack of empirical results at this stage will incur a penalty unless accompanied by a well-argued explanation of unforeseen obstacles and a credible mitigation plan.

Final report and presentation (during the exam period, 20% and 10%, respectively). Submit a polished manuscript of five to six pages that integrates prior checkpoints into a coherent narrative: motivation, related work, theoretical or architectural underpinnings, experimental design, results, analysis, and conclusions. Write with the precision and structure expected by a top-tier ML venue; include links to a public repository that hosts executable code,

data-processing scripts, and a README that allows any classmate to reproduce the core tables and figures in one command. Each team will also give a fifteen- to twenty-minute oral presentation, followed by five minutes of questions, showcasing the project's journey, insights, and broader significance.

Assignment	**Weightage**	**Released on**	**Due on**	**Grades Released By**
Paper Presentations	30%	Sep 15th, 5pm PT *first come, first serve*		Dec 10th, 11.59pm PT
Checkpoint 1 (Proposal)	10%	Sep 24th, 5pm PT	Oct 1st, 11.59pm PT	Oct 8th, 11.59pm PT
Checkpoint 2 (Baseline Implementation)	10%	Oct 8th, 5pm PT	Oct 22nd, 11.59pm PT	Oct 29th, 11.59pm PT
Checkpoint 3 (Midterm Progress)	10%	Oct 29th, 5pm PT	Nov 14th, 11.59pm PT	Nov 21th, 11.59pm PT
Participation Feedback	10%			Dec 10th, 11.59pm PT
Project Presentation	10%		Dec 5th, 11am to 1pm PT	Dec 10th, 11.59pm PT
Project Final Report	20%	Nov 21st, 5pm PT	Dec 12th, 11.59pm PT	Dec 19th, 11.59pm PT

This syllabus is subject to minor adjustments; any changes will be announced in class and on course Slack.