Overview

Throughout the course, you will complete a project where you will conduct original research with a focus on Machine Learning for Health (ML4Health). The readings and discussions throughout the course will help you with your project direction. We hope you pick a topic that is interesting, novel, and that motivates you. This should be fun!

The final project will be worth **70% of the course grade.**
You may work in groups of **up to 5 students** for this project. You must **work in a group of at least 2.** Reach out to the instructor or TAs if you have any issues finding a group.

Instruction

Your project must involve helping with a concrete healthcare problem through analysis of a dataset, please consult with the instructor/TA if you have questions regarding your project proposal.

To access papers/journals you can use the following resources:

- Passkey
- https://www.library.cornell.edu/

Data sets

Please find an existing dataset, or datasets, and conduct an analysis using the dataset to solve a health related problem (you can use lectures from class for motivation). Make sure your analysis is novel!

There are existing datasets/databases online that may be helpful for your project.

- MIMIC: https://mimic.mit.edu/
- eICU: https://eicu-crd.mit.edu/
- PPMI: https://www.ppmi-info.org/access-data-specimens/data
- iBKH: https://github.com/wcm-wanglab/iBKH
- StudentLife: https://studentlife.cs.dartmouth.edu/dataset.html
- CrossCheck: https://www.kaggle.com/dartweichen/crosscheck
- GDC: https://portal.gdc.cancer.gov/
- GEO: https://www.ncbi.nlm.nih.gov/geo/
- Single Cell Portal: https://singlecell.broadinstitute.org/single_cell#
- Kaggle: https://www.kaggle.com/

Many of these datasets **require you to submit an application for access**. Please start your project as early as possible if the dataset you intend to use requires you to apply for access.

Novelty

You must conduct a literature review to understand the state-of-the-art research relevant to your work. Reviews are conducted by searching databases using pre-specified keywords and inclusion criteria, and then providing an overview of the matching literature. You must clearly specify the novelty of your project. Some examples:

- Design a novel machine learning model to solve an existing healthcare problem and show your model can perform better than existing approaches you have reviewed;
- Apply an existing model to solve a new healthcare problem where machine learning algorithms have not been applied much so far and you demonstrate the potential there (for example, https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2780274);
- Perform comparative effectiveness type of research to compare different solutions and discuss future directions.

Deliverables

Within **all deliverables**, you should include in-text citations to academic papers, and have a reference section formally listing all in-text citations at the end of the deliverable. If you are not sure how to formally cite work, please refer to the papers linked above as examples. We ask that you follow the ACM citation style and reference formatting guidelines. You can use a reference management software, such as Mendeley or Zotero, if you'd like.

Please make sure to cover all questions and bullet points below. You will be graded directly on how well you answered the specific questions/covered the criteria. Rubrics will be posted prior to the assignment submissions so you understand how each of the individual bullet points affects your grade.

You should submit one Proposal, Final Report, and Video for your group, but each individual student will be asked to conduct peer reviews. The deliverables should all be submitted **in PDF format** (except for the final submission video). Please **do not use a font smaller than 11pt**.

Proposal: 15% (10% proposal+ 5% presentation)

Due date: 2/22/2024, 11:59PM ET

Please make sure you have secured access for the data sets you will be working on

Length: 1000 words (excluding figures, tables, and references)

- Introduction
  - What problem are you going to solve?
  - Why is this problem important?

    - Why is machine learning promising?
- Methods
    - How are you going to solve this problem?
    - Please state how your data set is made available to you and show some summary statistics of your data
- Potential roadblocks and resolutions
    - What potential roadblocks might you encounter?
    - How do you plan to resolve these roadblocks?
    - Are there any specific questions or topics you would like to discuss with the instructor/TAs, that would help with your project?

Intermediate Report Submission (Report Only 15%)

Report due: 3/28/24, 11:59PM ET

This report can be built upon your initial proposal

Length: 2000 words (excluding figures, tables, and references)

- Introduction
    - What problem are you going to solve?
    - Why is this problem important?
    - Why is machine learning promising?
- Related work
    - What similar research (academic publications) has already been published on this topic?
    - What are the current gaps?
- Methods
    - How are you going to solve this problem?
    - How your research is novel
    - Data set introduction
    - Experiments setup
- Results
    - Preliminary results
- Discussions
    - What are the challenges so far?
    - What are the remaining tasks?
    - Are there any specific questions or topics you would like to discuss with the instructor/TAs, that would help with your project?

Final Submission 40% (Presentation 5%, Code 15%, Report 20%)

Presentation date: 5/2/2023 and 5/7/2023, In Class

Report & code due date: 5/14/2023, 11:59PM ET

Report

This report can be built upon your intermediate report

Length: 4000 words (excluding figures, tables, and references)

- Introduction
  - What problem are you going to solve?
  - Why is this problem important?
  - Why is machine learning promising?
  - What are the major contributions of your work?
- Related work
  - What similar research (academic publications) has already been published on this topic?
  - What are the current gaps?
  - How your proposed study can fill in those gaps?
- Methods. A reader should be able to replicate your work after reading the Methods.
  - How are you going to solve this problem?
  - How your research is novel
  - Data set introduction
  - Experiments setup
    - Inputs/outputs
    - Data preprocessing
    - Model selection
    - Model evaluation
    - Model interpretation
- Results
  - Provide details of your major findings.
  - Support these findings with tables/figures when necessary. A reader who is skimming your work should be able to understand your major findings by exclusively reading the tables/figures alone.
- Discussion
  - Please give a recap of your major contributions
  - What worked/did not work about your research?
  - Are there any nuances in the methodologies (eg, algorithms) you used, based upon your results?
  - Are there implications of your work for technology researchers or clinicians?
  - Are there ethics and privacy implications?
  - What are future research directions, based upon your contributions?

Code

Please send in your source codes with data as a zip file (or made them available on Github)

- There must be a readme file introducing all the files
- There must be executable programs that can generate all results in your final reports
- Please provide clear annotations within your code, especially the utility of each function and which functions are used to generate which part of results in your final report

Presentation

Presentation Length: 8 minutes

Please make sure your presentation answers the following questions:

- What prior research, idea, or innovation enabled your project? What future research directions are promising based upon your results?
- How can the ideas proposed by this research be used in the real world? What might the barriers to adoption be?
- How might this research help address gaps in other solutions or research you have seen in this space?