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 **40% 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.

Project Categories

Please choose a project category below. If you are interested in a different category, please consult with the instructor/TA *prior* to writing your project proposal.

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

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

Data analysis

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!

Example papers:

- https://academic.oup.com/brain/article/140/7/1959/3855005
- https://www.nature.com/articles/s41591-018-0213-5
- https://pubs.acs.org/doi/abs/10.1021/acs.jproteome.0c00316

Model development

Design a novel machine learning model and/or algorithm motivated by a particular health dataset or problem. Make sure your model is novel! Novelty = more than adding additional layers to a neural network. The motivation for the new model should be clearly defined by the health related problem you are trying to solve.

Example papers:

- https://dl.acm.org/doi/abs/10.1145/3097983.3097997

- https://www.sciencedirect.com/science/article/pii/S1532046419302813
- https://proceedings.neurips.cc/paper/2018/file/8d34201a5b85900908db6cae92723617-Paper.pdf

Literature review

Literature reviews provide a survey of previous work on a specific topic, and often provide authors' perspectives on the direction of the field. Reviews are conducted by searching databases using pre-specified keywords and inclusion criteria, and then providing an overview of the matching literature. Literature reviews for ML4Health could be conducted on a specific type of methodology (eg, ML algorithms), disease area (eg, depression), or subject area (eg, ethics and privacy).

Example papers:

- https://www.nature.com/articles/s42256-021-00373-4
- https://www.nature.com/articles/s41398-020-0780-3
- https://academic.oup.com/bib/article/21/3/919/5498046?login=true

For the difference between systematic review and scoping review, please refer to this paper. Either a systematic review or scoping review can be provided for the final report: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-018-0611-x

Perspectives

Perspectives present a novel understanding or concern regarding existing machine learning models for particular health related problems or scenarios. Note here the perspectives should *not* be a short viewpoint. We expect longer perspectives with backup evidence from data analysis or published papers.

Example papers:

- https://www.nature.com/articles/s42256-019-0048-x
- https://www.sciencedirect.com/science/article/pii/S2589750020301928
- https://www.nature.com/articles/s42256-021-00390-3

Potential Datasets

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#

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

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

Percentage of final grade: 10%

Due date: 3/21/2023, 11:59PM ET

Max Length: One page (excluding figures, tables, and references)

- Introduction
    - What problem are you going to solve?
    - Why is this problem important and novel?
    - What similar research (academic publications) has already been published on this topic?
- Methods
    - How are you going to solve this problem?
- 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?

Final Submission (Report + Presentation)

Percentage of final grade: 40% (Presentation 10%, Report 20%)

Presentation date: 5/4/2023, Class

Report due date: 5/12/2023, 11:59PM ET

Report

Max Length: Five pages (excluding figures, tables, and references)

- Introduction
    - What problem did you solve?
    - Why is this problem important and novel?
    - What similar research (academic publications) has already been published on this topic?
    - What are the major contributions of your work?
- Methods. A reader should be able to replicate your work after reading the Methods.
    - What dataset did you use for your project, or prior research did you analyze?
        - If you used a dataset, please give an overview of the dataset, how the data was collected, enrollment criteria, the raw features, any cleaning performed (eg, for missing data), and any calculations for derived features
        - If you are conducting a literature review or writing a perspective, please describe the methods you used for your literature search including the search keywords, inclusion/exclusion criteria, and your reasoning behind these choices.
    - Please detail the methodology behind your analyses (eg, what algorithms, if any, were used?)
- Results
    - Please have a paragraph that details each major finding.
    - Support these paragraphs 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?

Presentation

Presentation Length: 8 minutes

Please make sure your video 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?