Project Description: BDH-reproducibility-challenge

Jimeng Sun

***Abstract*—CSE6250 Big Data Analytics for Healthcare is a graduate level course focusing on practical big data technology and machine learning for health analytic applications. One big part of this course is to conduct a group project that reproduces some published work about machine learning/deep learning (ML/DL) in healthcare. While taking this challenge, you should understand, replicate, and evaluate some recent publications and provide an end-to-end coverage of the core implementation in the publication including data, machine learning algorithms, validation of outputs, etc. For the final project, with the help of an AI assistant, your team will need to reproduce an AI for a healthcare research paper from a predetermined list. Since we recognize that generative models (i.e ChatGPT, Claude, Gemini) are becoming increasingly adopted in workflows, we are asking each of you to record your use of generative models in the reproducibility process as the appendix part. We want to understand how people are using these tools, whether these LLMs are being used properly, and more importantly, are they helpful in an educational or project setting. I hope that the best projects (with some additional effort) can lead to a review at the best medical informatics venues. This document provides the project guideline such as expectation, timeline, deliverables.**

***Index Terms*—Machine learning/Deep learning, Healthcare, reproducibility,**

I. Paper Selection

You should select at least a paper from the provided paper pool and aim to replicate the main claim described in the paper. The objective is to assess if the experiments are reproducible, and to determine if the conclusions of the paper are supported by your findings. There are some considerations in choosing a paper to reproduce:

- You should find the problem tackled in the paper interested.
- You should be able to access the data you will need to reproduce the paper's experiments.
- You should choose paper whose computational requirements for reproducing the experiment is affordable to you.
- You cannot not choose a paper that you, course staff, or someone in your current or former lab has written.
- Even though the codebase in paper is open source, which is very common nowadays, you should not directly copy-use it. Instead, you should develop your own code. Of course, the codebase in original paper can be your reference.

Your project will try to reproduce the main experiments in your selection and assess the ease of reproducibility. The project result can be either positive (i.e. confirm reproducibility), or negative (i.e. explain what you were unable to reproduce, and identify all the questions that would need to be answered to reproduce the experiments). Both outcomes are acceptable and can earn full credit.

II. Teams

You **MUST** team up with another classmate, i.e., your team should be composed of **TWO** people. Individual project is not allowed. Both team members will receive the same grade on the project (except that on member contributes much less than the other, which will lead to penalty on the project grades of the inactive member).

III. Project Milestones

Next we summarize the timeline for your project in this semester. Please remember that **we don't allow late submission** for any project components.

Due Date	Task Description
Feb 10	Group formation & (candidate) paper selection
Feb 17	Project proposal
Apr 21	Final Submission (final paper + code + presentation)

IV. Project Requirements

A. Group formation & Paper selection

You should form a team with another classmate and your team should select 2-3 candidate papers from the paper pool. Not every paper in paper pool is feasible for reproduction, you should read the paper to see if you are able to reproduce it (whether you have data access and sufficient computational power) before you select. Course staff will share a spreadsheet for team registration & paper selection, it contains two sheets. team registration & paper selection

1) Team registration. Every team should register two team members with name, GT Username, GT email and you should also input the indices of your candidate papers.
2) Paper Selection. It will contain indices of papers in paper pool. You should input your team index and team member initials after the paper you want to reproduce so that everybody knows which paper has been selected already.

Note: Please watch the header while finishing the sheets and DO NOT change information of other group!

After finish two sheets, write the information up and submit on Gradescope. It can be in any format but must contain:

1) Names, GT usernames of two members.
2) Indices, titles and authors of selected papers.

- At this stage, your team should read those candidate papers roughly and construct basic understandings of them.
- To avoid that many groups choose one same paper, we restrict each paper with maximally TWO groups. The paper selection is first come first serve.
- Deliverable:

- Finish **TWO** sheets we mentioned before.
- Submit a **PDF** no longer than 1 page on Gradescope **at team level**. Use this instruction for how to submit at team level. Remember to include your teammate!

All requirements must be finished before deadline to earn full credits, fail to finish all of them, such as no team-up, not finish both sheets (finish only one or none), writeup submission on Gradescope is missed or does not include both members, etc, will make the your team get **zero** for that and your team has to re-choose candidate papers from whatever left.

B. Project Proposal

At this stage, you review those candidate papers and decide which paper to reproduce. Your proposal should cover the following contents about your project, in order to demonstrate you have thought carefully about the paper you are planning to reproduce, and in order to communicate your understanding of the work and its importance to someone who most likely has not read the paper (e.g. course staff). With the assistance of an LLM assistant, you can draft a project proposal for the reproduction of your selected AI healthcare paper.

For those not already using LLMs, the following are free to use:

- Claude-Haiku-3.5
- Gemini 1.5 and 2
- GPT-4o mini

Please keep track of: The LLM you used The prompts you used for understanding and summarizing the algorithms in the paper. Please keep the chat saved. Your insights and analysis on making prompts more effective.

Compute Limits: We recommend the usage of free Google Colab Colab has a T4 GPU with 16 GB of VRAM with 1 hr limit. This amount is typically enough for non-LLM research.

- Please follow the exact rubric and format below for your proposal report to summary to all candidate papers.
  - What is the general problem in this work (e.g. '10-day ICU readmission prediction')
  - What innovations are in this work (e.g. a new network structure/special feature construction/specific analysis to data)?
  - What advantages/disadvantages does the work have (e.g. accuracy to current problem is high/method is hard to be generalized)? What do you think could improve their method? Is their hypothesis legitimate?
  - What is the data used in this work? If the data is accessible, attach the link in your paper
  - Is the codebase provided? If so, attach inks to the codebase in your target paper (Github, Gitlab, Bitbucket, etc).
  - Discussion of the feasibility of the computation.
  - **What was the initial prompt that you used? What was the initial output of the LLM? Validate the LLM response. How correct, relevant, and helpful was the LLM? How many prompts did you use? If the initial prompt didn't work, what was wrong with it?** **(In appendix)**

Note: it is important and required to well organize this summary. Use a list would be favorable for graders to quickly get your point. Such as:

Paper 1: Index, Paper Title, Venue and Author(s)

1) Task: 10-day ICU readmission prediction.
2) Innovation: a new network structure is proposed.
3) Dis/Adv: model achieves high accuracy, which beats the bench mark, however, it contains abundant coefficient and takes long time to train.
4) Data Accessibility: Yes (Link)
5) Code Accessibility: Code is not provided by author

Paper 2: Index, Paper Title, Venue and Author(s)

1) Task: .....
2) Innovation: .....
3) Dis/Adv: .....
4) Data Accessibility: .....
5) Code Accessibility: .....

- Decide your target paper
  - Which paper in the candidate you will replicate.
  - Why you choose the paper.
  - What are the specific hypotheses from the paper that you plan to verify in your reproduction study?
  - Briefly state how you are assured that you can obtain appropriate data and computational resources including software and hardware demanded in the paper.
- Deliverables
  - A **PDF** no longer than **2 pages main content** + Appendix + References (The candidate papers and important references in them must be cited)
  - You do not need to include your responses to the questions in **color blue** in the proposal's main part. These responses should be attached to the proposal's appendix section.
  - Submit to **Gradescope** at team level, remember to include your teammates in submission.
  - Use your own words instead of description in original paper.

Note: It is encouraged to choose more than one paper or paper with greater difficulty to reproduce. Let's say if the paper you choose to reproduce is difficult, i.e., it requires large dataset and has complicated network structure and optimization method, perform greatly on provided metrics, you will be fine even though you could not reproduce with comparative results. Such work usually contains many details, and it is not easy to cover all of them. So that we would not be very harsh on the results you obtain. However, if the paper is relatively easy, we would expect that you obtain convincing results.

C. Final Report

At this stage, you should have completed the code development (including documents) and run multiple experiments to test all the hypothesis in the paper. You should make assessment on the reproducibility of the paper with supportive evidence that obtained from your own experiments.

- Complete all sections in Report Template.
- Upload your well-documented codebase to Github, Gitlab, etc and attach the link at beginning of your report.
- Upload your presentation video in Youtube, OneDrive, Google Drive, etc and attach the link at beginning of your report.
- Deliverables
  - A **PDF** no longer than **8 pages main content** + Appendix + References. This file should contain the links to your codebase and presentation video.
  - You do not need to include your responses to the questions in **color blue** in the final report. These responses should be attached to the final report's appendix section.
  - We recommend using an AAAI latex template (Overleaf Pro is free to use with your GT email): Latex template link.
    - You can use the LLM to help you draft the submission. But, please do not only copy and paste.
  - Submit to **Gradescope** at team level, remember to include your teammates in submission.

D. Presentation

We expect a well-timed, well-presented presentation. You should clearly explain what the original paper is about (what the general problem is, what the specific approach taken was, and what the results claimed were), what you encountered when you attempted to reproduce the results, and what extensions/ablations you made.

- Your presentation will be 5-8 minutes long. It should delivered through online platform (Youtube, OneDrive, Google Drive, and any other places that grader can access easily).
- Explain the general problem clearly.
- Explain the specific approach taken in the paper clearly.
- Show your results and compare with the results in paper.
- Why do your results match with the paper results?
- Why are your results worse/better than the ones in paper?
- Explain the extensions/ablations you made
- Good visuals are important here. Text should be in large font, figure and tables are captioned with description.
- Deliverables
  - Attach the video link in final report
  - Zip your slides with report and submit it on **Canvas** at individual level, i.e., both members should submit the zip.

V. Report Template

ML/DL in healthcare usually requires many data processing (ETL) and modeling, in project, you can use any open source packages to do it rather than invent wheels by yourself. A very pwoerful package is PyHealth (https://github.com/zzachw/PyHealth) , which is developed by Dr. Sun's group. It is managed on Github and integrates ETL and modeling. You are welcomed to fork or clone for your project development.

1) Abstract
   - Link to your video
     - No points will be awarded if the link is not clickable or does not direct to your public project video
   - Link to Public GitHub Repo
     - No points will be awarded if the link is not clickable or does not direct to your public project video
2) Introduction
   - A clear, high-level description of what the original paper is about and what is its contribution to the wider research space. Cite the original paper.
3) Scope of Reproducibility
   - List all hypotheses from the paper you will test and corresponding experiments you will run.
4) Methodology
   - Dataset description
     - Source of the data: where the data is collected, provide the link if possible; if the data is synthetic or self-generated, explain how.
     - Statistics: dataset size, cross validation split, label distribution, etc
     - How do you use the data: change the class labels, split the dataset to train/valid/test, refining the dataset
     - Please use LLMs to help in writing data preprocessing code
       a) **What was the initial prompt that you used? What was the initial output of the LLM? Validate the LLM response. How correct, relevant and helpful was the LLM? How many prompts did you use? If the initial prompt did not work, what was wrong with it?**(In appendix)
   - Model Description
     - Includes a citation to the original paper
     - Includes link to the original paper's repo (if applicable)
     - Model architecture: layer number/size/type, activation function, etc
     - Training objectives: loss function, optimizer, weight of each loss term, etc
     - Use LLMs to help with the implementation of the model used
       a) **What was the initial prompt that you used? What was the initial output of the LLM? Validate the LLM response. How correct, relevant and helpful was the LLM? How**

many prompts did you use? If the initial prompt did not work, what was wrong with it?(In appendix)

- Others: whether the model is pretrained, Monte Carlo simulation for uncertainty analysis, etc

5) Training
   - Computational Implementation
     - Report at least 3 types of requirements such as type of hardware, average runtime for each epoch, total number of trials, GPU hrs used, and training epochs.
   - Includes Training Details
     - Loss functions
     - Please use LLMs to help write code for the training loop.
       a) **What was the initial prompt that you used? What was the initial output of the LLM? Validate the LLM response. How correct, relevant and helpful was the LLM? How many prompts did you use? If the initial prompt did not work, what was wrong with it?**(In appendix)
6) Evaluation
   - Please use LLMs to help identify and write code for metrics and evaluations.
     a) **What was the initial prompt that you used? What was the initial output of the LLM? Validate the LLM response. How correct, relevant and helpful was the LLM? How many prompts did you use? If the initial prompt did not work, what was wrong with it?**(In appendix)
7) Results
   - Report results for all experiments that you run:
     - specific numbers (accuracy, AUC, RMSE, etc)
     - figures (loss shrinkage, outputs from GAN, annotation or label of sample pictures, etc)
   - Comparison with the hypothesis and results from the original paper.
   - Explain why the results may be the same or different.
   - Additional Extensions or Ablations
     - Use LLMs to help brainstorm at least one new extension - new dataset, new loss function, removing a part of the model (ablation study), etc.
       a) What extensions did the LLM come up with? How valid are they?
       b) What were your insights and analysis on making prompts effective for brainstorming?
     - Use LLMs to help implement your planned extension(s) and validate them. Include results and a discussion.
8) Discussion
   - Make an assessment of whether the paper is reproducible or not.
   - Explain why it is not reproducible if your results are kind of negative.
   - Describe 'What was easy' and 'What was difficult' during reproduction.
   - Make suggestions to the author or other reproducers on how to improve the reproducibility.
9) Authors' Contributions
   - Please clearly mention the workload distribution between the group members.