Human Collaboration with Generative AI Models for Reproduction Study: "Reproducible Survival Prediction with SEER Cancer Data"

Anh Nguyen (Matta Nguyen)
University of Illinois Urbana-champaign
anhn4@illinois.edu

Abstract

This proposal outlines a 6-week reproduction study of "Reproducible Survival Prediction with SEER Cancer Data" by Hegselmann et al., leveraging generative AI assistance while maintaining scientific rigor. The study aims to validate the paper's reproducibility claims using updated data and models while contributing to transparency in healthcare machine learning research.

Objective

To reproduce and validate the methodology and results of the paper "Reproducible Survival Prediction with SEER Cancer Data" by Hegselmann et al., while establishing a framework for reproducible cancer survival prediction with up-to-date data research within a 6-week timeframe.

Key Points to be Established

- Validate reproducibility claims of the original paper with updated data and models
- Establish verified baseline results for cancer survival prediction
- Contribute to transparency in healthcare ML research
- (Optional): Offer another approach using different ML models

Summary of the Paper

During the initial support from Anthropic Claude 3.5 Haiku with the System Prompt "Help producing answers as you're a medical researcher explaining concepts for machine learning students who know nothing about medicine," there are key notes conducted from the initial review of the document:

- The paper notes that while machine learning applications in this domain are increasing, particularly using the SEER (Surveillance, Epidemiology, and End Results) database, fundamental issues remain: the lack of reproducibility in research methodology and results.
- The paper's novelty lies not in proposing new prediction methods but in establishing a framework for reproducible cancer survival prediction using SEER data. The authors utilized a systematic approach using logistic regression and multilayer perceptron (MLP) models, including a novel MLP with embedding variation.
- The study utilizes SEER data from 2004-2009, focusing on breast and lung cancer cases. While the SEER database requires authorization from the National Cancer Institute for access, the authors provide SEER*Stat session files and open-source code to ensure reproducibility. Their approach compares minimal data preprocessing with 1-n encoding of categorical inputs, demonstrating that 1-n encoding consistently improves performance across all experiments.
- Results showed that for breast cancer, logistic regression with 1-n encoding performed similarly to MLP models, while for lung cancer, MLP models demonstrated slightly superior performance. However, the authors emphasize that their contribution lies not in superior prediction performance but in establishing reproducible benchmarks and honest reporting of results, addressing a fundamental need in healthcare ML research.

Technical Approach

Phase 1: Data Foundation and Setup - Week 1

The plan for the first week is to focus on accessing the data and structure project foundation including GitHub repository and utilize LLMs to generate initial data processing code.

Phase 2: Implementation - Weeks 2-4

With LLM assistance, the purpose of this phase is to implement the paper's core models: logistic regression baseline, MLP architectures, and MLPEmb variations. Each implementation will follow a structured process: LLM-generated code, manual review, testing, and documentation. While not committing in original delivery due to the constrain with timeline, this project can expand to build a new model that was not included in the original paper to validate the find.

Phase 3: Analysis and Deliverable Preparation - Weeks 5-6

The final two weeks focus on analysis, visualization, and deliverable preparation. With the support of LLMs, the delivery could include generating visualization code and con-

duct ablation studies, while carefully documenting on the findings. This phase includes preparing our 4-minute presentation, finalizing GitHub documentation, and completing our PyHealth contribution. Special attention will be given to documenting the role of LLMs in our reproduction process, including limitations and learned best practices.

References

Reproducible Survival Prediction with SEER Cancer Data,Hegselmann, Stefan and Greulich, Leonard and Varghese, Julian and Dugas, Martin