Thesis Project Portfolio

Image Processing Tool for Quantifying Immunostained Sections of Fibrotic Cardiac Tissue

(Technical Report)

Machine Learning in Healthcare: How Strict Data Collection Policies Lead to Misrepresentational Models

(STS Research Paper)

An Undergraduate Thesis

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Jakub Lipowski

Spring, 2022

Department of Biomedical Engineering

Table of Contents

Sociotechnical Synthesis

Machine learning is becoming an increasingly relevant tool across academic and industrial applications. One key application is image processing. The capstone design project presented in this portfolio implements machine learning incorporated into an image processing software to classify and identify cells in images of injured heart tissue from rats. The end goal is to quantify spatiotemporal dynamics between cell populations to help identify targets for enhanced regeneration of injured heart tissue. While applications of machine learning, like in this design project, may appear limitless, it is also important to examine the social and ethical considerations surrounding artificial intelligence. Every machine learning model needs to be trained with high quality data that can generalize to the public. However, health data collection regulations in the United States make it difficult to obtain sufficient data for the training of an ethical model. The STS research in this portfolio examines how obstacles to data collection combined with a history of an inequitable healthcare system cause bias in machine learning technologies, which is extremely problematic in an industry like healthcare.

The technical report describes a novel automated image processing tool that quickly and accurately quantifies cell populations within immunostained sections of tissue that has suffered a heart attack, also known as a myocardial infarction (MI). The healing process following MI consists of several multicell interactions, namely between monocytes, macrophages, and fibroblasts. It is important to understand the spatiotemporal dynamics between these cell types in order to identify key targets for therapeutic intervention to enhance cardiac tissue regeneration. The software developed as part of the capstone project is equipped with a feature meant to quantify call populations within the vicinity of key structures such as blood vessels where much of the inflammatory response in MI is thought to originate. In addition, the software can quantify cell populations in entire tissue sections. The image processing tool features a user-friendly graphical user interface, a backend machine learning model, and a user manual meant to help users run the code from a platform like GitHub. In addition, the data obtained by this software can be used to inform future research on the spatiotemporal dynamics among cell populations.

As the world's industries enter a data revolution, leaders have to consider the ethical implications of data usage and artificial intelligence system development. Healthcare, the United States' largest industry, is certainly no exception. This STS research answers the question: "How can the discrepancy between a need for representative training data and strict data collection polices within an unequal healthcare system be reconciled to minimize the risks of biased machine learning models in the United States?" Machine learning models in healthcare need to be trained with data representative of all demographics, otherwise these models cannot generalize to all patients, which leads to fatal decisions. Rigid regulations concerning collection of health data imposed by the Health Insurance Portability and Accountability Act (HIPAA) make it difficult to obtain reliable data from healthcare institutions. Furthermore, inequalities embedded in the United States' healthcare system skew the data that is available towards specific demographics. Risk analysis is used to explain hesitation around loosening data collection policies. Reservations concerning health data privacy is risk analysis in action. However, in the process of risk analysis, society examines itself, which eventually leads to reform. Through analysis of cases where machine learning models in healthcare resulted in failure due to misrepresentational training sets, this research will shed light on how pieces of existing HIPAA policies could be amended to ensure such cases to not occur again. Before society can reap the benefits of machine learning in healthcare, it must ensure that artificial intelligence systems are tailored to all existing demographics.

Joining a technical implementation of machine learning with a discussion of the limitations surrounding machine learning applications in the United States healthcare system helps develop a holistic impression of the subject. If the technical aspect of this portfolio was completed disjointly, it would be easy to overlook sociotechnical implications of the artificial intelligence being used. After all, when writing code for a software, the priorities lie in debugging. However, the analysis of limitations of machine learning in healthcare caused by data collection practices creates a more high-level overview of the technical project. It contextualizes the image processing software by encouraging the consideration of how data for future iterations of the project will be collected, and which communities the project will impact.