THESIS PROJECT PORTFOLIO

Designing Machine Learning Methods for RNA-Sequencing Analysis of Atherosclerosis

(Technical Report)

Adapting to the Age of Data-Driven Medicine: Analysis of Healthcare Systems on the Deployment of Artificial Intelligence in Medical Imaging

(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

Jainam H. Modh

Spring, 2022

Department of Biomedical Engineering

TABLE OF CONTENTS

The global artificial intelligence (AI) and machine learning (ML) market is growing by 40-50% each year and is projected to exceed $300 billion by 2025. Amongst all major industries however, medicine and healthcare are the most difficult yet important domains of application for AI. Modeling the complexity of individual patient's health patterns, tailoring medical decisions and care, and providing accurate predictions of a patient's clinical outcomes is a difficult yet necessary path towards reaching the upcoming gold standard of personalized, data-driven medicine. Specifically, complex diseases such as atherosclerosis and cardiovascular disease require intricate and expansive diagnostic tools and treatments. Machine learning methods attempt to handle and investigate the nuances of these diseases by allowing computers to learn from medical data, which includes both genetic and imaging data, and perform tasks aimed towards gaining further insight into the disease and improving medical outcomes. This portfolio discusses and attempts to solve some of the challenges faced in the development and deployment of machine learning in healthcare and medicine. Both the STS research paper and technical project focus on identifying and improving upon machine learning applications so that they will be better utilized and understood by doctors and researchers while improving health outcomes for patients.

Atherosclerosis is one of the primary causes of cardiovascular disease, which accounts for 32% of global deaths and costs the United States $312.6 billion each year. Heritability estimates for atherosclerosis and its associated diseases vary between 40% to 70%, suggesting a strong genetic contribution to the disease. Investigations need to be conducted on how genetic variations manifest as changes in gene expression profiles at the early stages of atherosclerotic development. Thus, the technical project aims to leverage machine learning methods to analyze tissue-level RNA-seq data of coronary and aortic atherosclerosis in young adults and predict a

sample's pathology given its transcriptomic profile. Various feature selection methods were implemented to extract biologically relevant gene sets that were then used to train XGBoost classifiers to differentiate between early and late stages of atherosclerosis. The best performing models were trained on the top 90 and 120 differentially expressed genes yielding an accuracy of 85.96% and 86.54% for aortic and coronary samples respectively. Non-linear dimensionality reduction and gene ontology enrichment analysis were used to characterize the transcriptomic profiles representative of the gene sets determined by the feature selection methods. The results shown in this project lay the groundwork for the application of machine learning methods in the analysis of RNA-seq data at the tissue level which could eventually lead to applications in personal and preventative treatments for atherosclerotic development in young adults.

AI and ML tools are also vital in the field of medical imaging, which is being overwhelmed with the ever-increasing amount of complex and rich data that is being collected to make medical diagnoses and clinical decisions. AI and ML are currently being researched and integrated into this field by automating image analysis; aiding in disease detection, classification, and localization; and facilitating informed and efficient decision-making. Although the development of these tools is occurring at a rapid rate, deployment of these technologies in the modern healthcare system is proving to be extremely difficult. The STS research paper aims to understand and analyze the problems faced by healthcare systems and its constituents in the deployment of AI/ML-based medical imaging applications while providing insights into the steps which need to be taken to successfully deploy these technologies and adapt to the new age of data-driven medicine. An actor network was constructed to analyze the four constituents of the healthcare system - patients, radiologists, hospitals, and regulatory agencies - in relation to the flow of medical and clinical information. Next, AI/ML-based medical imaging software was

introduced to the same actor network and its effects on the network were investigated. Significant insight was obtained in how AI and ML technology affected and strained the flow of information through the healthcare system such as the inaccuracies caused by biased patient data collection methods and changes in the roles of radiologists when processing data. Additionally, several solutions were proposed to assist the system in overcoming these barriers in deployment to increase the application of AI and ML in healthcare in the near future.

Significant amounts of research were conducted over the past year regarding these two projects. Both projects were ultimately quite successful, producing impactful results that solve important problems related to machine learning and healthcare. However, I ended up taking on too much for both my projects as their scope was too large and infeasible in the given timeline. I only completed two-thirds of my initial goals set out for my technical project and my STS research paper would benefit if my claims were backed with more evidence. I request future researchers to contribute to projects related to machine learning in healthcare as there is a tremendous amount of work to be done in this field that is impactful and disruptive.