Thesis Portfolio

A review on the need for better explainability with increasing reliance on machine generated medical diagnostics

(Technical Report)

Potential Technical Solutions to Mitigate the Effects of Bias on Machine Learning Algorithms

(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

Eric Armstrong
Spring, 2021

Department of Computer Science

Table of Contents

Sociotechnical Synthesis

The introduction of machine learning into more and more fields has allowed for massive growth in many industries by allowing more people to have access to jobs, healthcare, legal solutions etc. This is due to the speed and volume of prediction that can now be done with machine learning models. However, when these models are difficult to understand, it is difficult to validate that the decisions they make are made for the right reasons, based on logical features in the data versus potentially unintentional patterns. This is particularly important for the introduction of machine learning into the medical field, and lots of research is being done on the best technical ways to allow health care professionals to work with and understand how a machine diagnosis may have been made, in order for hospitals and doctors to avoid legal trouble with medical malpractice. Being able to understand predictions is also crucial in ensuring biases in the data are not influencing the decision making of automated systems. While biases will eventually show themselves over time from observing the automated predictions, it is important to find them as early as possible so that minority groups are not discriminated against, especially in medical, legal, and job hiring areas where machine learning is widely used and has significant impact on someone's quality of life.

It is important to recognize that there is no quick fix to either of these issues particularly with the social and legal aspects of both. Just changing models to force fairness in decision making or altering the data to be unbiased is not going to change the underlying biases in society that are affecting people every day. These biases need to be brought up and discussed in personal and professional environments so that people are more conscious and future generations are better educated. And while there are some solutions to allowing machine learning into the medical field with models that train around black box models to show how different inputs affect the outputs, currently the most feasible being human in the loop systems where doctors validate machine decisions and have easy to understand data that indicates how a decision was made, there would need to be an overhaul of training and cost structures at every hospital in the country, alongside legal changes to ensure doctors, hospitals, and patients are treated with the highest care possible to

avoid medical malpractice and ensure trust. Ultimately, I think it is clear that while machine learning has led to many great things for society it also has led to some unintended consequences. For machine learning systems to be fully accepted in all areas, more research needs to be focused on transparency of black boxes, and bias mitigation, and there may also need to be more regulations in place to ensure these new techniques are effectively being used.