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

(Technical Paper)

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

(STS Paper)

A Thesis Prospectus Submitted to the

Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements of the Degree
Bachelor of Science, School of Engineering

Eric Armstrong
Spring, 2021

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Signature _____ Date _____
Eric Armstrong

Approved _____ Date _____
Anil Vullikanti, Department of Computer Science

Approved _____ Date _____
Bryn Seabrook, Department of Engineering and Society

Introduction

With machine learning on the rise within businesses, many problems are being solved that were not feasible before; however, these algorithms depend on very large amounts of data that are often unrepresentative of the population (Manyika, 2019). A recent study on some of the leading facial recognition software found that nearly 100% of white males were correctly identified while non-white females were recognized correctly only 65-79% of the time (Mames, 2018). This bias is not something that is coded into the algorithms, but a result of how math centric algorithms draw conclusions from biased data. There are multiple approaches to addressing this problem, but the two that will be the focus of this STS portfolio are using fairness metrics, such as requiring that models have equal predictive value across groups, and using advanced machine learning techniques to generate new "fake" data that is indistinguishable from real data to offset the biases at the data level. However, each of these approaches have their own problems, some of which make their use untenable. Due to the presence of bias, the user of an algorithm must understand how a machine came to a decision in order to trust its prediction. To help users trust their machine learning models, lots of research is currently being done on how to make models more explainable and interpretable for the user. Particularly to help medical professionals understand machine predicted diagnostics by explaining how a given prediction was made. This is relevant because machine learning diagnostics are outperforming doctors but the lack of explanation for such diagnostics could affect the quality of medical diagnostics in the long run and have unintended consequences in regards to safety standards, due to current medical malpractice law. For example, current models are identifying melanoma in 95% of images whereas doctors are currently only able to identify 86.6% and around 5% of adults receive misdiagnosis from doctors, which creates the need for additional treatment (Froomkin, Kerr, Pineau, 2019). The superiority of machine predicted diagnostics could require doctors to make diagnoses alongside AI to help validate their predictions and ultimately reduce the rate of misdiagnosis. If models lack explainability and doctors cannot understand why certain diagnoses are being made, they will not be able to effectively assign a treatment, leading to a decrease in the quality of care. For this reason, such models must be able to explain their decisions. By synthesizing current research in medical law and explainability in regards to machine learning, there will be a better idea of where we currently stand and where further research must be done.

Technical Topic

Medical malpractice occurs when a hospital, doctor or other health care professional, through a negligent act or omission, causes an injury to a patient. The negligence might be the result of errors in diagnosis, treatment, aftercare or health management (ABPLA, 2020). As machine learning diagnostics continue to improve and are consistently able to outperform medical professionals, medical malpractice law could cause the new standard of care to require that doctors make diagnoses alongside a high-quality machine learning diagnosis. (Froomkin, Kerr, Pineau, 2019). However, medical malpractice encompasses more than diagnosis. If the injury was due to the medical professionals' error in providing a treatment, it could still be considered malpractice. Consequently, if machine learning models are too difficult to be understood by doctors, they may not know how or why a certain diagnosis was made, and may neglect to provide the best treatment possible. Since treatment strategies are often not as effective when deployed in clinical practice compared to preliminary evaluation, there could be a

decrease in the quality of care, which is the primary purpose of the medical malpractice law to begin with. (Froomkin, Kerr, Pineau, 2019). Therefore, doctors and hospitals must be able to trust this new technology to accept it. In order for machine learning diagnostics to be readily accepted by doctors, more transparency is needed for medical professionals to understand which factors were most significant in determining the output and why. There are many different approaches to make models explainable, but in order for a model to be explainable from a legal perspective, a set of rules must be defined, particularly with what the model should guarantee (e.g. trustworthiness of the model and comprehensibility for medical professionals) (Guidotti, Monreale, Ruggieri, Turini, Giannotti, Pedreschi, 2018). Due to the large number of techniques that have tried to explain predictions of models, the analysis will be on models that have focused primarily on medical image analysis and medical diagnostics and that seem to be the most popular within the current research community. In addition to potential techniques that can be employed, it is important to examine proposed changes to medical liability rules in order to accommodate the introduction of machine learning, as well as preventing it from replacing doctors entirely and instead requiring physicians to be in the loop (Froomkin, Kerr, Pineau, 2019). Such a model, often called human-in-the-loop (HITL) hybrid requires keeping the clinical expert in control of the process and responsible for the final decisions but with a reduced workload and an added safety net (Singh, Sengupta, Lakshminarayanan, 2020). The goal for the literature review is that by comparing multiple papers regarding general explainability, explainability of medical image analysis models, and papers regarding legality, trends could be identified in where the papers overlap and where they disagree. From there, the discussion will move towards being able to frame how current medical litigation will have to be altered alongside needed progressions in explainability of models to safely and effectively allow the introduction of machine learning diagnostics into the medical field, ultimately improving the well-being for both medical professionals and patients.

STS Topic

As one of the biggest buzzwords over the last few years, machine learning is a very promising technology in almost every aspect of society, however, this technology remains very prejudiced to many groups of people due to biases within training data and poorly monitored algorithms. These biases particularly affect non-white, non-male, and lower-income populations (Kleinberg et al., 2020). Machine learning is an iterative process that tries to improve itself every iteration by identifying patterns in the training data that lead to the best performance based on the attributes in the data (race, income, location, gender, etc.), where performance could be defined as accuracy, precision, F-score, etc (Chouldechova, 2017). In an investigative article by the news site ProPublica, a criminal justice algorithm used in Broward County, Florida, mislabeled African-American defendants as “high risk” at nearly twice the rate it mislabeled white defendants. Other research has found that training natural language processing models on news articles can lead them to exhibit gender stereotypes. These biases are due to how machine learning systems learn to make decisions based on training data, which can include biased human decisions or reflect historical or social inequities (Manyika, 2019). However, sometimes bias can come from unexpected sources, and finding what causes these biases is much more difficult. For example, Amazon had to stop using a hiring algorithm after it realized it was favoring words typically found on men’s resumes over women’s (Guidotti et al., 2018). This bias would not have

been obvious to someone who thought they had taken all precautions to make candidates as anonymous as possible by removing gender, race and other identifying attributes. Part of the reason biases are hard to detect is due to the black box nature of the algorithms, making it hard to tell what the algorithm is actually learning from data. For example, a famous example is a classifier trained to recognize wolves and husky dogs. The classifier ended up basing its predictions to classify a wolf solely on the presence of snow in the background and very little do actually do with the appearance of the animal. There are multiple proposed solutions to mitigating bias in machine learning algorithms. One such solution is the introduction of fairness metrics to ensure that models are not discriminating against certain groups. However, one problem is that there is no one definition of "fairness" (Chouldechova, 2017). There are three common metrics agreed upon in most literature, yet there is no method that can satisfy these three conditions simultaneously (Kleinberg et al., 2016). Additionally, there are guidelines set forth by companies such as Google and IBM that help businesses continually mitigate biases in AI by providing frameworks and recommended practices (Manyika, 2019). A second method addresses the problem at a data level instead of imposing fairness on the model. It involves using conditional Generative Adversarial Networks (cGANs) to generate new synthetic, fair data with selective properties from the original data, as well as a framework for analyzing data biases, which is important for understanding the amount and type of data that need to be synthetically sampled and labeled for each population group. (Abusitta, et al., 2019). GANs are a special type of deep learning model that can be used to generate or output new examples that plausibly could have been drawn from the original dataset and can even be altered to have desired characteristics (Brownlee, 2019). This approach to mitigating bias is better in that general accuracy is improved rather than reduced, since more training data is being used, and as a general rule more data is always better. While these methods to solve bias in algorithms and data are a solution to mitigate biased predictions, they are an example of a technological fix for a much larger and long-term societal issue that stems from the history of systemic prejudice against women and minority groups that has only been made easier and automated due to the rise of machine learning. These fixes also do not change the fact that altering a dataset does not mean that new incoming data will be unbiased due to human perception of these biases being the same. The technological fix framework was introduced by Byron P. Newberry and is focused primarily around asking if a given issue can or should be "fixed" with technology or left alone. Some critiques of technological fix are that even if these solutions are failing to address an underlying problem, that they still can greatly benefit a group of people, and accordingly I will address both sides of this argument in my analysis. Additionally, biased models and data can very easily be viewed as a wicked problem. In our current political climate, race and sex have become political issues that is polarizing the nation, leading to lots of heated controversy about what the best solution may be. Viewing bias in machine learning as a wicked problem, many people could argue that machine learning has benefited everyone with new technologies, but understanding how to create a sustainable system that can adapt to new visible problems that arise is crucial, especially when it is something that is starting to affect all aspects of our lives. Wicked problem framing was originally introduced by Horst Rittel and Melvin M. Webber. Although there is no distinct definition of a wicked problem, it is a problem typically unable to be solved due to the conflict of perspectives on how to solve the issue and that there is no iterative process of alleviating the issue, which essentially defines the critique of the framework.

Methodology

Research Question: How effective are some of the different approaches to mitigating the biases in ML algorithms? To answer the research question, the Wicked Problem Framing and Literature Review methodologies will be used. Beginning with background, context will be given for some major examples of how machine learning has shown clear bias in our society even in recent years with our new complex models. Seager's Wicked Problem Framing technique gathers evidence to reveal relationships between actions and consequences by showing how historical actions of social inequality are affecting modern systems (Seager, Selinger, & Wiek, 2012). Following the background will be how the cause of the bias is not in the algorithms themselves, however, but is due to biases that have existed in society that have merely transferred into the models. The use of literature review will lead the discussion of different researchers' solutions to mitigate the biases in these algorithms. The two main approaches will be a model based, fairness approach and a data generation approach to unbias the data in order to unbias the model. The discussion will be focused on the benefits and drawbacks of both approaches and analyze how effective they would be in practice. There will also be discussion about research about the explainability of models in an attempt to discover where the biases are coming from within the data that might not be obvious from a human perspective. The sources for the review will primarily be recent research papers, published in the last couple years, as well as the fairness metrics that have been published by major companies.

Conclusion

This proposal covers the potential for machine learning to enter the medical field to supplement medical diagnostics, while addressing potential challenges such as solutions for explainability specifically within a medical context and the need for change in current medical litigation. The analysis will focus on legal issues regarding medical malpractice and standard of care and how under the current system, if machine learning is proven to be consistently better than medical professionals in diagnosing, a human diagnosis could no longer be considered to meet the criteria of standard of care and could be malpractice. The goal is that by analyzing multiple sources for explainability, machine learning diagnostics, and machine learning in a medical context, that a consensus could be found between them that could show the optimal direction for progress moving forward. On a similar note of explainable machine learning, the STS research paper will consider multiple proposed solutions to mitigate the bias in machine learning models. It will primarily consider fairness metrics and the use of cGANs and how effective they may be in mitigating bias and producing fair and accurate results for all groups. It will also touch on how explainability could help identify sources of bias in data that would otherwise be very difficult to identify. By reducing or effectively eliminating the bias in machine learning models, society will become fairer in many aspects due to the applications of machine learning being found almost everywhere.

References

Abusitta A., & Aïmeur E., & Wahab O. A. (2019). Generative Adversarial Networks for Mitigating Biases in Machine Learning Systems. Retrieved September 29th, 2020, from

https://deepai.org/publication/generative-adversarial-networks-for-mitigating-biases-inmachine-learning-systems

American Board of Professional Liability Attorneys (2020). What is Medical Malpractice?. Retrieved from https://www.abpla.org/what-is-malpractice

Chiappa S. (2019). Path-Specific Counterfactual Fairness. Retrieved October 15th, 2020, from https://aaai.org/ojs/index.php/AAAI/article/view/4777

Chouldechova A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Retrieved October 15th, 2020, from https://arxiv.org/pdf/1703.00056.pdf

Card D. (2017). The "black box" metaphor in machine learning. Retrieved October 15th, 2020, from https://towardsdatascience.com/the-black-box-metaphor-in-machine-learning4e57a3a1d2b0#:~:text=The%20black%20box%20metaphor%20dates,Skinner%20concep tualized%20minds%20in%20general.

Manyika J. (2019, October 25). What Do We Do About the Biases in AI? Retrieved September 29, 2020, from https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai

Brownlee J. (2019). A Gentle Introduction to Generative Adversarial Networks (GANs). Retrieved October 15th, 2020, from https://machinelearningmastery.com/what-aregenerative-adversarial-networks-gans/

Kleinberg J., & Ludwig J., & Mullainathan S., & Sunstein C. (2019). Discrimination in the Age of Algorithms. Retrieved October 15th, 2020, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3329669#references-widget

Kleinberg J., & Mullainathan S., & Raghavan M. (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores. Retrieved October 15th, 2020, from https://arxiv.org/abs/1609.05807

Mames. (2018, December 11). Impact of Algorithmic Bias on Society. Retrieved September 30, 2020, from https://blogs.ischool.berkeley.edu/w231/2018/12/11/impact-of-algorithmicbias-on-society/

Guidotti R., & Monreale A., & Ruggieri S., & Turini F., & Giannotti F., & Pedreschi D. (2018). A Survey of Methods for Explaining Black Box Models, Retrieved October 26th, 2020 from https://doi.org/10.1145/3236009

Iriondo R. (2020). What is Machine Learning?. Retrieved October 15th, 2020, from https://medium.com/towards-artificial-intelligence/what-is-machine-learning-mlb58162f97ec7

Schwenke C., & Schering A. (2007). True Positives, True Negatives, False Positives, False Negatives. Retrieved October 26th, 2020, from https://doi.org/10.1002/9780471462422.eoct021

Seager, T., Selinger, E., & Wiek, A. (2012). Sustainable Engineering Science for Resolving Wicked Problems. Journal of Agricultural and Environmental Ethics, 25(4), 467–484. https://doi.org/10.1007/s10806-011-9342-2

Karn U. (2016). An Intuitive Explanation of Convolutional Neural Networks. Retrieved from https://ujjwalkarn.me/2016/08/11/intuitive-explanationconvnets/#:~:text=Convolutional%20Neural%20Networks%20(ConvNets%20or, robots% 20and%20self%20driving%20cars.

Varshney S. (2018). Introducing AI Fairness 360. Retrieved October 15th, 2020, from https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/