The Use of SHAP to Address AI Biases in Healthcare

Jennifer Qiu
Queen's University
20jqq@queensu.ca

Michelle Shi
Queen's University
22tj18@queensu.ca

Shrika Vejandla
Queen's University
22nh43@queensu.ca

***Abstract*—The rise of AI has led to a need for refining and surveilling AI systems for errors and biases. This paper emphasizes the importance of supervising AI systems and also highlights the multi-faceted approach required to mitigate biases effectively. By combining a thorough literature review, rigorous research, insightful analysis, and algorithmic models, this paper offers a comprehensive framework for identifying and rectifying biases within healthcare AI. The achieved accuracy rates of 85%-89% signify a significant step forward, underscoring the feasibility of debiasing AI technology in healthcare settings. Moreover, these results serve as a promising foundation for future endeavours in this domain, suggesting further enhancements and advancements in addressing biases within AI systems.**

I. Introduction

Artificial intelligence (AI) poses immense potential concerning applications in diverse fields. The use of AI in healthcare is on the rise, posing innumerable potential applications from diagnosis to prognostication and treatment implemented in a clinical workflow, typically teaming with clinicians to streamline their work. However, ethical concerns surrounding the use of AI in healthcare feature topics including but not limited to biases, a lack of expertise of AI in healthcare professionals, and privacy and confidentiality infringement. The goal of this project is to identify, analyze, and explore how biases in healthcare AI and algorithms may be mitigated through a literature review, coding, and human-ML augmentation approach. Biases in healthcare AI and algorithms directly affect patients, healthcare professionals, hospitals, and the healthcare system as a whole. As the article by Norori et al. discusses, [Norori et al., 2023], disparities in healthcare start at the clinical data level, which then has the potential to create errors in medical decisions for certain groups, and is ultimately amplified with AI technology. Bias at all levels of healthcare and AI development has the potential to create serious complications and inequities in healthcare, highlighting the importance of fostering an awareness of these biases and working to mitigate data affected by systemic bias.

A. Motivation

The focus of this paper is on identifying, analyzing and eliminating biases (with algorithmic models). The literature review focuses on fairness/debiasing and human-ML augmentation in healthcare AI and algorithms. Many studies on this topic in the past have approached fairness as a set of governance procedures to mitigate unintended harm, which does address the aspect of fairness and reducing negative outcomes [Sikstrom et al., 2022]. However, there are very few scholarly works that work on understanding what exactly fairness/unfairness is in healthcare AI and algorithms [Sikstrom et al., 2022]. This paper focuses on biases but also focuses on the aspects of data collection, data analysis and how data is interpreted and instrumentalized within algorithms. These aspects, along with debiasing techniques, are more likely to address root issues in these technologies. This paper's topic is an emerging area in both AI and healthcare. Healthcare professionals must be able to trust the technology they are using, and this is only possible if the technology is free of biases and root problems.

B. Problem Definition

It has been previously demonstrated that AI applications in healthcare may amplify systemic biases from being trained on data that is inherently biased. Past examples include AI algorithms being used to predict the future risk of breast cancer potentially suffering from a performance gap wherein black patients were more likely to be incorrectly assigned as having a "low risk" [Mittermaier et al., 2023]. This arises due to sampling bias, the third of the five stages of the dataset lifecycle—with the other four being creation, design, collection and processing [Gao, 2021]. This is further applied in this project's predictive model, where having taken a breast cancer dataset from Kaggle; it is shown in the "Sampling Bias" section that 85% of the dataset is patients of the Caucasian race [Shi, 2024], which induces potential bias in the predictions. There are many ways to go about solving this problem. One way is to use SHAP (SHapley Additive exPlanations) values, which essentially take each feature of a dataset and determine its importance in predicting the final outcome—the output. Through this, we can "identify if certain features disproportionately affect particular groups" [Awan, 2023]. Another way to solve this problem is by Synthetic Minority Oversampling Techniques (SMOTE) which is explored further in this paper. We can measure the success of these solutions by seeing how well we can identify, analyze and eliminate biases in healthcare systems; we can also look at how user-friendly the solutions are; after all, our goal is to have AI work well with healthcare professionals in eliminating biases, so effective communication between the two parties is crucial. This paper aims to show that mitigating AI biases in healthcare through

the use of SHAP values and human-ML augmentation will produce a more equitable, diverse, and inclusive healthcare system.

II. Background and Related Work

A. Research Questions

1) What are the most effective methods to de-bias data and AI?
   - One method that can be used is called SHAP (SHapley Additive exPlanations) which helps to identify if biases are present and how they affect certain populations.
2) Which root problems need to be addressed in terms of bias in healthcare AI?
   - Root problems in healthcare AI typically arise due to systemic biases, including broad political processes, governing bodies, and their respective stakeholders that implement policies which may result in health inequities, or certain populations being generally underrepresented in health research.
3) How do we integrate ML into clinical contexts?
4) What are specific methods to enhance model fairness and prediction accuracy?
   - This could range from using a RandomForestClassifier() method or Logistic Regression; different methods have different benefits but certain ones work better with our project.

B. Contributions

The main contributions of this paper are summarized below:

1) We do a literature review of research on this topic, such as currently used AI for medical condition detection (sepsis).
2) We include a technical aspect with easy-to-understand visual graphs to help connect those who may not be as knowledgeable in the field of machine learning to our research project.
3) The main root problems causing biases in healthcare AI are addressed and ethical implications are discussed.

Provide an overview of the AI technology or application, including its technical basis, current uses, and potential future developments. Review existing literature on its ethical, social, and legal implications, and discuss related ethical frameworks or guidelines.

C. Related Works

1) **Out with AI, in with the psychiatrist: a preference for human-derived clinical decision support in depression care** [Maslej et al., 2023]
   AI-based methods are being developed to process and summarize clinical notes, namely transformer-based language models that extract information from clinical text [Maslej et al., 2023]. It is speculated that further applications will involve clinical support tools (CSTs) capable of generating AI-based summaries of clinical notes [Maslej et al., 2023]. However, validation studies have previously indicated that the success of AI-based CSTs in experimental settings may not necessarily work in real-life applications – specifically, there have been previous examples of tools giving potentially harmful recommendations for testing patients with pneumonia [Maslej et al., 2023]. Moreover, improving AI accuracy and debiasing may not necessarily translate to enhanced clinical performance, since contextual factors, such as the clinician's perspective on AI will shape interactions [Maslej et al., 2023]. For instance, psychiatry is a field in which it has been argued that AI is likely to outperform human prognostication since the field features an assessment of highly heterogeneous pathological underpinnings [Maslej et al., 2023]. It was shown that psychiatrists' ratings for treatment recommendations were less favourable when the perceived source was AI, when the recommendations were correct in the hypothetical scenario presented [Maslej et al., 2023]. This hence shows that highly accurate AI CSTs are rendered futile if they are not accompanied with careful trust and teaming with the clinician. Our algorithm would ideally be hence implemented in a context that prompts the clinician to reflect critically on AI information through cognitive methods such as delaying the AI prompt [Maslej et al., 2023].
2) **Predictive care: a protocol for a computational ethnographic approach to building fair models of inpatient violence in emergency psychiatry** [Sikstrom et al., 2023]
   Amidst the increasing awareness that algorithms can amplify systemic inequities from being trained on biased datasets, it is unclear how removing biased features from training data will impact other features, and doing so does not necessarily address underlying systemic contexts that led to the bias [Sikstrom et al., 2023]. This pilot proposal suggests computational ethnography is a means of integrating machine learning into risk assessment that could impact acute psychiatric care [Sikstrom et al., 2023]. Namely, learning how electronic health record data is compiled and used to predict the risk of violence and aggression may be conducted by leveraging patients' sociodemographic and behavioural characteristics to improve machine learning algorithms [Sikstrom et al., 2023]. This is crucial to incorporate since there have not been studies examining which patient groups may be potentially over-represented in false positive predictions, despite existing evidence of biases that could lead to perceptions of risk in patients that are intersectional [Sikstrom et al., 2023]. Our approach similarly looks to incorporate demographic data as opposed to the numeric metrics or ratings on their own.
3) **Conceptualizing fairness: three pillars for medical algorithms and health equity** [Sikstrom et al., 2022]
   This paper speaks to our broad approach in implement-

ing means of de-biasing algorithms. Specifically, it is noted that rather than working to de-bias to validate algorithms after they are constructed, it may be more helpful to address root issues by paying attention to how data are collected, learning about what kinds of data make up larger dataset, and learning about how data are interpreted and instrumentalized within algorithmic systems [Sikstrom et al., 2022]. This particularly speaks to three pillars: transparency, impartiality, and inclusion as components of 'fairness', which often may be a tradeoff [Sikstrom et al., 2022]. Namely, it is mentioned that enhancing feature representations with latent embeddings or applying neural networks can improve the ability to predict health outcomes, but it can also make models less transparent, which may serve as a detriment in the ability to examine how the algorithm incorporated patients [Sikstrom et al., 2022]. Our goal is to similarly investigate the extent to which data reflect partiality based on the populations that are represented.

4) **Human-machine teaming is key to AI adoption: clinicians' experiences with a deployed machine learning system** [Henry et al., 2022]
Integrating machine learning could be challenging particularly in time-constrained clinical contexts, particularly if clinicians struggle to trust algorithmic systems if they are not able to effectively understand the specific logic behind an alert or recommendation [Henry et al., 2022]. In this study, clinicians did not differentiate operations between machine learning-based and conventional clinical decision support systems, and were also generally responsive to its alerts and integrated them into their diagnostic process [Henry et al., 2022]. That said, some clinicians did express concerns such as potential for over-reliance on automated systems which could potentially degrade their clinical abilities in the long run [Henry et al., 2022]. Our project seeks to supplement clinician abilities by fostering affecting human-AI teaming such that there are features embedded to prompt clinicians to justify their decision and describe their understanding of the algorithm's suggestion.
5) **Artificial Intelligence for Early Sepsis Detection: A Word of Caution** [Schinkel et al., 2023]
This article demonstrates the relationship between AI and sepsis. Sepsis, known to affect millions of people worldwide, is generally a condition that is difficult to diagnose in its early stages due to its inconsistency in appearance [Schinkel et al., 2023]. So medical professionals and researchers decided to integrate AI into their healthcare system and "develop automated systems that provide timely alerts and make physicians aware of imminent sepsis" [Schinkel et al., 2023]. While this did prove to be effective in treating the early stages of sepsis, it also posed problems as these "timely alerts" resulted in misuse of antibiotics [Schinkel et al., 2023]. Furthermore, sepsis is more complicated and individualistic than something AI algorithms, which are based on patterns and correlations, can handle [Schinkel et al., 2023]. This relates to our paper in the sense that we are also addressing inequitable issues that come with AI; while it is used with good intentions, there are always unintended consequences. In many instances in healthcare, those unintended consequences will be biased towards certain populations. With bias comes things like improper treatment and care for minority groups and misdiagnoses. This is why it is important to address these consequences, as we are discussing in this paper.
6) **On Improving Fairness of AI Models with Synthetic Minority Oversampling Techniques** [Zhou et al., 2023]
Biased AI models are problematic as they reflect the biases present in data collection, developers, and society. They lead to the amplification of existing social inequities. One technique that aims to improve the fairness of AI models is the Synthetic Minority Oversampling Technique (SMOTE). In comparison to other AI techniques, SMOTE can effectively improve fairness even with a large bias presence- regardless of the AI algorithm [Zhou et al., 2023]. SMOTE works by randomly fabricating a new sample along the line segment between an instance x and one of its random neighbours ($x^(k)$) in the feature space [Zhou et al., 2023]. Specifically, the Fair-SMOTE algorithm balances subgroup data distributions to ensure that the privileged and underprivileged groups have an equal number of positive and negative instances [Zhou et al., 2023]. Therefore, the parameterized data sampling method produces optimal fairness predictions with a small loss in its predictive power. This specific technique is proposed to enhance model fairness and prediction accuracy. It is considered a pre-processing step [Zhou et al., 2023]. It addresses the cause of algorithmic bias that exists within data and favours the privileged group while decisions are made [Zhou et al., 2023]. The results of Zhou et al.'s paper demonstrate that SMOTE has excellent utility and privacy, which is critical in healthcare-related technologies.
7) **Artificial intelligence in healthcare: Complementing, not replacing, doctors and healthcare providers** [Sezgin, 2023]
The use of AI in clinical contexts is revolutionary, as it contributes to improved diagnostic accuracy, better treatment planning and decisions, and improved patient outcomes [Sezgin, 2023]. Along with this comes concerns about AI replacing healthcare professionals. The related paper and this paper focus on the use of AI in healthcare as human-ML (machine learning) augmentation, not replacement. The related paper mentions the human-in-the-loop (HITL) approach, which is centered on having human expertise guide, communicate, and supervise AI systems [Sezgin, 2023]. This approach ensures that AI systems are free of error and bias.

The paper by Sezgin focuses on the benefits of human-ML augmentation and suggests that decisions made by healthcare professionals guided by AI may be/become more accurate than those without AI [Sezgin, 2023]. It suggests that collaborative AI adoption is more realistic than the human replacement of AI and that healthcare organizations should be responsible for ensuring AI tools have undergone rigorous evaluation for safety and effectiveness [Sezgin, 2023]. Also discussed is the potential for legal, infrastructure, privacy, and security implications. One of the focuses of this paper is human-ML augmentation in healthcare, and ethical implications will be further discussed in detail.

8) **Providing Care: Intrinsic Human–Machine Teams and Data** [Russell and Kumar, 2022]
There are very few scholarly works that look at methods for adapting quantitative health data features with the help of human expertise. The related paper proposes an entropy-based construct that is combined with quantitative measures in a critical clinical event (CCE) [Russell and Kumar, 2022]. The methods mentioned would ultimately build trust in AI based clinical decision support systems (CDSS) which would improve human-ML augmentation outcomes [Russell and Kumar, 2022]. Our project focuses on the uses and limitations of AI-ML augmentation in healthcare and clinical (medical student) perspectives on this topic.

III. Methodology

A. About the Datasets

The breast cancer dataset (taken from [Breast Cancer, 2022]) consists of numerous patients, all with breast cancer. The main feature of interest is the one labelled "Survival Months", in which a patient is given several months to live based on the data in this dataset's other features: stages, grades, tumour size, nodes, estrogen, and progesterone status, along with race, age, and marital status, and whether the patient is dead or alive. As for potential bias factors, there are three: race, age, and marital status. While age and marital status do not appear to have much bias, there is a major sampling bias when it comes to race. As mentioned before, 85% of the dataset consists of patients of the Caucasian race, while the remaining 15% is made up of Black patients and "Other" patients, which include Asian American Pacific Islanders (AAPI) and Indigenous Peoples. This may lead the predictive model astray as it may consider race to be a significant factor that plays into the length of survival for a patient. Moreover, this dataset does not contain breast cancer patients of Hispanic origin, which only further proves that there is sampling bias in this dataset. When it comes to preprocessing, because the breast cancer dataset only has entries of patients with breast cancer, it is impossible to train a predictive model to predict whether a patient has breast cancer. So instead the target feature—the output—was set to "Survival Months", and the values, which used to range from 1 to 107, were preprocessed to give binary values instead—1 and 0. Any patient with a 1 as a value for "Survival Months" meant that they were supposed to live longer than the average, and any patient with a 0 meant that they would not live longer than the average. Other preprocessing was done, including dropping the "Marital Status" column as it appeared to have minimal effect on the training model's accuracy, and changing all the categorical data into numerical data using one hot encoding. All this was done in hopes of making training the predictive model easier.

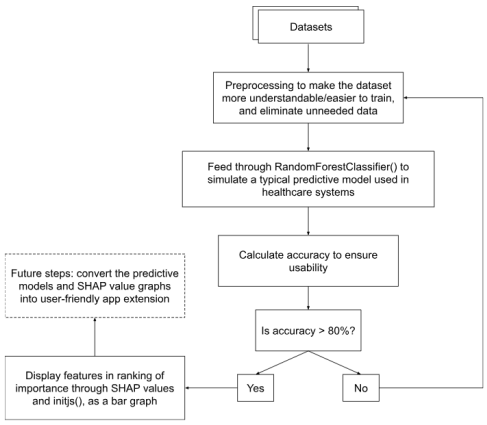

Fig. 1. A simple flow diagram that describes the idea of this project.

The heart attack dataset (taken from [Heart Attack, 2021]), like the breast cancer dataset, consists of many patients. The features include many medical factors: exercise-induced angina, number of major vessels, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting electrocardiographic results, and the maximum heart rate achieved. This dataset was much easier to work with; an "output" feature was already given in binary values, with 1 dictating that the respective patient had a higher chance of getting a heart attack, while 0 meant that the patient had less of a chance of getting a heart attack. Additionally, there are two probable bias factors: sex and age. It appears that both play a fair role in dictating whether a patient is at risk of a heart attack. A sampling bias test concluded that there is indeed a bias in this dataset, with over two-thirds of the dataset being patients of sex "1" and the remaining third of the dataset being patients of sex "0". By taking into account the fact that males are generally much more favoured as medical research clients, it is most likely that sex "1" refers to males and sex "0" refers to females. As for the preprocessing, the dataset had already included numerical data throughout, with a helpful "output" feature encoded in binary values to determine whether a patient had a high chance of getting a heart attack. Therefore no preprocessing was needed for this dataset.

B. Experiment Set-Up

To recall, the purpose of our project is to find ways to mitigate bias in healthcare systems, created by AI. So our thought process was to simulate what a typical predictive model would look like in a healthcare setting, and then use SHAP values to point out any potential bias the model has.

To start, we wanted to create a predictive model using the medical datasets described in the above section. We planned on using RandomForestClassifier() and Logistic Regression to train the predictive models. The reason for using RandomForestClassifier() is that this method essentially splits datasets into factors and makes decisions based on the values of those factors. This works in our favour because we want to highlight certain features of a dataset that may indicate potential bias in predictive modelling. For example, if the model weighs the "value" of a patient's race heavier than the stage of a tumour, there may be a problem with the model's decisions and prediction for how long said patient will survive. As for Logistic Regression, this method is best used when the target output includes binary values, 1 and 0. With the heart attack dataset, the output is conveniently converted into binary values, so we believed that using Logistic Regression could prove to be fruitful.

With both datasets, these two methods yielded similar accuracies, with Logistic Regression displaying a slightly higher accuracy. However, to implement SHAP values, the RandomForestClassifier() method had to be used, as SHAP does not support Logistic Regression as of now. This makes sense, as SHAP values take each feature a dataset offers and rank them based on their correlation strength with the output. For example, if there is a strong relationship between a patient's sex and their level of risk for a heart attack, the model will consider a patient's sex to be of higher importance than something with slightly less of a correlation like one's resting blood pressure. To display these rankings, we used shap.initjs() and shap.summary_plot(), functions that convert SHAP values into a bar graph with the highest-ranking features at the top and the bar lengths representing their average impact on the model's decisions.

C. Evaluation Methods

Considering that our AI implementation occurs within the context of a clinical workflow, an ideal implementation would also involve producing recommendations, results, or suggestions that provide a novel, informative perspective that can be acted upon. Most importantly, this suggestion should work to provide a suggestion that supplements the clinician's expertise and knowledge, leading to a more well-informed final decision made, or a suggestion that effectively benefits clinician time. Moreover, the implementation should somehow better allocate healthcare resources to vulnerable or underrepresented populations. Our main reason for using SHAP values and displaying them in bar graphs is so that healthcare professionals can easily point out whether a factor should be weighed heavier or whether the relation between a bias and a medical condition should be considered as important; if the model's prediction differs from the healthcare professional's decision, they can look at the possible causes of this difference. Most likely they will find that the dataset is inherently biased due to the sampling size of various populations. In other words, this project will be successful if a healthcare professional is able to work well with the predictive model to ensure that bias toward certain populations is mitigated to achieve equity throughout the healthcare system.

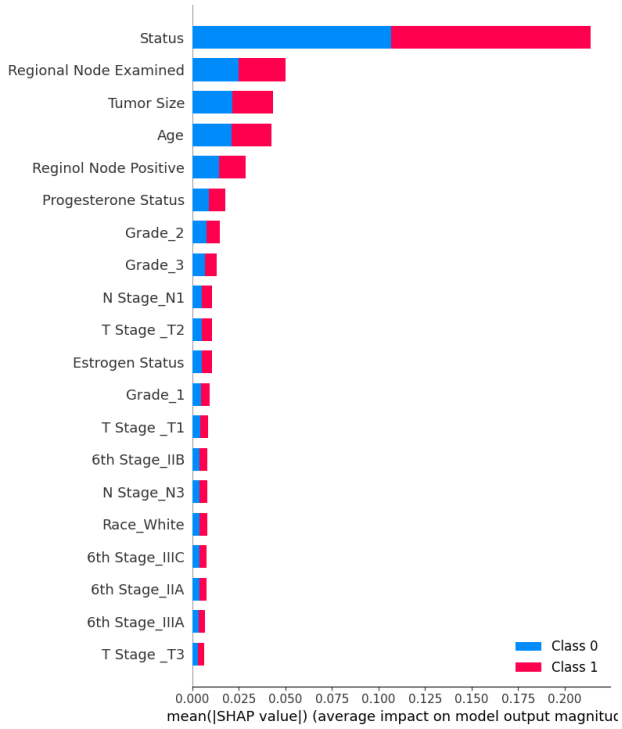

Fig. 2. A summary plot describing the impacts of each feature on the model's prediction for a breast cancer patient's survival months.

D. Replication package

Links to Kaggle are included here.

- The dataset used for breast cancer patients.
- The dataset used for the likelihood of heart attacks.
- The predictive model for the breast cancer dataset.
- The predictive model for the heart attack dataset.

IV. Results and Discussion

As previously mentioned, it appears that using the RandomForestClassifier() method produces the most benefits for our project; with this method, we can implement SHAP values and visualize said values using bar graphs so that healthcare professionals can determine whether there is bias towards a certain patient and take the steps needed to give the patient proper care. Through using RandomForestClassifier(), we were successfully able to create predictive models that mirrored their respective datasets with 85%-89% accuracy. We will further discuss the outcomes of this project in the following paragraphs.

When analyzing Figure 2, it appears that the feature deemed to be most important is the "status" feature. To recall, the "status" feature returns either a value of 1 or 0, with 1 signifying "alive" and 0 meaning "dead". The predictive model takes this into account so much that it is considered the most important feature in the entire dataset—in other words, the model found the strongest correlation between the "status"

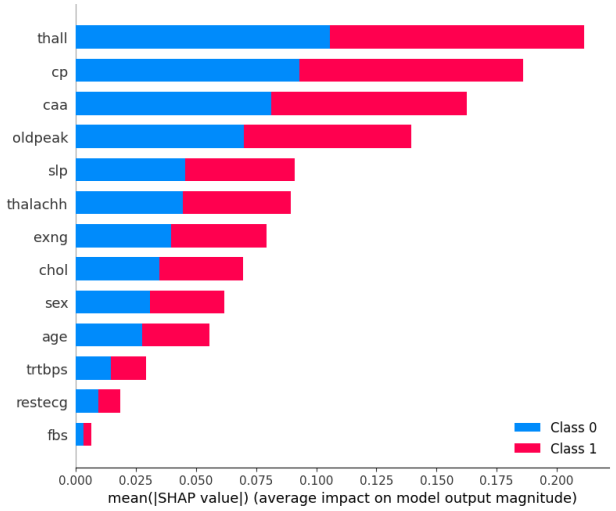

Fig. 3. A summary plot describing the impacts of each feature on the model's prediction for a patient's likelihood of getting a heart attack.

feature and the "Survival Months" target output. This makes sense, as one must be alive to survive for longer than the average survival months, whereas if one was dead they would automatically get a 0 value. While this makes sense, it does not help our project in regards to finding and mitigating biases in healthcare systems. It appears to be too important of a feature to simply take away during preprocessing, however, as the accuracy of our predictive model decreased by a significant amount.

Furthermore, the "Race_White" feature is deemed to be more important than other medical factors such as whether a patient is in their 6th stage and what their T (tumour) stage is. This may be due to sampling bias, as 85% of patients in this dataset are White. One might also realize that "Race_Black" and "Race_Other" do not appear on this graph, as their significance in predicting a breast cancer patient's survival months is considered infinitesimal compared to the rest of the features. This indicates the presence of bias, as covered in the "About the Datasets" section.

Figure 3 displays a similar bar graph to Figure 2, though the significance of each feature appears to be more well-rounded. According to this figure, the most important feature appears to be the patient's thalassemia levels, which are essentially a blood disorder that targets hemoglobin levels. This makes sense, as hemoglobin helps carry oxygen in red blood cells to different parts of the body; if the hemoglobin levels are dangerously low or high, something is bound to happen with the heart.

However, the feature we want to take a closer look at is the "sex" feature. As mentioned before in the "About the Dataset" section, there is an overwhelming difference between sex "1" and sex "0". This indicates a glaring sampling bias in which the males are given more priority and variety over females. This puts females at a disadvantage and therefore they may be at risk of misdiagnosis or improper care and treatment.

A. *Ethical Considerations*

The Kaggle notebooks involved in this project have been linked in the replication section. The methods and strategies used to create the models has been explained to ensure replicability. Therefore the model is transparent for people not involved in the project and the inclusion of the model background in this paper demonstrates its explainability.

Since the highest model accuracy is 89%, there is still room for the model to make mistakes. These mistakes could lead to the misidentification of biases or the possibility of bias being undetected. Until it reaches near 100% accuracy, it is unethical to completely trust the models to do their job. In addition, models need to be supervised and edited by humans (human-ML augmentation) to screen for biases, errors, and utility. The nature of AI in healthcare at the time of this paper does not allow for independent AI models or technologies to be used alone or replace healthcare professionals. Healthcare professionals have the necessary and relevant training to make informed decisions about patient care, and understand the logic behind these decisions. Although AI is advancing, it often does not understand the logic behind healthcare decisions and suggests unhelpful or harmful recommendations based on errors.

Privacy is a critical topic in healthcare. The project suggests an oversampling technique (SMOTE) as one method to mitigate bias. Similar oversampling or synthetic techniques can be used to replace real patient data to protect patient privacy. These techniques would have to have the proper training and supervision to ensure accuracy and usability, however it has promising potential as an alternative method instead of using private patient data. This also addresses the issue of health data piracy and hacking, as flaws in security in AI may lead to private health data being stolen. If the data used is synthetic, it would be a representation of real data and not real-world data, so the real data would be safe from malintent.

The model is incapable of harm, to begin with; it simply takes in data and puts out a visualization of what the data looks like. If the model were to display an unwanted output (i.e. a wrong prediction), it can be handled by the healthcare professionals themselves. After all, this model cannot act on its own decisions and is therefore powerless, so the professionals can choose to not agree with its decisions.

One concern with the model is the potential to create a more equitable system for those physicians and healthcare professionals who use it, and the disparity in health if other physicians do not. In that case, those who use the model would be promoting equitable health conditions while there would be a gap in equity in those who do not use it.

V. Conclusion and Future Work

While we were able to successfully display the data of the predictive models in a comprehensive bar graph stating the supposed importance of each feature, we will need to convert this into a more user-friendly platform, such as an app or a simple desktop extension. Since healthcare is gradually becoming more technology-integrated, it would be efficient

for professionals to be able to analyze a patient's medical condition by simply pulling up an extension to check for potential biases against the individual.

Moreover, to facilitate clinician-AI teaming such that biases such as mistrust in AI can be mitigated to some extent, we would want to implement features such as timers that delay the AI prompt or embedding features that involve justification of the clinician's final decision within the electronic medical record system or dashboard. Doing so may also facilitate future research that conducts natural language processing analysis deriving common themes among clinicians

A. Limitations

While this project would aid those practicing Western medicine, implementing AI in healthcare may not be as acceptable in different cultures, such as the Indigenous and countries in the East. For example, while we believe that using AI and other advanced technology may expand the benefits of healthcare, the Indigenous use a more holistic approach through the use of "ceremonies, plant, animal or mineral-based medicines, energetic therapies and physical or hands-on techniques" (here). Since our project does not cover these aspects of medicine and healthcare, we will not be able to uncover any potential biases these practices may have towards certain groups within the Indigenous.

Furthermore, this project does not find a solution to eliminating these biases in healthcare; instead, it simply addresses them and brings them to light. Ultimately, the healthcare professionals still have to come up with the solutions themselves. For example, in the heart attack dataset, it is clear in the "Sampling Bias" section that there is a significant gap between the two "sex" values, 1 and 0—most likely male and female respectively. Therefore it is implied that not enough research is being put into female anatomy since datasets do not include many female clients. While this project does highlight this inequity, it cannot take away this inequity. In the end, healthcare professionals still must acknowledge these biases and find ways to mitigate and eventually eliminate these biases.

VI. Acknowledgements

The team would like to give a special thanks to Aghia Mokhber, Alice Li and Bonnie Yang, three medical students at the Queen's School of Medicine who consulted their perspectives on this project.

References

[Awan, 2023] Awan, A. (2023). An introduction to shap values and machine learning interpretability. *DataCamp*.

[Breast Cancer, 2022] Breast Cancer (2022).

[Gao, 2021] Gao, A. (2021). Getting to the root of data bias in ai. *BCGAMMA*.

[Heart Attack, 2021] Heart Attack (2021).

[Henry et al., 2022] Henry, K. E., Kornfield, R., Sridharan, A., Linton, R. C., Groh, C., Wang, T., Wu, A., Mutlu, B., and Saria, S. (2022). Human–machine teaming is key to ai adoption: clinicians' experiences with a deployed machine learning system. *npj Digital Medicine*, 5.

[Maslej et al., 2023] Maslej, M. M., Kloiber, S., Ghassemi, M., Yu, J., and Hill, S. L. (2023). Out with ai, in with the psychiatrist: a preference for human-derived clinical decision support in depression care. *Translational Psychiatry*, 13:1–9.

[Mittermaier et al., 2023] Mittermaier, M., Raza, M. M., and Kvedar, J. C. (2023). Bias in ai-based models for medical applications: challenges and mitigation strategies. *npj Digital Medicine*, 6:1–3.

[Norori et al., 2023] Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., and Tzovara, A. (2023). Addressing bias in big data and ai for health care: A call for open science. *Patterns (N Y)*.

[Russell and Kumar, 2022] Russell, S. and Kumar, A. (2022). Providing care: Intrinsic human–machine teams and data. *Entropy*, 24:0–1369.

[Schinkel et al., 2023] Schinkel, M., van der Poll, T., and Wiersinga, W. J. (2023). Artificial intelligence for early sepsis detection: A word of caution. *American Journal of Respiratory and Critical Care Medicine*.

[Sezgin, 2023] Sezgin, E. (2023). Artificial intelligence in healthcare: Complementing, not replacing, doctors and healthcare providers. *DIGITAL HEALTH*, 9.

[Shi, 2024] Shi, M. (2024). Bc_detection_qmind.

[Sikstrom et al., 2023] Sikstrom, L., Maslej, M. M., Findlay, Z., Strudwick, G., Hui, K., Zaheer, J., Hill, S. L., and Buchman, D. Z. (2023). Predictive care: a protocol for a computational ethnographic approach to building fair models of inpatient violence in emergency psychiatry. *BMJ Open*, 13.

[Sikstrom et al., 2022] Sikstrom, L., Maslej, M. M., Hui, K., Findlay, Z., Buchman, D. Z., and Hill, S. L. (2022). Conceptualising fairness: three pillars for medical algorithms and health equity. *BMJ Health Care Informatics*.

[Zhou et al., 2023] Zhou, Y., Kantarcioglu, M., and Clifton, C. (2023). On improving fairness of ai models with synthetic minority oversampling techniques. *SIAM*.