Don't Go Breakin' My Heart:
Fairness and Transparency in Automated Heart Disease Diagnostics

1017 Responsible Data Science
Pedro Galarza and Jack Epstein

I. Background: general information about your chosen ADS

The advent of large scale data collection and storage has given way to a new era in information that is revolutionizing a variety of fields. Due to the sheer volume and availability of data many organizations seeking quantitative solutions have begun employing Automated Decision Systems, or ADS's. There is an obvious intersection between data science and the healthcare industry, and ADS's are already being leveraged to predict diseases, make public health decisions, assess patient risk, allocate specialized care, and much more. While the promises of AI in medicine are exciting, there are numerous risks emerging from these systems. Problems of bias, fairness, equitability and privacy–already rampant within our healthcare system–have the potential to be amplified by ADS systems capable of making high-impact decisions about individuals' health at scale [1,2].

One particular application of ADS's that has received much attention and research is machine learning models for the diagnosis of cardiovascular diseases. Cardiovascular diseases are the leading cause of mortality around the globe according to the WHO, with coronary artery disease representing a large portion. Methods for diagnosing coronary artery disease can often be invasive and risky. ADS systems have the potential to provide an accurate, accessible, and non-invasive diagnostic method for CAD. There is a rich literature exploring machine learning techniques that have been developed around data sets for exactly this purpose [3,4].

One of the foundational data sets on which many early CAD machine learning diagnostic systems implemented is the UCI heart disease data set released with the purpose for pioneering predictive models [5]. The data set has 303 patients with 76 associated features, with the "goal" feature being a binary indicator of heart disease. Only 14 medically relevant features have been released, the redacted features are mostly medical attributes less relevant to CAD, but also include identifying information like social security numbers and names. One feature that is conspicuously missing from both the original data set and the redacted on is race. Futhermore the attribute "sex" present in released data set which exposes a natural tension between a feature's utility and its protected status.

In this paper, we seek to perform a case study of a publicly available prediction system called "12 ML Models + Visualization (92% Accuracy)" which employs an array of standard machine learning techniques to predict CAD using the UCI dataset much like an ADS would [6]. While perhaps not as sophisticated as some of the academic models, the notebook is implementable locally and allows for the exploration of the benefits and drawbacks of these diagnostic ADS systems. Furthermore, because the notebook uses the industry standard scikit-learn packages, our analysis seeks to show the advantages and drawbacks of these out of box methods which are implemented every day on similar problems.

As hinted, we intend to consider sex as a protected class and explore how men and women maybe treated differently by this classifier. There is an inherent tradeoff between accuracy and fairness that is especially contentious in healthcare domains. Sex as feature contains important and relevant medical information for an accurate diagnosis, however in the context of machine learning its status as a protected class may be overlooked and lead to classifieres with large accuracy gaps between the sexes. We're hoping that this ADS will be a useful test case in exploring this tension and reveal some of the risks diagnostic automated decision systems pose toward protected classes.

Furthermore, diagnostic ADS systems in practice are usually deployed under the guidance of medical professionals. For this reason it's not only important for these systems to be accurate, but also trustworthy and accessible to their users (doctors, nurses, etc). By implementing different transparency methods and metrics we hope to also further understand the relationship between complexity, fairness, interpretability, and accuracy.

II. Input and output

As mentioned above, the data for this ADS comes from the UCI Machine Learning Repository, where it undergoes most of the aggregation, anonymization and cleaning. Initially the data comes from hospital patients from a combination of the following: Hungarian Institute of Cardiology, University Hospital in Zurich, University Hospital in Basel and V.A. Medical Center in Cleveland [5]. As mentioned, the original dataset has 76 attributes including personally identifying information (PII), however this has all been removed before the ADS owner accesses this data.

The 13 remaining input features are listed in Table 1 below. All of the categorical features have been numericized and there are no missing values in the dataset once it reaches Kaggle.

Feature Name	Description	Input Space	Mean	Range
age	Patient's age	R	54.37	29-77
sex	Patient's sex, M=1	{0,1}	0.68	0-1
cp	chest pain type (0: typical angina, 1: atypical angina, 2: non-anginal pain, 3: asymptomatic)	{0,1,2,3}	0.97	0-3
trestbps	resting blood pressure (in mm Hg on admission to the hospital)	R	131.62	94-200
chol	serum cholesterol in mg/dl	R	246.26	126-564
fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)	{0,1}	0.15	0-1

restcg	resting electrocardiographic results: 0: normal, 1: having ST-T wave abnormality, 2: showing probable or definite left ventricular hypertrophy	{0,1,2}	0.53	0-2
thalac	maximum heart rate achieved	R	149.65	71-202
exang	exercise induced angina (1 = yes; 0 = no)	{0,1}	0.33	0-1
oldpeak	ST depression induced by exercise relative to rest	R	1.04	0.0-6.2
slope	the slope of the peak exercise ST segment. 0: upsloping, 1: flat, 2: downsloping	{0,1,2}	1.40	0-2
ca	number of major vessels (0-3) colored by fluoroscopy	{0,1,2,3,4}	0.73	0-4
thal	blood disorder called thalassemia (0=NULL, 1=fixed defect, 2=normal blood flow, 3=reversible defect	{0,1,2,3}	2.31	0-3

Table 1: List of all features

The target variable in this dataset is binary, with the positive class representing a diagnosis of angiographic disease. More specifically, someone is classified as positive for the disease if their artery diameter has narrowed by more than 50%. The ADS solely outputs the binary prediction {0,1}, however, the underlying code is capable of producing probabilistic outputs, which will help during the evaluation stage. In practice, when this ADS predicts a positive result, this is an indicator for someone to seek immediate medical help. In terms of distribution, this is a relatively balanced dataset, with 54% of the instances in the positive class. This base rate will change considerably when crosstabbed with certain target variables, as we will explain below.

Along with the issue of limited sample size, it is important to note that this dataset is not representative of the general population and while the output of the ADS is to be predictive for all people, the inputs are biased in many ways. First, the data used is of hospital patients meaning these are people who selected to get medical treatment when they felt they needed it. While this could potentially bias towards people who have access to care, the more obvious issue is that this biases the dataset towards people who either have heart problems or feel unwell enough to get treated for heart problems.

This self-selected group leads the data to skew in other, more tangible ways. As we can see in the initial plots [see Figure 3], this dataset skews much older than the general population. The median age is 55, while the global median age is 30 and the median age in the United States is 38, both far lower than our dataset [7]. Another clear sign this data does not represent the general population is the prevalence of chest pain. While a rough estimate, about a quarter of adults experience chest pain [8], which is far lower than the 92.5% experiencing some sort of pain in this dataset. This is unsurprising, as chest pain is likely one of the primary reasons why patients seek medical attention. Perhaps most important given the context of our paper, is that

this dataset is over two-thirds male. The US and global population is just about half male/female, so this does not properly represent the general population. Despite common misconceptions, heart disease is quite prevalent in women [9] and therefore we explore whether such a system puts women at a disadvantage as they are underrepresented in this dataset.

We can view these distributions overlaid with the target variable to get additional key insights into this dataset. We can see in Figure 2 that while the dataset skews more heavily male, this difference is much more extreme with the negative class distribution than with the positive class distribution. This is caused by differing base rates between the sexes, where men in this dataset have heart disease 45% of the time, compared to 75% for women. In the same figure, we see an interesting phenomenon with chest pain, where the positive cases over-index in having non-anginal chest pain, while the negative class over-index in typical angina. This may seem counterintuitive, but could be explained by people with chest pains being more likely to seek medical help yet a typical angina is different from our target variable.

In Figure 4, we see the pairwise correlations of all continuous features, overlaid with the target variable. While not a specific insight around correlation, we see another interesting phenomenon around the age distributions, which is further highlighted in Figure 3. While conventional wisdom would say that the positive class should skew older, we see the opposite -- the negative class clearly skews older, while the positive class has a relatively more normal shape. This is likely due to the fact that older people are more likely to suffer from heart disease so are also more likely to seek out preventative care.

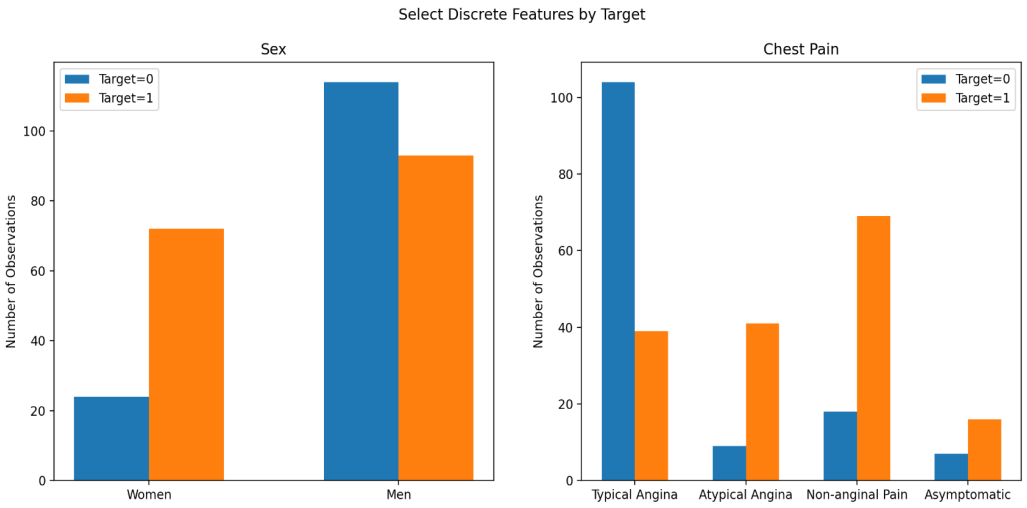

Figure 2: Distributions for Sex and Chest Pain

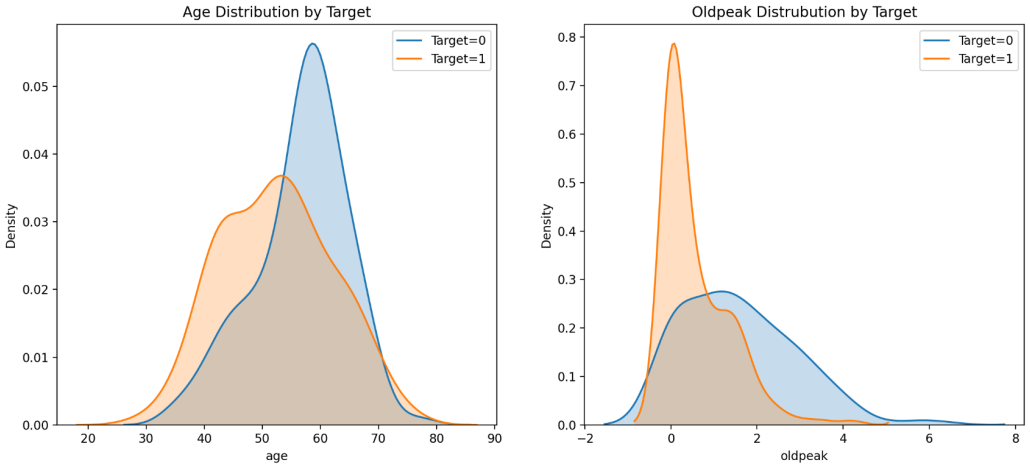

Figure 3: Distributions for Age and Old Peak

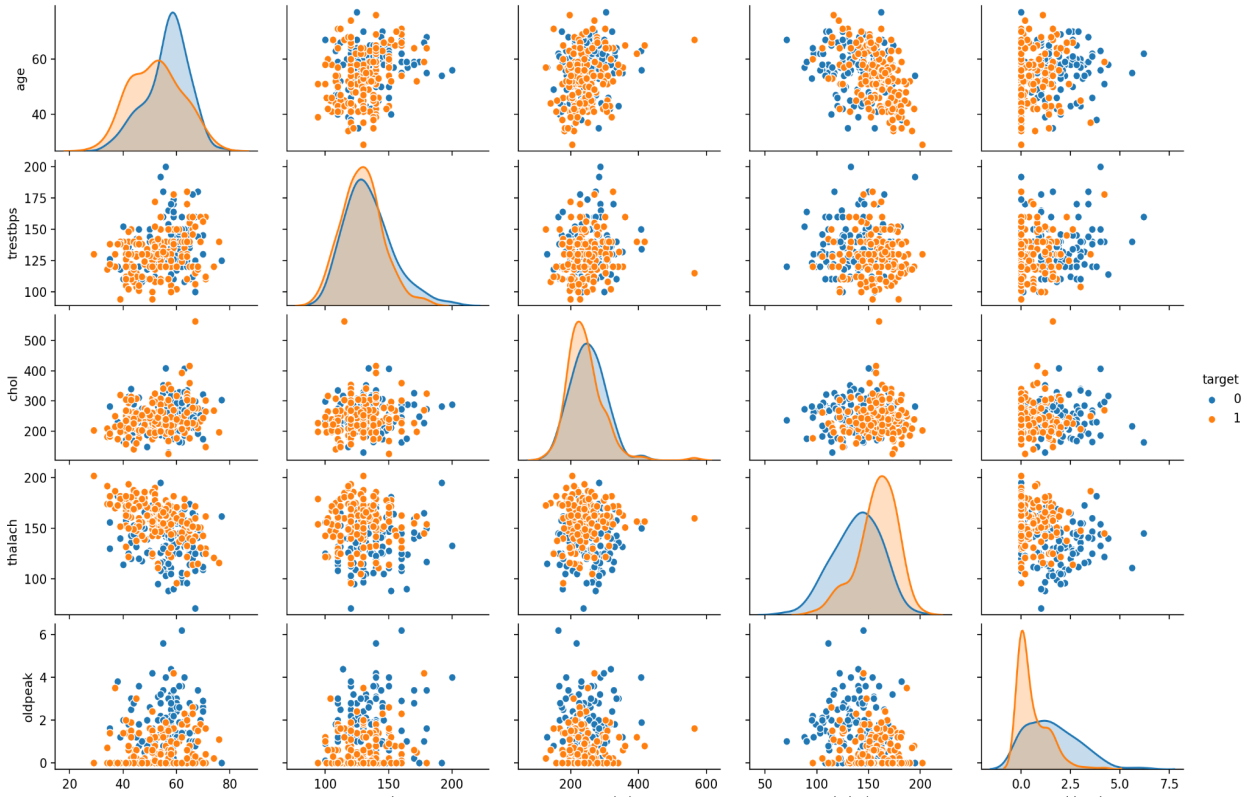

Figure 4: Pairwise Distributions for all Continuous Features

III. Implementation and validation

This ADS appears to take a more minimalist approach to the machine learning process. The owner takes the cleaned data from the UCI dataset and does very little additional cleaning and preprocessing. The dataset enters the system with no missing values and the chosen 13 features numericized. The system owner applies a pre-processing technique of normalizing the data, however does not go through the proper methodology here. Rather than normalizing features, the owner normalizes the entire dataset. Additionally, the owner scales all data together before splitting into training, testing and validation sets, which affects the independence of the three splits.

This has two key effects. First, any features that need to be scaled relative to each other, like oldpeak and trestbps, do not get appropriately rescaled. Fortunately, this does not hurt model performance significantly as the reported accuracy is still quite strong. Second, this changes the input space of certain features. In the most important case, this changes sex from a binary feature to a continuous feature. Luckily for us, sex=0 remains at 0, therefore we are able to map this feature back to it's original form when using tools such as AIF360, which require certain attributes be binary.

In terms of model implementation, the ADS owner keeps it relatively simple and focuses more on testing various families of functions rather than fine-tuning a specific model. After the scaling, as mentioned above, the owner splits the data into training, testing and validation data with a 66.67%, 16.67%, 16.67% split between the three. The owner then tests various types of models, from linear and tree based models, to neural nets, using almost exclusively the out-of-bag hyperparameters. There is no hyperparameters testing in this ADS, meaning the owner does not take full advantage of having both testing and validation data. The owner essentially treats these two datasets as independent test sets used for scoring model accuracy. We exploit this decision in order to make our fairness and transparency analyses more robust by combining these two datasets.

The ADS owner ultimately chooses the winning model based on accuracy of the test set. While recall, precision and F1 score are all included in the ADS [see Table 5], the winning model -- the Multilayer Perceptron Neural Net -- is chosen based on overall test accuracy. It is worth noting, however, that this choice is somewhat arbitrary, as this model does not perform as well on the validation set, and therefore the strong accuracy may be due in some part to the small size of the test set. For the sake of consistency, it is the declared winner by the ADS owner, therefore we choose this model for all fairness and transparency analyses.

Model**	Test Accuracy	Test Recall	Test Precision	Test F1
Neural Net*	**0.92**	**0.857**	**1.00**	**0.923**
Gaussian Naive Bayes	0.86	0.84	0.875	0.857
Bernoulli Naive Bayes	0.82	0.80	0.833	0.816
Multinomial Naive Bayes	0.48	0.48	1.00	0.649
Logistic Regression	0.60	0.55	0.917	0.687

Decision Tree	0.78	0.76	0.792	0.776
Random Forest	0.90	0.828	1.00	0.906
Gradient Boosted Trees	0.88	0.80	1.00	0.889
XGBoost	0.88	0.80	1.00	0.889
Support Vector Machine	0.66	0.595	0.917	0.721
K-nearest Neighbors	0.68	0.633	0.792	0.704
SGDClassifier - SVM	0.78	0.741	0.833	0.826

Table 5: Performance Metrics by Model

* Winning Model -- used for ADS

**All models use standard SKLearn parameters

IV. Outcome: Fairness and Performance Tradeoff

Introduction

There is an inherent tradeoff between overall model performance and the model fairness for all ADS systems, notwithstanding how this tradeoff is balanced is highly application specific. In the attached notebook [1017_FinalProject_Fairness.ipynb] we explore how this tradeoff manifests itself across several metrics for male and female sub populations. We are interested in how this model performs across 3 different evaluation frameworks. (1) First we evaluate the original model using the designated test data. (2) Next we notice that the author never utilizes their validation data during training so we append it to the test data set in order to build a larger evaluation set for more robust holdout metrics. (3) Lastly we explore how a disparate impact remover algorithm changes the metrics on the larger hold out set. Three different metrics were chosen to evaluate the performance of the model—False Negative Rate, Accuracy (0-1), False Positive Rate [see Table 6 for descriptions]. To evaluate the fairness of the model we compute the disparity between males and females for each of these metrics, we also measure the disparate impact of the model for these groups. Below we outline how each of these metrics should be considered within our diagnostic framework. Each model's performance for each of these metrics is described in Table 8.

Metrics and Results

Metric	Description
False Negative Rate	This is the key performance indicator for most diagnostic models. The key negative impact from this model is misclassifying a positive instance of CAD. The cost associated with undiagnosed CAD is much higher than misclassified negative instances. The harm

	incurred upon a particular subgroup can thus be described by this metric's disparity between the privileged and unprivileged groups.
False Positive Rate	In a diagnostic application this is probably the least important error. While the cost of diagnosing someone with CAD when they are healthy is non-zero, it is far from being outweighed by a false negative. A good model should seek to minimize this metric (and it's disparity across sub-groups) but it should not be prioritized over a good FNR.
Accuracy	This is the standard 0/1 loss, or the percentage of correct guesses. This is the KPI for which the owner of this ADS used to evaluate the model. For our purposes it is a useful general metric, however it should be baselined by the distribution of targets.
Disparate Impact	This metric describes how represented each protected class is in the model's prediction. The Ideal Disparate impact score of 1 indicates both classes are equally represented. For ADS systems where there is no verifiable contrapositive and the target assignment is an opportunity (loan, college acceptance) this a good metric for fairness. In our case, however, our target is a verifiable attribute and representation in the predicted positive outcome is not the primary way subjects are impacted by this model.

Table 6: Key metrics and descriptions

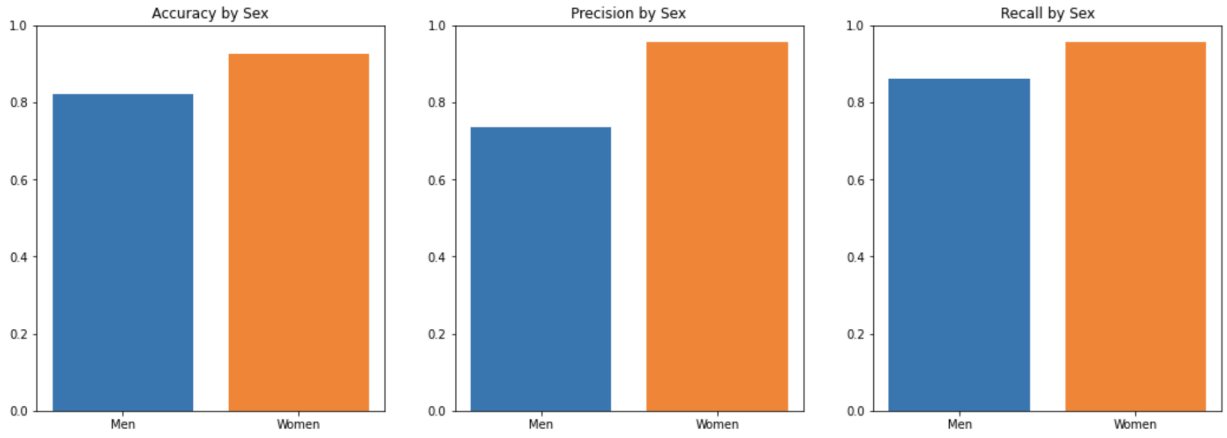

Figure 7: Performance Metrics by Sex

Model Framework	Train/Test	Train/Hold-out	Disparate Impact Remover
Overall Accuracy	0.92	0.85	0.83
Male Accuracy	0.917	0.822	0.795

Female Accuracy	0.929	0.926	0.926
Disparate Impact	1.929	1.829	1.636
Overall FPR	0.154	0.208	0.271
FPR Difference	0.203	0.0455	0.023
Overall FNR	0.0	0.096	0.077
FNR Difference	0.0	-0.094	-0.060

Table 8: Fairness Metrics

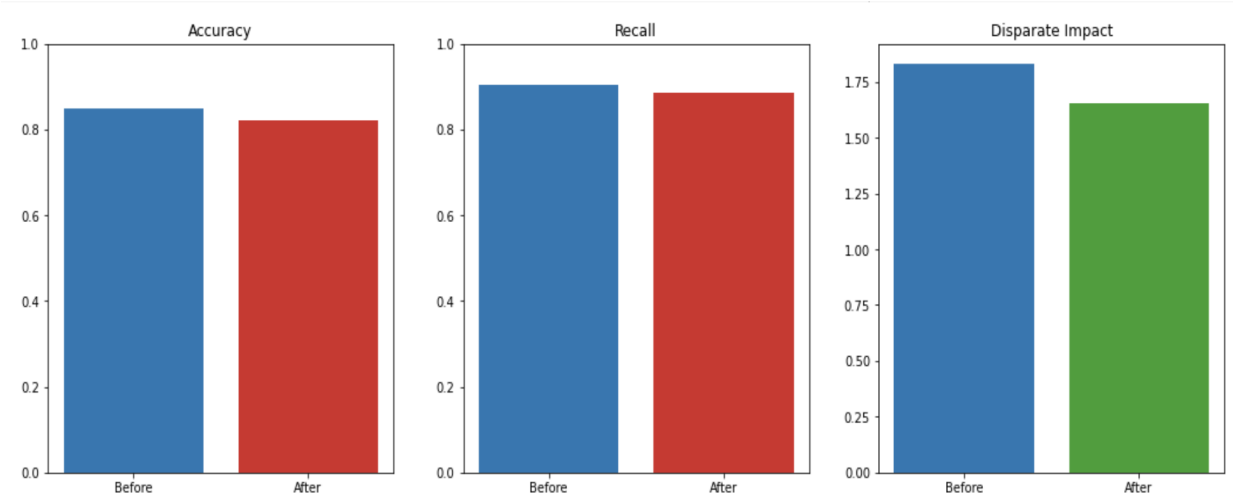

Figure 9: Performance and Fairness Updates Using Disparate Impact Remover

Discussion

The first thing to note is that overall accuracy of the model decreases when evaluated on a larger hold out set. This drop in performance is consistent across all metrics except FPR disparity (which is negligible in our context). This suggests the first flaw in the ADS: the training data and learning pipeline was not sufficient to build a generalizable model. The author's published statistics with the smaller test set can be thought of as inflated, because of this we will move forward by only considering the evaluation results on the larger hold out set.

For the metrics prioritized (FNR and accuracy) there is a considerable disparity between men and women. Women have higher accuracies and lower false negative rates than men. But does this constitute a biased automated decision system? Despite metric disparities that affect patient outcomes it is difficult to call this ADS unfair and it's even more difficult to correct the disparities that exist. Here sex exists as both a protected class and a biological feature with real correlations with the target. In a medical context sacrificing accuracy for fairness is not recommended. If, hypothetically, this model was less accurate for women or had a higher FNR for women due in part to a smaller sample size, then this would be problematic, as women would be further at risk due to a sampling bias. Fortunately, this is not the case.

Not only, should performance be prioritized in this tradeoff but we also show that fairness processing techniques do little to improve the disparity in the model. In the attached collab notebook we demonstrate a baseline implementation of the Disparate Impact Remover algorithm on the data set and evaluate the results on the larger hold out data set. While we see a considerable improvement in the disparate impact ratio, and a small improvement of the FNR disparity, this comes at the cost of both overall accuracy and male accuracy. As outlined earlier disparate impact is not the metric we seek to optimize and this compromise is not advantageous.

While we argue that prioritizing fairness in model construction is not conducive to the application, we insist that fairness metrics are crucial in elucidating some of the disparities that exist in the model. Transparency about how these models are biased and how decisions are made can ultimately lead to corrective implementations of these decisions. This is especially true in health care diagnostic settings where the expertise of a medical professional is often combined with an ADS system to make conclusive decisions.

V. Outcome: Transparency

Introduction

For a diagnosis as important as heart disease, accuracy and fairness are not enough. We need to understand the mechanics behind any prediction, both to build faith with the patients and to audit any disparities. In the attached collab notebook [1017_FinalProject_Transparency_SHAP.ipynb], we use the SHAP explainer to derive both global and local explanations for the ADS, as well as propose a methodology to improve the model after extracting information about features from our misclassified examples. We then reevaluate the ADS using the same performance and fairness metrics mentioned previously.

Metrics and Results:

We must use the SHAP explainer on a subsample (n=75) of our training set due to the complexity of the model, but it is sufficient to glean the general mechanics of the model. A superficial analysis of the SHAP plot gives a fair amount of confidence in the model. Feature values are correlated positively or negatively in a consistent manner with SHAP values for most instances as indicated by the gradation of the feature values.

Following this bird's eye view analysis, we can zoom into a specific instance showing the full decomposition of SHAP contribution. In Figure 10 we display the decomposition for a female who was given a false positive assignment. In the collab notebook, the decomposition for several other individuals is displayed. It is immediately noticeable that features commonly making the largest contributions in these examples are 'oldpeak,' 'ca,' and 'cp.' Though not necessarily in the same direction each time.

Unfortunately because the ADS we are investigating uses scikit-learn's MLP classifier, there is no feature importance method to determine the global weight of each feature. Thus for a more systematic approach to determining the feature importances we take the sum of the SHAP values across classifications. We can extend this logic to determine the feature most

responsible for the misclassification of individuals by applying the method to a misclassified subpopulation of interest.

In Figure 11 and Figure 12 we show the relative importance of each feature globally and for the negative misclassification of males, respectively. The features 'oldpeak,' 'ca,' and 'cp' all contribute to global classification, as well as true negative classification. However the "oldpeak" feature plays a more significant role in the misclassification of males than in the global classification. Using this we propose a refinement of this model by removing the 'oldpeak' feature from training. This transparency guided model rescues 2 false positive miss classifications for men. So while the method does not yield the expected improvement of the false negative rate disparity, we do see metric improvement (Figure 13, Table 14) and demonstrate the methods feasibility.

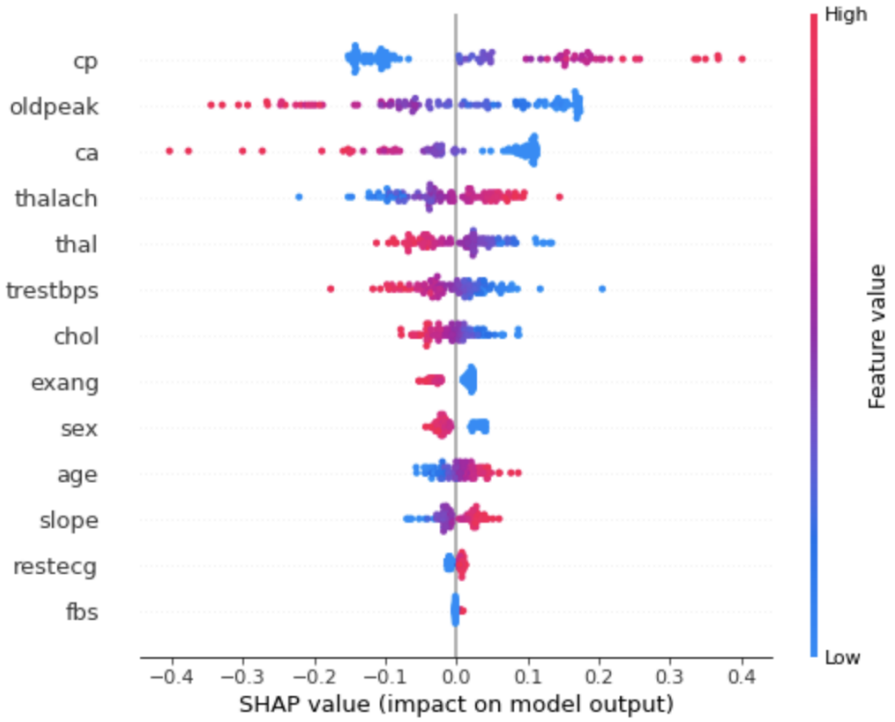

Figure 9: SHAP Values for individual data points across features.

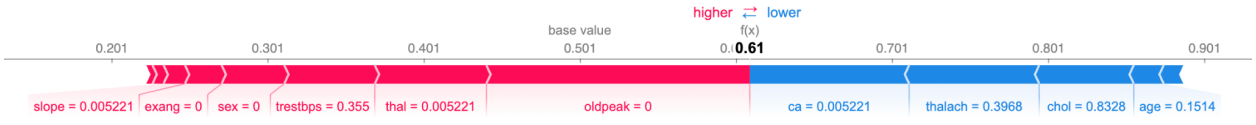

Figure 10: SHAP Explainer Decomposition for individual Case. Here the individual is a female given a false negative classification.

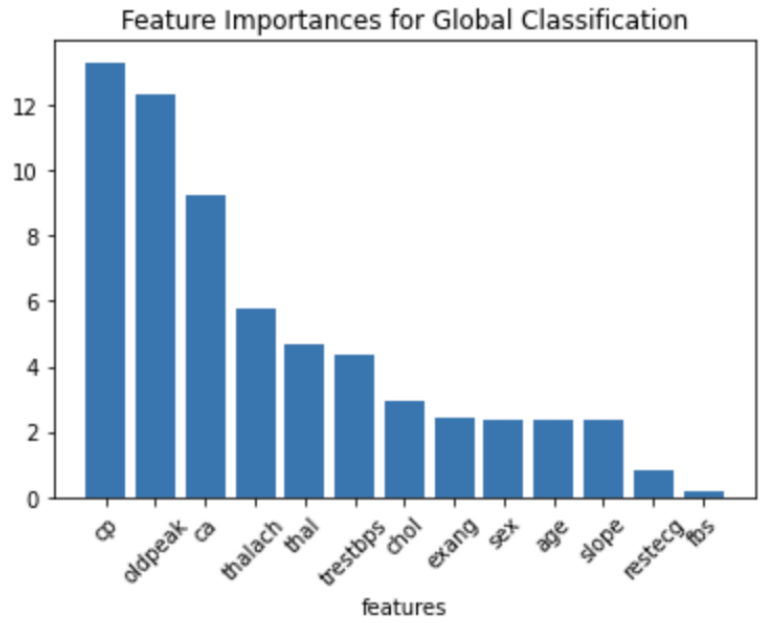

Figure 11: Feature Importances as Determined by the total sum of feature SHAP magnitudes for all test points.

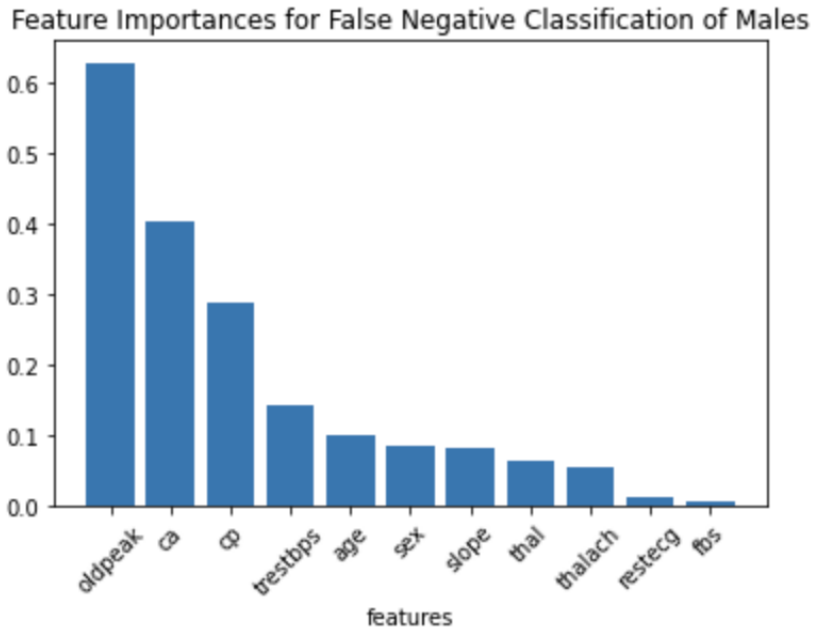

Figure 12: Feature Importances as Determined by the total sum of feature SHAP magnitudes for all male test points with false negative classifications.

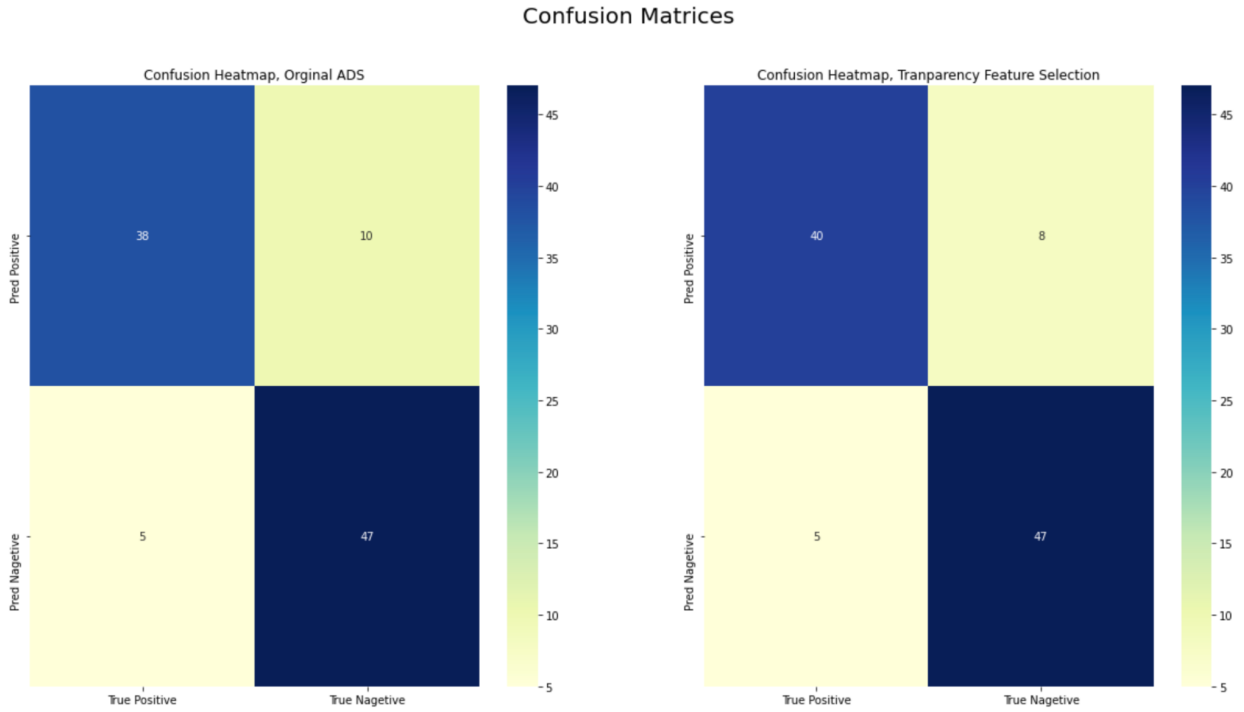

Figure 13: Heat map representation of confusion matrices for the ADS results on the hold out set and the Transparency Guided ADS.

Model Framework	**Train/Hold-out**	**Transparency Guided Feature Selection (no 'oldpeak')**
Overall Accuracy	0.85	0.87
Male Accuracy	0.822	0.849
Female Accuracy	0.926	0.926
Disparate Impact	1.829	1.67
Overall FPR	0.208	0.167
FPR Difference	0.0455	-.18
Overall FNR	0.096	0.96
FNR Difference	-0.094	-0.0165

Table 14: Metric comparison between original ADS results and Transparency Guided ADS system.

Discussion

Because diagnostic ADSs are employed by healthcare professionals and affect real patient outcomes transparency methods are indispensable. First health care professionals need to be able to trust the model to use effectively, being able to explain which features contribute to classification globally, as well as local predictions allow these professionals to

cross reference their own rule-based methods for an augmented diagnostic work flow. Secondly, when patient well being is at stake it is crucial for recipients to be able to audit the results from these systems in case there are concerns.

While the model selected by this author of this system achieves the highest accuracy as compared to the other 11, it is also the least interpretable. There is a valid discussion to be had about what is the appropriate tradeoff between transparency and accuracy of a model. Can a loss in model accuracy in exchange for transparency be rescued by a more effective implementation? This should be decided on a case by case basis, but it is clear that serious efforts should be made to bridge the gap between automated systems and rule-based ones.

In our exploration we provide baseline transparency methods for the ADSin question. A health care professional can decide whether feature importances and misclassification trends are supported by current theories, and interpret the classification accordingly. Additionally building these transparency models allows users to incorporate their individual knowledge of features not considered by the data set. For example this ADS does not use race, family history, or marital status–all features that are correlated with specific health outcomes–in its predictions. A physician presumably has access to this information and can incorporate the general ADS classification into a holistic and individualized health plan.

VI. Summary

Ultimately, the intention of the data set seems to allow for the exploration of baseline ML models for diagnostic application. For this purpose, the data is appropriate for the task at hand. While we argue there are issues with the sample size, the chosen features offer a good variety, ranging from demographic data to more specific measurements of heart health.

When considering the questions of fairness and accuracy, it is worth acknowledging which stakeholders benefit from improvements in each metric. Accuracy is the most straightforward -- both the ADS owner and society benefit from strong performance here. With a strong accuracy, the ADS owner can point to the quality of the model to help build trust. More importantly, it is vital that patients who are receiving diagnoses from this tool need the confidence that this prediction is correct. However, on this note, we also argue patients stand to benefit even more from strong recall. As we discussed above, patients need to know when they need help. If a patient goes to get checked and they are told they do not have heart disease, this needs to be correct. A false negative gives a false sense of security and can lead to deadly results. When referring to fairness metrics, we treat women as the protected class, however, optimizing towards a more optimal disparate impact score or towards more even accuracy between the sexes would benefit men more so than women, as this ADS is more accurate and has higher recall for women.

Taking into consideration the broader societal impact of such a system, we can answer two basic questions. First, whether this system is ready to be deployed in a real-world setting and second, whether we believe more broadly that heart disease is a problem suited for a machine learning approach. We argue this current system should not be deployed in public. While it passes early checks regarding both accuracy and fairness, we have concerns about the robustness of the model and its overall broader implementation. Our concerns boil down to two major areas: sample size and model building.

Given how small the dataset already is, it gets even smaller when overlaying with features like sex. This is an issue with the given test set of 50 instances, and also our bigger scoring dataset which still only has 100 instances. This means there is an inherent variance in all metrics, both in terms of fairness and accuracy which need to be taken with a grain of salt. Due to this, we argue that while the implementation of this system is both accurate and fair, it is not robust. This is most clear when we look at the accuracy score on our combined test+validation set which falls to 85%. One option the ADS owner could have taken to try and mitigate this problem would be through cross-validation, where all data points are used to train and evaluate multiple iterations of a model. This is generally a good approach when test data is limited, which could lead to variance in scores and predictions.

The sample size alone would be enough to give us pause into releasing this ADS, however, the previously discussed issues around pre-processing and lack of hyperparameter tuning give us pause as well. Perhaps the owner was satisfied with the initial accuracy scores, yet we would recommend further testing before we commit to such a system. Plus, as discussed in the transparency section, Neural Nets are highly complex and not interpretable. If the ADS owner had tuned a more interpretable model and achieved similar accuracies, we argue this would be more fit for deployment.

With all of that being said, we argue there is an appropriate space for machine learning in the world of medicine. Systems of this sort have the potential to revolutionize medical treatment and accessibility and there is a societal benefit to building systems like these that can complement current diagnostic methods. These systems can solve issues of scalability and human error, and in an ideal setting, can give people a diagnosis directly at home without needing to go and pay to see a doctor. With the right balance of responsible machine learning efforts and human oversight, this is a problem that can be vastly improved by the technology we have at hand.

VII. References

1. Burger, Mitchell. “The Risk to Population Health Equity Posed by Automated Decision Systems: A Narrative Review”. *Cornell University,* 2020, https://arxiv.org/abs/2001.06615
2. Monteith, S., Glenn, T. Automated Decision-Making and Big Data: Concerns for People With Mental Illness. Curr Psychiatry Rep 18, 112 (2016). https://doi.org/10.1007/s11920-016-0746-6
3. Roohallah Alizadehsani, Moloud Abdar, Mohamad Roshanzamir, Abbas Khosravi, Parham M. Kebria, Fahime Khozeimeh, Saeid Nahavandi, Nizal Sarrafzadegan, U. Rajendra Acharya, “Machine learning-based coronary artery disease diagnosis: A comprehensive review”. *Science Direct*, 2019. https://www.sciencedirect.com/science/article/abs/pii/S001048251930215X
4. Jesmin Nahar, Tasadduq Imam, Kevin S. Tickle, Yi-Ping Phoebe Chen, “Association rule mining to detect factors which contribute to heart disease in males and females”, *Science Direct*, 2013. https://www.sciencedirect.com/science/article/pii/S095741741200989X
5. Aha, David W, “Heart Disease Dataset - UCI”, *UCI Machine Learning Repository*. 2018. https://archive.ics.uci.edu/ml/datasets/Heart+Disease
6. Cal, Baris, “12 ML Models + Visualization (92% Accuracy)”. *Kaggle,* 2019, https://www.kaggle.com/bariscal/12-ml-models-visualization-92-accuracy
7. “World Population Prospects 2019”. *United Nations*, 2019, https://population.un.org/wpp/Download/Standard/Population/
8. Germany, Judy. “6 Facts About Chest Pain”. *Rush University*, 2021, https://www.rush.edu/news/6-facts-about-chest-pain
9. “Women and Heart Disease”. *Center for Disease Control (CDC),* 2020, https://www.cdc.gov/heartdisease/women.htm