Sparse Feature Interactions for Interpretable Healthcare Decision-Making

James Enouen,[1] **Yan Liu,** [1]

[1] University of Southern California - Department of Computer Science
enouen@usc.edu, yanliu.cs@usc.edu

Abstract

In the present day, machine learning and artificial intelligence are applied to nearly every aspect of our daily lives. As these fields are increasingly embedded within consequential fields like self-driving cars, recidivism prediction, credit loaning, and healthcare advising, it becomes increasingly clear that criteria beyond accuracy are required for these multifaceted applications. It has become ever-pressing to address the needs of interpretability, robustness, safety, and fairness for these decision-critical applications. Often interpretability is seen as a cornerstone to these auxiliary requirements by allowing a practitioner to gain insights into the model's decision-making.

This work focuses on developing an interpretable model which blends the inductive biases of a deep neural network with the statistical guarantees of traditional machine learning methods. This new model not only achieves better performance on the MIMIC dataset than the other interpretable methods but also the powerful machine learning methods like kernel machines, neural networks, and boosting machines. We additionally apply this methodology to a causal inference dataset where we can perform scientific discovery on the underlying data in conjunction with our model interpretations. We look at how causal models could offer a better paradigm for studying the impact of algorithmic decision-making and show preliminary results investigating fairness considerations within this framework. We hope this work can inspire future work at the intersection of causality, interpretability, and healthcare.

1 Introduction

Over the past decade, deep learning has achieved significant success in solving challenging AI problems like computer vision, language processing, and game playing (He et al. 2015; Vaswani and et al. 2017; Silver and et al. 2017). As more AI models serve as important solutions and advisory tools for critical decisions with serious downstream consequences, we are left with a series of auxiliary goals to confront: interpretability, robustness, safety, and fairness. These goals are extremely salient in the domain of healthcare when we put the livelihood of an individual human being under the care of a machine learning system. Often interpretability is seen as an important proxy for these goals by allowing a human auditor to review the algorithmic reasoning and compare its logic to these supplementary objectives. In this way, interpretability is viewed as the cornerstone for nearly all of these complementary goals of machine learning.

Many of the works in interpretable machine learning focus on explaining a large black-box decision model by finding the most important features or using local approximations. Fewer works have been looking to make truly interpretable 'glassbox' models for fear of losing the great accuracy which comes from deep learning. This paper leverages recent work in the theory of feature interactions to provide a model which harnesses a powerful neural network to guide the construction of a corresponding interpretable model. We find that this new model has almost the same level of sparsity and interpretability as some of the simplest spline models, but leverages modern deep learning to get drastically improved results. We apply this model to the healthcare dataset of MIMIC and show improved results compared to not only interpretable methods, but also stronger baselines like neural networks and boosting machines. A full depiction of this model's decision-making is provided in the appendix.

We next apply this powerful and interpretable model to the econometric framework of causal inference, where our goal is to uncover the causal influence of a specific treatment. Our interpretable model allows us to make scientific inferences on the causality of our dataset. The golden standard for the causal inference framework are completely randomized clinical trials where we randomly assign the treatment to each patient. This setting is of great importance to the healthcare domain for studying the positive influence of certain drugs or treatments. In many serious studies regarding disease and critical care, however, we must deal with non-randomized experiments where providing optimal care to patients is paramount, introducing many confounders. We suggest first investigating the econometric setting where randomized data is abundant for learning causality under these constraints.

This paradigm shift also has positive consequences for the study of algorithm fairness, which is still struggling to define reasonable and all-encompassing metrics for fairness. Group notions of fairness are easy to measure but conflict with each other and most reasonable causal notions; causality-based fairness on the other hand develops rigorous metrics but is

nearly impossible to measure in practical settings. Causal inference toes the line between these two settings by being both easy to measure and causally motivated. We later show preliminary results in this direction. These ideas offer insights into a new paradigm of studying fairness and interpretability for algorithmic decision-making algorithms applied directly to individual-level patients for healthcare.

2 Preliminaries

Notation We will denote an n-dimensional input as x with its i-th component as x_i; its corresponding output will be denoted y. We will primarily consider one-dimensional y as is the case in regression and binary classification. We will use $f(x)$ to denote the function or model used to approximate y. We will denote subsets of the feature set by $\mathcal{I} = \{i_1, \ldots, i_{|\mathcal{I}|}\} \subseteq [n] := \{1, \ldots, n\}$. Its cardinality will be denoted $|\mathcal{I}|$, its complement $\setminus\mathcal{I}$, and its power set $\mathcal{P}(\mathcal{I})$. For $x \in \mathbb{R}^n$, we define:

$$(x_{\mathcal{I}})_i = \begin{cases} x_i & \text{if } i \in \mathcal{I} \\ 0 & \text{otherwise} \end{cases}$$

Feature Interactions The goal of *feature interaction detection* is to uncover the groups of features which depend on one another. That is to find the sets $\mathcal{I} \subset [n]$ such that $\omega(\mathcal{I})$ are positive and large, where $\omega(\mathcal{I})$ is defined as:

$$\omega(\mathcal{I}) := \mathbb{E}_x \left[\frac{\partial^{|\mathcal{I}|} f(x)}{\partial x_{i_1} \partial x_{i_2} \ldots \partial x_{i_{|\mathcal{I}|}}} \right]^2 > 0. \tag{1}$$

This equation corresponds to a 'statistical (non-additive) feature interaction' meaning that a function f cannot be decomposed into a sum of $|\mathcal{I}|$ arbitrary subfunctions f_i, which each exclude a corresponding interaction variable (Friedman and Popescu 2008; Sorokina et al. 2008; Tsang et al. 2018): $f(x) \neq \sum_{i \in \mathcal{I}} f_i(x_{\{1,2,\ldots,n\}\setminus i})$. In other words, the features $\{x_i : i \in \mathcal{I}\}$ must all be simultaneously known by at least one additive function to be able to predict the output $f(x)$. For example, $f(x_1, x_2) = x_1 x_2$ **does** have an additive feature interaction whereas $f(x_1, x_2) = \log(x_1 x_2)$ **does not** because it can be decomposed as $\log(x_1) + \log(x_2)$.

3 Methods

Feature Interaction Detection

Recent work in feature interactions has focused on detecting and measuring feature interactions from large-scale data. Three of the most popular and generally applicable of these feature interaction detection algorithms are the Shapley Taylor Interaction Index (STI) (Dhamdhere, Agarwal, and Sundararajan 2019), Integrated Hessians (Janizek, Sturmfels, and Lee 2020), and Archipelago (Tsang, Rambhatla, and Liu 2020). We choose Archipelago because of its speed relative to the other available methods. Details of how this algorithm estimates interaction strength are left to the appendix.

Generalized Additive Models

We consider the generalized additive model (GAM), a generalization of simple linear regression (Hastie and Tibshirani 1990)

$$g(y) = f_1(x_1) + \cdots + f_n(x_n) \tag{2}$$

where the f_i are arbitrary learned functions. g is called a link function which will be the identity function for regression and inverse-sigmoid for classification. This original formulation with only degree one feature interactions will henceforth be referred to as GAM-1. We adapt the original definition to also model a set of arbitrary interactions via:

$$g(y) = \sum_i f_i(x_i) + \sum_t f_{\mathcal{I}_t}(x_{\mathcal{I}_t}) \tag{3}$$

The first summation term corresponds to the classical GAM-1. The second term, however, extends GAMs to full capacity models which can represent complex nonlinear dependencies of arbitrary feature sets. For instance, if our set of interactions $\{\mathcal{I}_t\}_{t=1}^T$ includes the complete set $[n]$, then our model has *exactly* the same capacity as the underlying functional model we choose for the 'shape functions' f_i and $f_{\mathcal{I}}$ (splines, random forests, deep neural networks, etc.)

We will use GAM-K to refer to a GAM whose highest order interaction in $\{\mathcal{I}_t\}_{t=1}^T$ is of cardinality K. (i.e. $|\mathcal{I}_t| \leq K$ $\forall t \in [T]$.) For instance, the GA^{2}M model would be a GAM-2 since the interaction sets are all possible feature pairs: $\{\{i, j\} : i < j \in [n]\}$.

In our experiments, we use neural networks to fit our shape functions and introduce a novel feature interaction selection algorithm to maintain computational tractability. Details of the algorithm are left to the appendix for brevity and we refer to the resulting model as a Sparse Interaction Additive Network (SIAN). We will use the notation SIAN-K in the same way as GAM-K to denote the largest interaction.

Causal Inference

In this paper, we focus on the Neyman-Rubin causal model (Rosenbaum and Rubin 1983; Holland 1986) which is the model used for medical trials, policy interventions, and economic planning. In this framework, each individual is described by x -their 'features' or 'covariates' -including the *sensitive* features like race and gender as well as some *non-sensitive* features like (for a medical example) resting heart rate or (for a loan approval) past credit history. We will also consider the binary treatment variable T which either has value 0, corresponding to no treatment/ baseline, or value 1, corresponding to receiving the treatment (e.g. receiving the drug.) We then measure the outcome of the individual Y (e.g. did the patient feel better). Unfortunately, the crux of the issue is that we can only ever observe one of the two possible outcomes (Y_0 or Y_1) from either treatment 0 or treatment 1. Regardless, our goal is to estimate how the outcome Y depends on the covariates X. The individual treatment effect (ITE) is defined as the difference in expected value by choosing treatment 1 over treatment 0 as a function of the observed covariates x:

$$ITE(x) = \mathbb{E}[Y|X = x, T = 1] - \mathbb{E}[Y|X = x, T = 0] \tag{4}$$

We use this model to estimate the likelihood of donating blood when offered a small monetary incentive. In our results section, we visualize the discoveries made by our interpretable model using the causal dataset.

Algorithm 1: Feature Interaction Selection

Input: Trained prediction model f(x), and validation dataset $X^* = \{x_1, \ldots, x_v\}$

Parameter: Cutoff index K, cutoff threshold τ

Output: $\mathcal{I}$, a set of subsets of indices corresponding to approximately all feature interactions with index below K and strength below τ.

Set $\mathcal{I} \leftarrow \{\{i\} : i \in [n]\} \cup \emptyset$ // The true detected interactions so far
Set $\mathcal{J} \leftarrow \{\{i, j\} : i, j \in [n]; i < j\}$ // The next set of interactions to check
$k \leftarrow 2$
while $k \leq K$ **do**
  place holder
  for J in $\mathcal{J}$ **do**
    $\omega(J) \leftarrow 0$
    for x $\in X^*$ **do**
      Compute $\omega_J(x)$ with Archipelago
      $\omega(J) \leftarrow \omega(J) + \omega_J(x)$
    end for
    $\omega(J) \leftarrow \omega(J)/|X^*|$
    if $\omega(J) > \tau$ **then**
      $\mathcal{I} \leftarrow \mathcal{I} \cup J$
    end if
  end for
  $k \leftarrow k + 1$
  $\mathcal{J} \leftarrow \{J : J \in \mathcal{P}([n]); |J| = k;$
  $\frac{1}{2^k} \sum_{I \subseteq J} [1_{I \in \mathcal{I}}] > \theta\}$
end while
return $\mathcal{I}$

4 Datasets

Our experiments will focus on two primary datasets. The first is the publicly available MIMIC Health Care dataset which predicts mortality from a vast number of patient covariates having around 30 features, 32,000 samples, and a 9% positivity rate. The second is a blood bank's randomized trial which attempted to see the influence of a campaign offering previous donors a small monetary incentive with respect to likelihood to donate. This dataset has 34 covariates and 60,000 potential donors and a positivity rate of around 1%. We evaluate using both the area under the receiver operating characteristic (AUROC) and the area under the precision-recall curve (AUPRC) metrics. For the econometrics experiment we additionally look at economic performance by appraising blood donations and causal fairness metrics for both treatment and outcome.

5 Results

Across all three of these machine learning datasets, we find that there are consistently SIAN models which can easily match the AUROC and AUPRC performance of deep neural networks, kernel machines, and random forests.

We can see visualized in Figure 1 the risk of mortality as we increase age or maximum heart rate. Both of these show interpretable and expectable trends learned from our medical dataset.

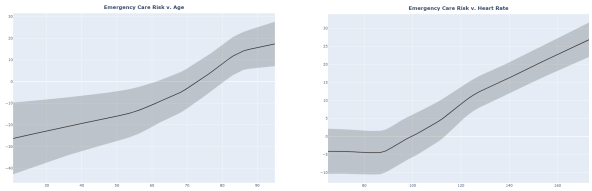

Figure 1: Mortality Risk vs. Age (left) and Heart Rate (right)

Model	AUROC		AUPRC	
	Early	Late	Early	Late
SAPS II	0.792	–	0.281	–
SOFA	0.703	–	0.225	–
svm	0.778	–	0.359	–
rf	0.789	–	0.363	–
xgb	0.803	–	**0.369**	–
ebm (ga^{2}m)	**0.816**	–	0.367	–
dnn	0.812	0.777	0.340	0.309
dnn-L	**0.816**	0.787	0.365	0.326
sian-1	**0.826**	0.824	0.369	0.368
sian-2xs	0.821	0.805	**0.384**	0.362
sian-2	0.820	0.805	0.372	0.351
sian-3	0.821	0.806	0.371	0.347
sian-5	0.809	0.787	0.367	0.338
sian-10	0.801	0.784	0.354	0.322

Table 1: MIMIC: Model Performance (Early/Late Stopping)

For the MIMIC dataset, we can see the models' performances in terms of both AUROC and AUPRC in Table 1. We also included the interpretable medical baselines of SOFA and SAPS II which are simple logic-based scoring methods (Le Gall, Lemeshow, and Saulnier 1993; Vincent et al. 1997). We next included the prolific machine learning methods of kernel machine, random forest, and deep neural network. For all of the network based approaches on this dataset, we evaluate both the best, early-stopped network and also the network which is trained for a long period of time and stopped at a specific number of epochs (300).

We can see that the DNN is comparable with SIAN's which have feature interaction sizes of 1-3. We can moreover see how there is a drop in performance when the DNN is trained for a longer period of time, showing its greater dependency on early-stopping. The SIAN-1 is almost completely resilient to late-stopping and the fall-off from late-stopping gradually increases alongside the size of the feature interactions.

We can see that the best AUROC model is the sian-1 and the best AUPRC model is the sian-2xs model, but both models consistently outperform the deep neural network, kernel machine, random forest, and boosting machine. Consequently, it becomes clear that the sian-1 or sian-2xs are the best performing models to be used for the application system. This is wonderfully convenient for a healthcare dataset, because these two models are by far the most interpretable networks. In the appendix D, we provide a complete vi-

sualization of this decision model which is able to outperform deep neural networks, kernel vector machines, random forests, and boosting machines.

In our causal inference dataset, we discover a salient trend in the data dependent on potentially sensitive features in Figure 2. We investigate how these discoveries influence the Treatment Fairness and Outcome Fairness of the decision algorithm.

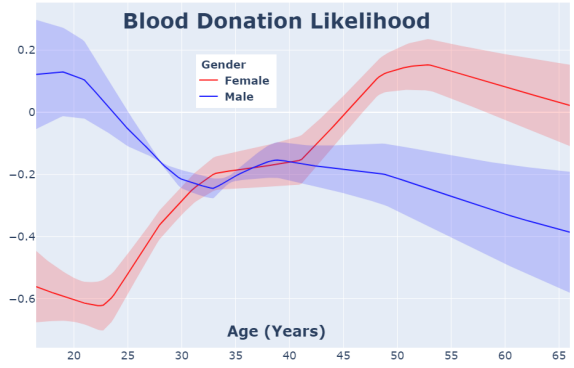

Figure 2: Age and Gender Feature Interaction Shape Plot

	donation percentage	economic benefit	AUROC T0	AUROC T1
none	0.500%	28.8± 5.3 ¢	–	–
all	0.981%	36.9± 0.2 ¢	–	–
random	0.740%	33.3± 1.1 ¢	–	–
boosting	0.970%	36.9± 0.2 ¢	50.0	69.1
svm	**1.000%**	**40.7± 4.7 ¢**	71.5	60.8
dnn	0.893%	36.8± 0.9 ¢	70.4	**75.8**
dnn	0.846%	37.7± 1.7 ¢	**73.4**	**75.8**
sian-1	0.706%	27.5± 1.3 ¢	68.0	73.8
sian-2	0.890%	**42.4± 0.2 ¢**	69.7	75.3

Table 2: Blood Donation: Performance Results

	donation percentage	economic benefit	gender TF	gender OF	age TF	age OF
sian-2	0.891%	**42.4± 0.2 ¢**	63.8	72.0	**95.8**	67.1
sian-2 (no gender)	0.900%	41.8± 0.5¢	**97.3**	71.9	88.5	59.3
sian-2 (no age)	0.837%	40.5± 0.5¢	74.9	69.4	91.5	**73.5**
sian-2 (no gender or age)	0.845%	40.4± 0.5¢	**97.6**	64.9	**95.3**	72.1

Table 3: Blood Donation: Remove Sensitive Features

In Table 2, we can see that SIAN is still competitive with all other machine learning methods and although SIAN does not outperform all models for AUROC or donation percentage, it still dominates the other models in terms of the economic benefit from evaluation. In Table 3, we see how removing sensitive shape functions like the one depicted in Figure 2 influences the performance of the model. We can see that the performance gradually degrades as we remove the information from the model and that the treatment fairness increases; however, the outcome fairness does not seem to have such a clear relationship. Future work will need to investigate this relationship causal fairness and algorithmic decision-making and develop methods which can consistently ameliorate the unfair decisions of any model type.

6 Limitations and Future Work

The primary limitation of the current work is the model's restriction to the static setting and further work is needed to integrate the model into a time series setting where we evaluate clinical measurements over time to get the best possible prediction. Another current limitation is the longer training time compared to other methods, although this is of minimal importance at test time in the domain of healthcare.

Further understanding of how the benefits of SIAN scale with the dimensionality of the dataset and the number of available samples is of great importance, especially in application to the causal inference setting where the number of test subjects may be limited greatly. Future work will need to develop the application domain of econometrics and causal inference before we know the exact practicality of this model when facing the greater number of confounders and lack of purely randomized experiments which we will face in medical applications. When these goals are achieved, SIAN and other interpretable models are promising candidates for safety-critical decision-making models in both healthcare and beyond.

Multiple experiments confirm that SIAN models can easily produce powerful and interpretable machine learning models which can consistently match the performance of state-of-the-art benchmarks like deep neural networks, kernel machines, random forests, and boosting machines. Hopefully, future work will continue to use SIAN and other interpretable models to improve performance in interpretable healthcare and further develop our understanding of machine learning alongside causal inference and fairness in medical domains.

References

Ambrosino, R.; Buchanan, B. G.; Cooper, G. F.; and Fine, M. J. 1995. The use of misclassification costs to learn rule-based decision support models for cost-effective hospital admission strategies. *Proc Annu Symp Comput Appl Med Care*, 304(8).

Dhamdhere, K.; Agarwal, A.; and Sundararajan, M. 2019. The Shapley Taylor Interaction Index. *arXiv preprint arXiv:1902.05622*.

Duchi, J.; Hazan, E.; and Singer, Y. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul): 2121–2159.

Friedman, J. H.; and Popescu, B. E. 2008. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 916–954.

Hastie, T. J.; and Tibshirani, R. J. 1990. Generalized additive models.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385.

Holland, P. W. 1986. Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396): 945–960.

Janizek, J. D.; Sturmfels, P.; and Lee, S.-I. 2020. Explaining Explanations: Axiomatic Feature Interactions for Deep Networks. *arXiv preprint arXiv:2002.04138*.

Le Gall, J.-R.; Lemeshow, S.; and Saulnier, F. 1993. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *JAMA*, 270(24).

Rosenbaum, P. R.; and Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1): 41–55.

Silver, D.; and et al. 2017. Mastering the game of Go without human knowledge. *Nature*, 550: 354–359.

Sorokina, D.; Caruana, R.; Riedewald, M.; and Fink, D. 2008. Detecting statistical interactions with additive groves of trees. In *Proceedings of the 25th international conference on Machine learning*, 1000–1007. ACM.

Tsang, M.; Liu, H.; Purushotham, S.; Murali, P.; and Liu, Y. 2018. Neural Interaction Transparency (NIT): Disentangling Learned Interactions for Improved Interpretability. In *Advances in Neural Information Processing Systems*, 5804–5813.

Tsang, M.; Rambhatla, S.; and Liu, Y. 2020. How does this interaction affect me? Interpretable attribution for feature interactions. *arXiv preprint arXiv:2006.10965*.

Vaswani, A.; and et al. 2017. Attention Is All You Need. *CoRR*, abs/1706.03762.

Vincent, J. L.; Moreno, R.; Takala, J.; Willatts, S.; Mendonça, A. D.; Bruining, H.; Reinhart, C. K.; Suter, P. M.; and Thijs, L. G. 1997. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. *Intensive Care Med.*, 22(7): 707–710.

A Algorithmic Details

Practical Details

For the baseline DNNs we are using hidden layer sizes (256,128,64) with ReLU activations. For the GAM subnetworks we are using hidden layer sizes (16,12,8) with ReLU activations. We also consider DNNs of sizes (1024,512,256) and (128,64,32) when appropriate (referred to as DNN-L and DNN-S respectively.) The parameters for the FIS algorithm are the degree of interactions K and the feature interaction strength threshold τ. In our experiment results, our GAMs are decorated with their interaction degree K and the magnitude of τ (e.g. gam2xs refers to a second-degree GAM with a very large τ corresponding to an extra-small set of feature interactions, gam3xl refers to a third-degree GAM with a very small $\tau = 0$). GAMs which are not decorated with a size indicator are using the default τ parameter for their dataset. We train all networks using Adagrad (Duchi, Hazan, and Singer 2011) with a learning rate of 5e-3.

Archipelago

Archipelago quickly estimates the feature interaction strength by approximating the Hessian of the model given a target data instance x^* and a baseline data instance x'. The secant approximation of the Hessian which is used by Archipelago is defined:

$$\omega_{i,j}(x) = \Big(\frac{1}{h_i h_j}(f(x^{\star}_{\{i,j\}}+x_{\setminus\{i,j\}})-f(x'_{\{i\}}+x^{\star}_{\{j\}}+x_{\setminus\{i,j\}}) - f(x^{\star}_{\{i\}}+x'_{\{j\}}+x_{\setminus\{i,j\}})+f(x'_{\{i,j\}}+x_{\setminus\{i,j\}}))\Big)^2 \quad (5)$$

where $h_i = x^*_i - x'_i$, $h_j = x^*_j - x'_j$, x^* is a target data instance (a member of our data population) and x' is a baseline data instance (usually the all zeroes vector).

The most thorough way to identify this feature interaction is using many different 'context vectors' x. Because of this, we theoretically have $\bar{\omega}_{i,j} = \mathbb{E}_{x\in\mathcal{X}}[\omega_{i,j}(x)]$ where the expectation is over the 'full context' of all possible combinations of the target x^* and the baseline x'. We only consider the following approximation which greatly lessens the computation: $\bar{\omega}_{i,j} := \frac{1}{2}(\omega_{i,j}(x^*) + \omega_{i,j}(x'))$.

This generates a score $\bar{\omega}_{i,j}$ for every possible feature pair $\{(i,j) : i \neq j \in [n]\}$. Higher degree interactions are calculated in the exact same way using a similar secant approximation of the higher order derivatives.

Feature Interaction Selection Algorithm Below in Algorithm 2, we show how to select the feature interaction sets from all possible subsets using the inductive biases of a trained neural network.

Fairness Metrics

We use treatment fairness and outcome fairness as the two primary measurement for fairness. For the rest of our discussion, we will only study the case of binary variables S which need protection, although more general cases are still applicable.

Algorithm 2: Feature Interaction Selection

Input: Trained prediction model f(x), and validation dataset $X^* = \{x_1, \ldots, x_v\}$

Parameter: Cutoff index K, cutoff threshold τ

Output: $\mathcal{I}$, a set of subsets of indices corresponding to approximately all feature interactions with index below K and strength below τ.

Set $\mathcal{I} \leftarrow \{\{i\} : i \in [n]\} \cup \emptyset$ // The true detected interactions so far
Set $\mathcal{J} \leftarrow \{\{i, j\} : i, j \in [n]; i < j\}$ // The next set of interactions to check
$k \leftarrow 2$
while $k \leq K$ **do**
  place holder
  for J in $\mathcal{J}$ **do**
    $\omega(J) \leftarrow 0$
    for x $\in X^*$ **do**
      Compute $\omega_J(x)$ with Archipelago
      $\omega(J) \leftarrow \omega(J) + \omega_J(x)$
    end for
    $\omega(J) \leftarrow \omega(J)/|X^*|$
    if $\omega(J) > \tau$ **then**
      $\mathcal{I} \leftarrow \mathcal{I} \cup J$
    end if
  end for
  $k \leftarrow k + 1$
  $\mathcal{J} \leftarrow \{J : J \in \mathcal{P}([n]); |J| = k\}$; $\frac{1}{2^{|J|}} \sum_{I \subseteq J} [1_{I \in \mathcal{I}}] > \theta\}$
end while
return $\mathcal{I}$

Treatment Fairness For treatment fairness, we simply ask for 'Statistical Parity' on the treatment variables. That is to say we ask that the same percentage of each group is given the treatment.

$$P(T = 1|S = 1) = P(T = 1|S = 0) \tag{6}$$

Outcome Fairness For outcome fairness, we use a measure similar to 'Predictive Parity' where within the treated members, we have the same percentages of each group taking the outcome action.

$$P(Y = 1|T = 1, S = 1) = P(Y = 1|T = 1, S = 0) \tag{7}$$

We evaluate both of these parity equations using the $p\%$-rule for binary features. This measurement simply takes the lowest ratio between the two percentage values for each sensitive class. 100% score would then correspond to the equalities above. This rule comes out of the current legal framework which claims anything below 80% is possibly discrimination. We will refer to the $p\%$ values for these two metrics as TF and OF accordingly. We will define TF and OF below as a function of an underlying distribution $\mathcal{P}$ and a decision algorithm D as follows:

$$TF(D, \mathcal{P}) := \frac{\mathbb{E}_{X,Y \sim \mathcal{P}}[D(X)|S = 0]}{\mathbb{E}_{X',Y' \sim \mathcal{P}}[D(X')|S = 1]} \tag{8}$$

$$OF(D, \mathcal{P}) := \frac{\mathbb{E}_{X,Y \sim \mathcal{P}}[Y_{D(X)}|S = 0]}{\mathbb{E}_{X',Y' \sim \mathcal{P}}[Y'_{D(X')}|S = 1]} \tag{9}$$

Consequently, our $p\%$ rule corresponds to the optimization constraint that TF or OF belong to the interval $(1 - \varepsilon, \frac{1}{1-\varepsilon})$ or that their logarithms belong to $(-\varepsilon', \varepsilon')$. It is hence possible to reconsider these notions as an additive constraint instead of a multiplicative one.

B Visualizing Shape Functions

When we use GAM2 models we are also able to easily visualize their shape functions using heatmaps, making them incredibly interpretable models. For dimensions three and higher, it becomes increasingly difficult to find suitable representations of the information being uncovered by the GAM.

Below in Figure 3, we can see what the MIMIC gam2xs was thinking about the feature pairs it was told to look at in greater depth.

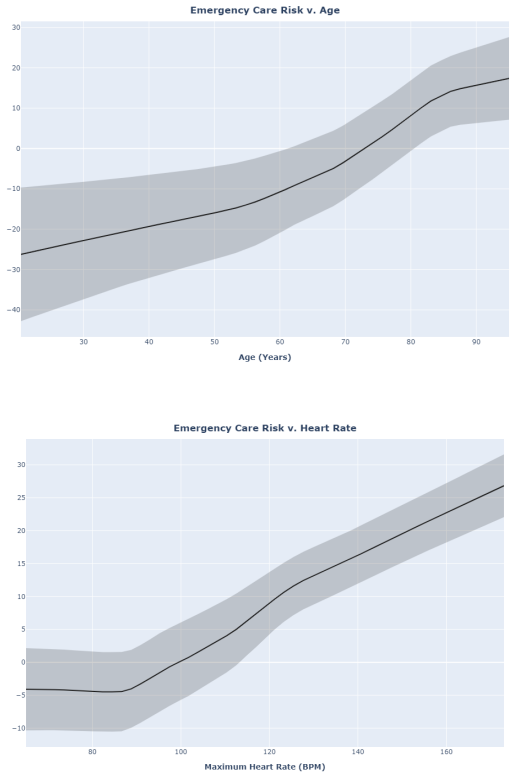

Figure 3: Mortality Risk vs. Age (left) and Heart Rate (right)

We can see in the first shape function that the maximum Blood Urea Nitrogen and maximum Blood Pressure have an interaction where if either one is high then there is likely a higher risk of mortality, but having both be relatively high is actually much less indicative of mortality. The second shape function shows that while older people are more likely to die, a high fever is actually more indicative of a problem in a mid-age patient (40-60 years old) than the risk in much older patients. The third shape function shows the relatively expected trend of higher risk with older patients in a mono-

tonic and relatively linear fashion. The fourth shape function shows that having a maximal Heart Rate below 100 is actually better for survivability, but as the maximal heart rate of a patient grows to 140 or 160, the risk of mortality grows nearly linear with the increase in heart rate.

These trends can be directly compared not only with the dataset, but also medical experts who can either agree or disagree with the conclusions of the model. These trends are extremely important in a medical setting where machine learning models are at risk of learning correlations instead of causations. Famous examples of this are when ML models thought that asthma made pneumonia less deadly or that temperatures above 40°C we're not at risk. Sometimes trends that are so obvious to humans occlude these trends from machine learning models which are trained in an observational fashion, precluding them from realizing that people in these groups are in fact so at risk that they are immediately given emergency treatment (Ambrosino et al. 1995; Tsang et al. 2018). Being able to fully interpret a GAM2 model is a key benefit of these models for application into safety-critical decision-making scenarios like health care.

C Healthcare: Interpretability and Causality

Interpretable machine learning models are of critical importance for the domain of healthcare. Even greater is the need for interpretability alongside causality to avoid correlation-based analyses showing that asthma reduces death risk from pneumonia or age reduces death risk after one hundred years. Statistically validated models are of extreme importance when there are lives on the line and we really need to make sure that the trends we are learning from our data hold up against our scientific understanding of the world. The possibility for machine and human collaboration in the medical field opens up vastly once we can marry this causal understanding with human-interpretable machine learning results.

Clinical trials are the golden standard of experimental design; however, they are extremely costly and can be restrictive in critical care settings where it becomes immoral to give a patient poor treatment. In this setting, we need to do more work to balance confounding variables with potentially unknown treatment distributions. It is clear that there are a great number of promises behind the door into the world of causal inference when we can handle these edge cases. For this reason, we propose using the cheap, large-scale experiments from econometrics as a test bed for understanding how machine learning works for these tasks. Only after rigorous testing in these low-impact applications can we then apply them to safety critical healthcare tasks.

In addition to the incentives from an interpretability perspective, there are also tangible benefits in the fairness domain. Currently, algorithm fairness is struggling to pin down good metrics for fairness and while most of the original group fairness notions were easy to compute, they have been discarded because of their inability to capture causal relationships from the data. Causality-based fairness metrics, however, require extremely rigorous assumptions which can seldom be made in practice. This leaves practitioners with no good options for measuring the biases of their algorithms. It seems that causal inference achieves a balance between these two extremes by providing metrics which are easy to evaluate and still measure causality from the data.

For these and other reasons, it seems that many discussions of the algorithmic impact on human lives should be shifted to this paradigm of causal inference where we view the algorithmic decision as a treatment which elicits downstream consequences. In healthcare especially, we envision this might be the only way to consistently and accurately evaluate the positive and negative impacts of implementing an algorithmic decision-making tool into the healthcare pipeline against a typical clinician over a variety of metrics with respect to safety, robustness, and fairness.

D THE ENTIRE GAM2XS MODEL SHAPES

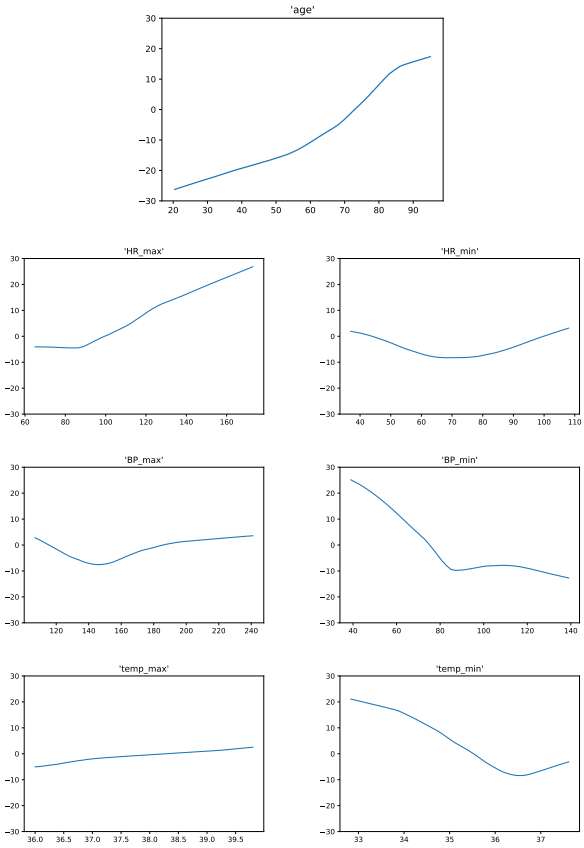

Figure 4: MIMIC 1D Shape Functions

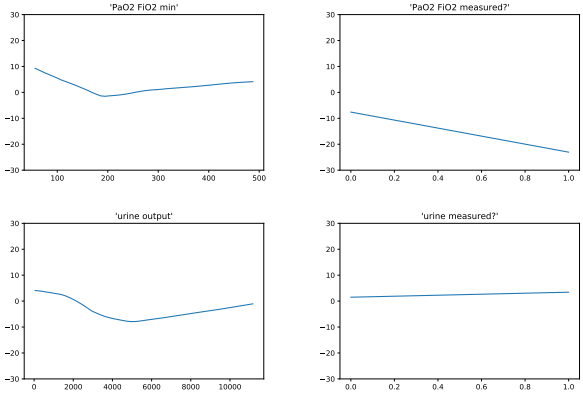

Figure 5: MIMIC 1D Shape Functions

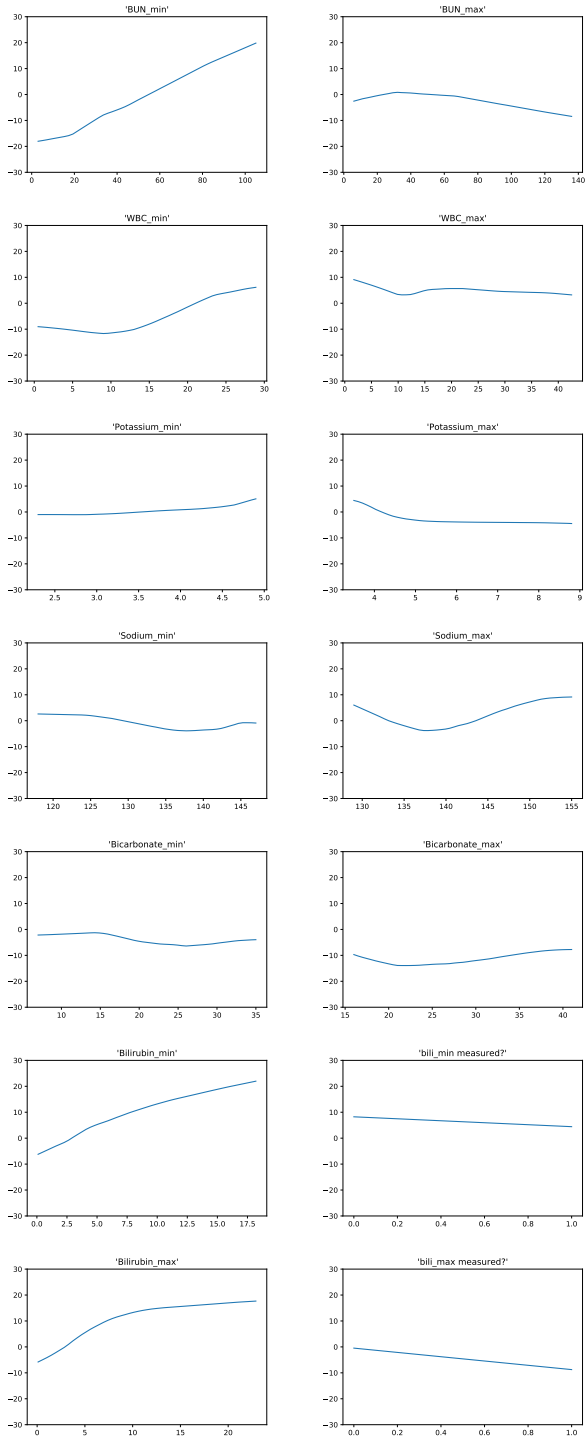

Figure 6: MIMIC 1D Shape Functions

.

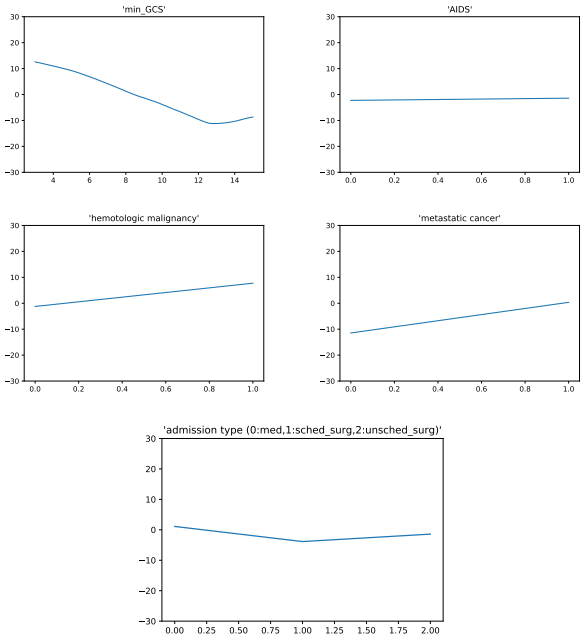

Figure 7: MIMIC 1D Shape Functions

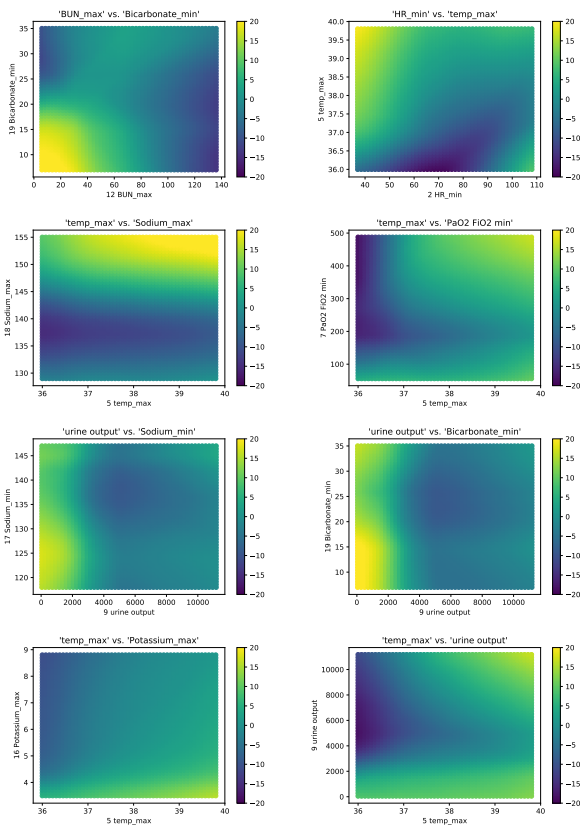

Figure 8: MIMIC 2D Shape Functions

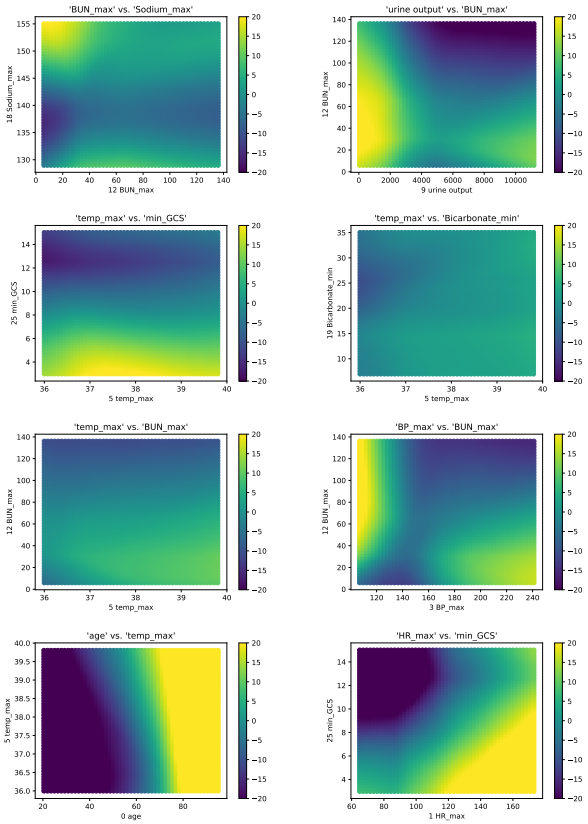

Figure 9: MIMIC 2D Shape Functions

E Code

A Python implementation using Pytorch will be made available at github.com/USC-Melady.