Predicting Angiographic Disease Status: Drawing the line between demographically decoupled and jointly trained models

Ananth Balashankar[1,2], Alyssa Lees[2], Srikanth Jagabathula[3], and Lakshminarayanan Subramanian[1,4,*]

[1]Courant Institute of Mathematical Sciences, New York University, New York, NY, USA

[2]Google AI, New York, NY, USA

[3]Stern School of Business, New York University, New York, NY, USA

[4]Department of Population Health, New York University School of Medicine, New York, NY, USA

[*]Corresponding author: lakshmi@cs.nyu.edu

December 30, 2021

Keywords: Demographic analysis, angiography, machine learning, gender bias, age groups

Word count: 3,979

Abstract

Objective

Application of machine learning predictive models for determining angiographic disease status have the risk of amplifying existing biases towards demographic groups based on age and gender. We formalize the underlying choice between demographically decoupled and jointly trained models and propose a framework that allows domain experts to balance equity with the goal of achieving the most accurate classifier for all demographic groups.

Materials and Methods

We propose an objective called Demographic Pareto Efficiency to discover classifiers for diagnosing angiographic disease status that optimize the demographic group accuracies for four groups based on age (>60, <=60) and gender (male, female). We discover predictive models on the Pareto frontier of group-level accuracy and aid domain experts in making efficient accuracy trade-offs.

Results

We outperform baseline classifiers incorporating min-max, adversarial and parity based notions of fairness both in overall accuracy and group-level accuracy by up to 9.7% and 9.6% respectively while retaining comparable levels of discrepancy between groups in the UCI Heart Disease dataset. Our approach searches for Pareto optimal group performance, whereas baseline approaches converge to non-pareto solutions, thus leaving room for improvement in group level accuracies.

Discussion

In determining angiographic disease status, machine learning predictive models need to maximize the accuracy of stratified demographic groups based on age and gender, by leveraging the benefits of transfer learning across groups through iterative and group-aware joint training, rather than maximizing an overall demographic group-agnostic accuracy measure.

Conclusion

Demographic Pareto Efficiency provides a framework to maximize prediction accuracy across demographic groups, while retaining fairness within a relaxation bound.

1 INTRODUCTION

Societal inequities have the real risk of being vastly exacerbated if machine learning algorithms do not take explicitly address issues of demographic inequity [1, 2, 3, 4, 5]. In the context of diagnosis of angiographic disease status, age and gender based demographic groups have been known to have prognostic differences in CT coronary angiography, with females below 60 years of age have the least predictive value [6]. Prior work have also shown that women have higher mortality from myocardial infarction, mostly at younger ages [7, 8, 9]. Given that different demographic groups based on age and gender have different profiles of heart disease, the problem of improving the predictive accuracy of diagnostics across such demographic groups has not been explicitly tackled as the primary objective. Instead, there is an emphasis on overall accuracy of patients when using Machine Learning (ML) based predictive models. In this paper, we define the notion of "Demographic Pareto Efficiency" (interchangeably referred to as pareto efficiency) as a guiding principle for domain experts to choose diagnostic models that improve predictive accuracy of angiographic disease status across demographic groups based on age and gender; and provide a methodology that discovers a larger set of ML models that consistently improve upon the predictive accuracy for all demographic groups. Specifically, our methodology makes the choice between learning separate decoupled models, one for each of the group, and a joint model trained on all groups based the main outcome measure of demographic group-level accuracies.

Improving equity in health is well studied and various philosophical notions of fairness exist (distributive, procedural, etc.) [10, 11, 12, 13] and the appropriateness of each definition depends on the ethical context in which they are applied. Theoretically, in an equitable world of perfect data, a classifier with perfect diagnostic accuracy across all subgroup populations may be created. Due to a variety of reasons including historical injustices [14], sampling bias [15], selection bias [16], label noise, among others, group populations are often not fully represented in commonly used real-world health datasets [17]. With such skewed data, [18] has shown that an unavoidable trade-off exists between group fairness and accuracy. With this trade-off, domain experts have to choose between coupled (jointly-trained on all groups) and decoupled (one model per group) models based on how well they balance the demographic group accuracies. While the benefits of decoupled models are known theoretically when we have large and diverse datasets [19], the impact of such models on group-level accuracy in diagnosing the angiographic disease status in patients remains unexplored. We investigate the role of decoupled training across demographic groups based on age and gender in the UCI Heart Disease dataset. Inspired by social science and welfare economics literature [20, 21, 22] (see S.I for detailed related work), we propose a novel methodology that combines decoupled group-wise models and use them to guide a jointly trained model to achieve demographic pareto efficiency [23]. .

Demographic Pareto Efficiency [23] is achieved when no single group performance can be improved without the degradation in performance of another group. The set of all such group level performances when plotted in a multi-dimensional graph (one group's performance per dimension), forms the Pareto frontier (like blue dots illustrated

Optimization Objective	**Operating Point**
Overall Accuracy	$opt_b = (0.63, 0.77)$
Strict Accuracy Equality[3]	$(0.60, 0.60)$
Adversarial [24]	$(0.73, 0.56)$
Mini-max [25]	$(0.68, 0.63)$
Pareto Efficiency (Ours)	$PE = (0.71, 0.63)$

Table 1: Preferred classifiers and their demographic group-level accuracy based on different objectives in Fig 1.

in a simulation shown in Figure 1). Ensuring that classifiers achieve Demographic Pareto Efficiency while balancing fairness constraints and accuracy has critical implications to the discussion about the unavoidable accuracy-fairness trade-offs in the real world [18]. For example, if domain practitioners are required to make a choice between two classifiers based on the demographic accuracy-fairness trade-off in predicting the Angiographic disease status, the comparison would be meaningful only if both those classifiers were on the Pareto frontier. Otherwise, the discussion of demographic accuracy and fairness trade-offs would be premature as there exists a third classifier which can achieve better group level accuracy and better medical outcomes. (e.g.: "Pareto Efficient Fairness" should be preferred over "Strict Accuracy Equality" in Table 1). Through our approach, we discover such Pareto efficient predictive models to be considered as candidates in determining the angiographic disease status of patients, and avoid unnecessary concessions in group level accuracy without significant degradation in fairness (demographic parity).

2 BACKGROUND

Problem Definition: The angiographic disease status is defined as a binary label (diseased or not) based on the fact if there is more than 50% diameter narrowing in any of the major blood vessels in a patient (lmt, ladprox, laddist, diag, cxmain, ramus, om1, om2, rcaprox, rcadist). To predict this angiographic disease status, we use 13 input attributes of the patient such as age, gender, chest pain type (typical angina, atypical angina, non-anginal and asymptomatic), resting blood pressure (mm Hg on admission to hospital), serum cholestrol (mg/dl), fasting blood sugar (binary >120 mg/dl), resting electrocardiographic results (normal, having ST-T wave abonormality, probable or definite left ventricular hypertrophy), maximum heart rate achieved, exercise induced angina (yes/no), ST depression induced by exercise relative to rest, slope of the peak exercise ST segment, number of major vessels colored by fluoroscopy, and detection of thalassemia (none, major or reversible defect). For this heart disease dataset, different machine learning algorithms may be trained on the same training dataset of patients, and may obtain different test accuracies on demographic groups as illustrated in Figure 1. For example, when we plot each ML algorithm as a separate point, where the value along the x and y axis indicate the test accuracy over demographic groups A and B respectively, we see that some algorithms perform poorly on both groups (annotated as alg-1..5 in grey) as compared to other algorithms (annotated as opt_b, overall, PE, opt_a in blue). The blue line indicates the Pareto frontier of the demographic group test accuracies, which defines the set of optimal choices a domain expert would have if they were to optimize for demographic group test accuracies. Note that each algorithm on the Pareto frontier when compared with another algorithm on the Pareto frontier, performs better on one demographic group and poorly on the other, but never poorly on both the demographic groups. Hence any choice the domain expert would make among these ML algorithms would need to trade-off one demographic group's accuracy for the other. While this choice depends on the domain expert and the context in which they operate (other remedial or diagnostic measures incorporated for specific demographic groups), the key takeaway from this illustration is that optimizing for overall accuracy without consideration of the demographic groups may lead us to one of the points on the Pareto frontier, but may not always be the one that the domain expert would choose given all options on the Pareto frontier. *Hence, our primary goal in this paper is to discover all the machine algorithms on the Pareto frontier, so that the domain expert is given a rich set of algorithms with Pareto optimal demographic group accuracies to choose from.* However, discovering the Pareto frontier is non-trivial as their discovery is driven by optimizing specific demographic group accuracies, while keeping the accuracy of other algorithms constant. Further, this problem is exacerbated by disparate sampling bias, epistemic, and aleatoric uncertainty about how angiographic disease status presents itself in different demographic groups. Thus, it is not known that a simplistic training objective such as improving overall accuracy is sufficient for discovering the full Pareto frontier. Thus, in order to discover the full Pareto frontier in a systematic approach, we present an iterative training methodology.

3 MATERIALS AND METHODS

We now formally define Demographic Pareto Efficiency and explain how to train a joint model that leverages the benefits of decoupled classifiers in discovering Pareto efficient classifiers on the Pareto frontier.

Definition 1. ***Demographic Pareto Efficiency****: We introduce Demographic Pareto Efficiency as a set of classifiers with respect to groups (defined by sensitive attributes). Demographic Pareto Efficient classifiers are defined as the set of classifiers where there does not exist another classifier which has better performance (for a defined performance metric such as accuracy, TPR, etc) across all groups.*

For groups $g \in |G|$, we denote $(f_1, f_2 ... f_{|G|})$ as the tuple of group performance metrics achieved by any Demographic Pareto Efficient classifier. The $f_1, f_2, ..f_{|G|}$ are group performance metric values that are to be maximized (e.g. accuracy, TPR). Formally, Demographic Pareto Efficiency states that no other classifier exists with performance metrics $(q_1, q_2 ... q_{|G|})$ such that $f_1 \leq q_1$ and $f_2 \leq q_2$ and .. $f_{|G|} \leq q_{|G|}$.

Definition 2. ***Pareto Loss****: Pareto Loss, ϵ_g, for a group g, is defined as the relative difference between the performance* *of a classifier for that group f_g and the optimal performance for the group f_{opt-g} across all discovered classifiers.*

$$\epsilon_g = 1 - \frac{f_g}{f_{opt-g}} \tag{1}$$

While the above formulation of optimizing the Pareto loss can lead to multiple Pareto efficient decoupled and jointly trained diagnostic models of angiographic disease, the domain expert has to choose a single classifier among them post-training. Ideally, choosing the most desirable classifier is left to the end, once the complete Pareto front has been discovered. As discovering the Pareto frontier itself is our problem statement, this can lead to a deadlock condition, where an effective choice between decoupled and jointly trained models cannot be made without making choices that lead to better exploration at training time.

Definition 3. ***Pareto Efficient Fairness****: We define a classifier as Pareto Efficient Fair (PEF) if it is Pareto Efficient and minimizes a weighted average of variance and absolute sum of the Pareto loss across groups.*

The definition of Pareto loss requires us to know the true optimal performance per group f_{opt-g} a priori, which may not be possible. Hence, we use a decoupled classifier to estimate these optimal values at each iteration of training.

Pareto Efficient Algorithm: A heuristic pseudo-optimal group accuracy f_{opt-g} for each group g is formulated by training a decoupled classifier M_g to minimize the cross-entropy loss $\mathcal{L}_{ce}$ on samples in group g from dataset D [26]. We then iteratively update f_{opt-g} if a better group accuracy is evaluated by a jointly trained model M on a held-out test set using the $eval$ function. A summary of the Pareto Efficient bias mitigation algorithm is presented in Algorithm 1, and the corresponding components are explained in detail below. This is an in-processing algorithm (as opposed to post-processing [27]) which trains a joint model M on all subgroups to minimize the Pareto Efficient fairness loss $\mathcal{L}_p$ in every batch by stochastic gradient descent. We strictly ensure that the mini-batch is representative of the group distributions by sampling group-wise batch samples proportionately. Our algorithm explicitly achieves *potentially optimal* performance for each of the groups by explicitly recognizing these differences [28] as opposed to ones which do so implicitly [29]. Now, we formally define our fairness based Pareto loss function $\mathcal{L}_p$ used in each iteration of our algorithm.

Consider a set of Demographic Pareto Efficient classifiers T_{PE}, with each classifier $t \in T_{PE}$ containing a tuple of Pareto losses $\mathcal{E}_{t,G} = (\epsilon_{t,1}, \epsilon_{t,2}, ...\epsilon_{t,|G|})$. The sample variance of Pareto loss across groups is denoted by $\sigma^2_{t,G}(\mathcal{E}_{t,G})$. The goal is to find the Pareto Efficient Fair classifier $t_{PE-fair}$ that minimizes the variance of Pareto losses among all groups. Since it is empirically difficult to find all the Demographic Pareto Efficient classifiers T_{PE} at each iteration of our algorithm, we relax this by approximating the Pareto classifiers as ones that have a low absolute sum of group Pareto losses ($\|\mathcal{E}_{t,G}\|_1$) among all classifiers $t \in T$. Since the classifier with the lowest absolute Pareto loss may not equate to the classifier that minimizes the variance of the Pareto loss across groups and vice-versa, we trade-off these

Algorithm 1: Iterative Pareto Efficient Bias Mitigation

```
$G$: set of sensitive groups, $D$: dataset, $D_g$: data of group $g \in G$
for $g \in G$ do
    $M_g = \arg\min \mathcal{L}_{ce}(D_g)$
    $f_{opt-g} = \text{eval}(M_g, D_g)$
    $f_g = \emptyset$
end for
while $\exists g \in G, f_g = \emptyset \vee f_g > f_{opt-g}$ do
    $f_{opt-g} = \max(f_g, f_{opt-g}), \forall g \in G$
    $train(M, \mathcal{L}_p(D))$
    $f_g = eval(M, D_g), \forall g \in G$
end while
return $M$
```

two minimization criterion using a Lagrangian factor α in the *Group Pareto Loss* as follows:

$$t_{PE-fair} = \underset{t \in T_{PE}}{\arg\min}\, \sigma^2_{t,G}(\mathcal{E}_{t,G}) \tag{2}$$

$$\approx \underset{t \in T}{\arg\min}\, \alpha \|\mathcal{E}_{t,G}\|_1 + (1-\alpha)\sigma^2_{t,G}(\mathcal{E}_{t,G}) \tag{3}$$

When $\alpha = 0$, the variance of Pareto loss is minimized, whereas, when $\alpha = 1$, we minimize the absolute Pareto loss. In all our experiments, we chose $\alpha = 0.5$ after cross-validation, however the domain expert in the angiographic disease diagnoses might chose another value based on the trade-off between variance and absolute sum of Pareto losses. By making this choice explicit, we can demand transparency from practitioners deploying diagnostic ML models about the trade-offs they made. A high α would force that each demographic group be as close as possible to it's optimal performance, whereas a low α would enforce that each group suffer similar Pareto losses as compared to their optimal group performance.

Augmented Pareto Loss: We now generalize our definitions for any binary diagnostic model. Here, the minimization criterion of the Group Pareto Loss, but we minimize the group Pareto Loss over the parameters of the binary classification model using stochastic batch gradient descent. The Group Pareto Loss is augmented with an appropriate loss weight (λ) via the Lagrangian dual formulation similar to [30]. As an example, the standard cross-entropy classification loss: $\mathcal{L}_{ce}$ [31] can be augmented to yield the Pareto Efficient Fairness Loss: $\mathcal{L}_p$. The penalty term weighted by λ is used to ensure that maximum overall accuracy can be achieved while minimizing a combination of the absolute Pareto loss and its variance. After cross-validation, we set $\lambda = 0.1$, but here too the domain expert might choose based on external factors that impact the relative weight of overall as compared to group-level accuracy. (see S.I for detailed methods)

$$\mathcal{L}_p = \mathcal{L}_{ce} + \lambda(\alpha \|\mathcal{E}_G\|_1 + (1-\alpha)\sigma_G^2(\mathcal{E}_G)) \tag{4}$$

4 RESULTS

Here, we predict health status as binary label (presence or absence of Heart Disease) using medical and demographic information, where we consider age (>60, <=60) and gender (male, female) to be the stratification variables. The intersection of these 2 variables are considered sensitive groups in our study.

The dataset consists of 920 patients from four hospitals of Cleveland Clinic Foundation; Hungarian Institute of Cardiology, Budapest, V.A. Medical Center, Long Beach, CA; and University Hospital, Zurich, Switzerland with a total of 75 attributes, out of which 13 attributes are used for predicting the binary label of angiographic disease status (0: <50% diameter narrowing, 1: >50% diameter narrowing). The number of samples in each of the four demographic groups Young/Male, Young/Female, Old/Male, Old/Female are 550, 149, 176 and 45 respectively. We split the dataset into a 10-fold train/test random stratified splits (train on 9 splits, and test on the remaining split, repeated 10 times) based on the demographic groups to ensure that the training and test data are sampled from the same distribution and that all demographic groups are represented as per the dataset. We compare our approach with the scaled versions of group fairness [3] and [24] for groups. In [3], the authors optimize for overall accuracy in the constrained setting of ensuring equal false positive rates. The method is generally applicable to other measures of performance. For comparison, we implement an objective to maximize overall accuracy along with a Lagrangian relaxation which adds a penalty for parity loss (deviation from the overall accuracy) for each group.

This baseline scenario is equivalent to optimizing for balanced accuracy across sub-groups or assuming that perfect group-level performance can be achieved (accuracy of 100%). Instead, in our iterative approach, we use a per-group decoupled classifier's pareto optimal performance as a training signal. In [24], the authors implement bias mitigation as a way of erasing the sensitive group membership by back-propagating negative gradients in a multi-headed feedforward neural network. In [25], they adopt a minimax objective that ensures that the least performing group has the highest accuracy possible. We evaluate by comparing these 4 techniques on the UCI Heart disease dataset. We perform a 10-fold cross validation and report the average accuracy across the 10 splits.

Preprocessing

Each entry in the dataset has been pre-processed using the one-hot encoding for categorical features and the Tensorflow bucketization library into 10 buckets for numeric features. The resulting embedding is concatenated and used as input to a 3-layer feedforward neural network with 256, 128 and 64 hidden units respectively. We trained each of the models

Model	**Accuracy**	**FPR**	**FNR**	**Parity Loss**	**Pareto Loss**
Baseline (no bias loss)	0.879	0.348	0.701	0.192	0.018
Equality Constraint[3]	0.870	0.381	**0.684**	**0.132**	0.123
Adversarial [24]	0.837	0.327	0.723	0.253	0.087
Min-max [25]	0.839	0.306	0.765	0.231	0.055
Pareto Efficient Fair Loss	**0.939**	**0.266**	0.690	0.198	**0.000**

Table 2: Comparison of test losses in UCI Heart Disease dataset. PEF optimizes Pareto loss, while [3] minimizes Parity loss. The higher parity loss for PEF does not mean degrading group performances, but instead improves each group. Also, PEF and [3] achieve best False Positive Rate (FPR) and False Negative Rate (FNR) respectively as a side-effect [32], despite not optimizing for it.

for 100 epochs and noticed that training and dev error plateaued. The test metric reported is the average of 10-fold demographic group stratified cross validation accuracy along with the corresponding error bars denoting one standard deviation. The group identifiers present in the datasets were used to aggregate group Pareto loss during training.

Demographic Group Performance

The UCI Heart Disease dataset predicts angiographic disease status as a binary label (presence or absence of Heart Disease) using medical and demographic information. Age is binarized at a threshold of 40 years between young and old individuals, and gender is given to be binary (male/female) and are assigned as sensitive variables. The intersection of these 2 variables are considered sensitive demographic groups in our study. In Figure 2, we present group level performances for the UCI Heart Disease Dataset. Our approach of incorporating pareto efficiency leads to improvements in group level accuracies for all groups of the data by an average of 9.6%. We see improvements in the accuracy of predicting the presence of Heart disease in Table 2 by an average of 9.7% and that the relaxation of the demographic parity loss performs better than strict fairness constraints (Figure 3). This implies that improving based on demographic pareto efficiency obtains a better overall accuracy than even the baseline which explicitly optimizes overall accuracy on a held-out test set. This non-trivial result is due to the fact that when optimizing for overall accuracy on a training dataset, predictive models may incorrectly assume that the patterns in the majority group (Young/Male) might generalize to other demographic groups. We overcome this issue, and ensure that the demographic groups' accuracies are improved in an iterative manner as outlined in Algorithm 1. Since we use the decoupled classifiers' accuracies to measure the Pareto losses, and ensure that we incrementally train the joint model in such a way as to improve the accuracies of each of the groups (the training will terminate if individual group accuracies cannot be improved). This in turn has improved the overall test accuracy by overcoming issues of overfitting to the majority demographic group.

Trade-off Parameters

The choice to optimize overall accuracy as opposed to group-specific pareto efficiency cannot be made blindly. Hence, it is important to understand the impact of λ, α on the group-level accuracy-fairness trade-off. In Figure 4, we do a parameter sweep across values from 0 to 1 in increments of 0.1 and notice the changes in the overall accuracy, and the group-specific accuracies, along with the corresponding Parity and Pareto losses associated with the test evaluation. Based on this grid, the optimal choice of parameters (λ, α) based on overall accuracy is $(0, 0)$, whereas for each of the four demographic groups are $(0.6, 0.1)$, $(0.1, 0.4)$, $(0, 0)$, $(0.3, 0.2)$; whereas the choice for optimizing parity loss is $(0.4, 0.4)$ and the one for pareto loss is $(0.9, 0.5)$. These trade-offs further illustrate the choice required to be made by domain practitioners when adopting a classifier for predicting angiographic disease status. Table 2 and Figure 2 values are plotted with these parameters into account. We see that in some groups (e.g. Young/Male), the baseline without fairness based bias loss is comparable to a solution that maximizes that group's accuracy. Such baselines although pareto efficient, lie outside the region of relaxation in fairness weight permitted (Fairness Weight = 1-Parity Loss) and are hence not desirable.

5 DISCUSSION AND SIGNIFICANCE

Jointly Trained vs Decoupled Models: The choice of decoupled models in healthcare diagnosis needs to be made with careful consideration of the stratification dimensions. Decoupled models may be applicable when membership in a demographic group has been shown to have clinical significance. If the objective as presented is to maximize individual group level accuracies, one might be tempted to train a model for each strata separately. Our paper demonstrates the need for joint training across demographic strata to achieve pareto efficient fairness. Purely decoupled classifiers are optimal only under certain conditions of distributional uniformity and availability of data [19]. However, our approach works under a real-world skewed data setting where the data for all demographic groups might not be available uniformly, thereby rendering decoupled classifiers to be sub-optimal. When stratified by the chosen set of demographic group attributes, if there is no predictive model in the desired fairness region, our approach performs no worse than existing equality based constraints as our Pareto loss will be dominated by the high variance in loss between groups. In this scenario, our model would hence chose a low absolute Pareto loss, provided that α, the hyperparameter to trade-off between variance and total value of the Pareto loss is appropriately fine-tuned. Hence, to leverage the benefits of transfer learning, as shown in our evaluation it might be beneficial to bootstrap with decoupled classifiers and train jointly.

Demographic group stratification: The stratification we choose to optimize performance by, depends on what domain experts believe is clinically significant for the disease status diagnoses. For example, age and gender are known to be significant in angiographic disease status in patients, and hence there is a possibility for us to learn different decoupled models. Other possible demographic group stratification can be done based on race and geographical location, as coronary artery disease has been shown to be harder to diagnose in black populations [33], and that there is a difference in angiographic profiles across patients from different geographical locations in Asia and South America [34, 35, 36, 37].

Other definitions of Fairness and Individual Accountability: Notions of pareto efficiency are compatible with assumptions of individual fairness. Utilizing our methodology, the domain expert can make an informed choice among different Pareto efficient models. We have demonstrated that achieving Demographic Pareto Efficiency has benefits and yields classifiers that outperform the baselines for overall and *all* group accuracy. Individual instance based fairness definitions often compare diagnosis and outcomes of one patient with similar patients in a dataset or counterfactual scenarios. However, defining the dimensions of similarity between individuals can be quite challenging for a specific disease type, and should consider the variations of disease prognosis along the same dimensions such as demographic information and co-morbidities. In the event where multiple demographically Pareto Fair operating points are discovered on the Pareto frontier [38], domain experts should choose the right operating point among them by incorporating other procedural steps to mitigate the discriminatory outcomes earlier in the process. Further, pareto efficiency improves the accuracy of minority and under-represented protected demographic groups when compared to unconstrained classifiers, which may implicitly allow the dominance of majority demographic groups when overall accuracy is optimized. While our methodology does not completely eliminate discriminatory biases, Demographic Pareto Efficiency and the choices around it can provide more transparency and understanding of the structural and socio-technical causes behind unfairly distributed datasets and models, which can improve the contestability of ML predictive models.

Other Heart Diseases: In addition to the diagnostic task of angiographic disease status, we see this choice between decoupled and jointly trained models emerge in other heart disease tasks too. In a cardiology study of over 4000 ER patients with cardiac event symptoms [17], no symptoms were found to be predictive of a heart attack in white women. In black males, only an unrelated symptom (diaphoresis) was found to be indicative of a future cardiac event with 95 percent confidence, while in white males, relevant features (left arm radiation, pressure, tightness) were detected as indicators with high accuracy.

6 CONCLUSION

The choice between decoupled and jointly trained diagnostic models for angiographic disease status is critical for positive health outcomes in demographic groups. We have shown that by optimizing for Demographic Pareto Efficiency, the choice between decoupled and jointly trained models can be further broken down to choice of classifiers that have Pareto optimal performance across the demographic groups. As the Pareto front is unknown, we show that by incorporating a heuristic based on Pareto Efficient Fairness in training a combination of decoupled and jointly trained models, we achieve better overall and individual demographic group level accuracy as compared to other constraints in decoupled and jointly trained models. We demonstrate empirically that our approach achieves Demographic Pareto Efficiency by improving overall and subgroup accuracy by up to 9.7% and 9.6% respectively in the UCI Heart Disease dataset.

Competing Interests and Funding

Prof. Subramanian and Prof. Jagabathula are co-founders of Velai Inc. They reports that Velai Inc broadly works in the area of socio-economic predictive models. Prof. Subramanian is also a co-founder of Entrupy Inc, and Gaius Networks Inc and has served as a consultant for the World Bank and the Governance Lab. Mr. Balashankar is a Ph.D student at New York University, and is also funded in part, by the Google Student Research Advising Program and a summer internship at Google AI with Dr. Lees. No other disclosures were reported.

References

[1] J. Vacca and H. Rosenthal, "A local law in relation to automated decision systems used by agencies," 2018. [Online]. Available: http://legistar.council.nyc.gov/LegislationDetail.aspx?ID=3137815&GUID=437A6A6D-62E1-47E2-9C42-461253F9C6D0

[2] P. J. Bickel, E. A. Hammel, and J. W. O'Connell, "Sex bias in graduate admissions: Data from berkeley," *Science*, vol. 187, no. 4175, pp. 398–404, 1975. [Online]. Available: http://science.sciencemag.org/content/187/4175/398

[3] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, "Men also like shopping: Reducing gender bias amplification using corpus-level constraints," in *EMNLP*, 2017.

[4] H. Heidari, C. Ferrari, K. Gummadi, and A. Krause, "Fairness behind a veil of ignorance: Welfare analysis for automated decision making," vol. abs/1806.04959, '18. [Online]. Available: http://arxiv.org/abs/1806.04959

[5] T. Bolukbasi, K. Chang, J. Y. Zou, V. Saligrama, and A. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," *CoRR*, vol. abs/1607.06520, 2016. [Online]. Available: http://arxiv.org/abs/1607.06520

[6] K. H. Yiu, F. R. de Graaf, J. D. Schuijf, J. M. van Werkhoven, N. A. Marsan, C. E. Veltman, A. de Roos, A. Pazhenkottil, L. J. Kroft, E. Boersma, B. Herzog, M. Leung, E. Maffei, D. Y. Leung, P. A. Kaufmann, F. Cademartiri, J. J. Bax, and J. W. Jukema, "Age- and gender-specific differences in the

prognostic value of ct coronary angiography," *Heart*, vol. 98, no. 3, pp. 232–237, 2012. [Online]. Available: https://heart.bmj.com/content/98/3/232

[7] N. Smilowitz, A. Mahajan, M. Roe, A. Hellkamp, K. Chiswell, M. Gulati, and H. Reynolds, "Mortality of myocardial infarction by sex, age, and obstructive coronary artery disease status in the action registry–gwtg (acute coronary treatment and intervention outcomes network registry–get with the guidelines)," *Circulation: Cardiovascular Quality and Outcomes*, vol. 10, p. e003443, 12 2017.

[8] A. M. Mahajan, H. Gandhi, N. R. Smilowitz, M. T. Roe, A. S. Hellkamp, K. Chiswell, M. Gulati, and H. R. Reynolds, "Seasonal and circadian patterns of myocardial infarction by coronary artery disease status and sex in the action registry-gwtg," *International Journal of Cardiology*, vol. 274, pp. 16–20, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167527318337781

[9] R. Prashanth, "Finding association between lipid profile and demographic and disease status of patients undergoing coronary angiography: A retrospective study in rural South India," *JOURNAL OF INDIAN COLLEGE OF CARDIOLOGY*, vol. 11, no. 2, pp. 62–65, 2021. [Online]. Available: https://www.joicc.org/article.asp?issn=1561-8811;year=2021;volume=11;issue=2;spage=62;epage=65;aulast=Prashanth;t=6

[10] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and abstraction in sociotechnical systems," ser. FAT* '19, 2019.

[11] N. Grgic-Hlaca, "The case for process fairness in learning : Feature selection for fair decision making," 2016.

[12] G. N. Rothblum and G. Yona, "Probably approximately metric-fair learning," *CoRR*, vol. abs/1803.03242, 2018. [Online]. Available: http://arxiv.org/abs/1803.03242

[13] D. Madras, E. Creager, T. Pitassi, and R. Zemel, "Fairness through causal awareness: Learning causal latent-variable models for biased data," ser. FAT* '19, 2019.

[14] N. Grgic-Hlaca, E. M. Redmiles, K. P. Gummadi, and A. Weller, "Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction," ser. WWW '18, 2018, pp. 903–912. [Online]. Available: https://doi.org/10.1145/3178876.3186138

[15] A. Chakraborty, J. Messias, F. Benevenuto, S. Ghosh, N. Ganguly, and K. P. Gummadi, "Who makes trends? understanding demographic biases in crowdsourced recommendations," *CoRR*, vol. abs/1704.00139, 2017. [Online]. Available: http://arxiv.org/abs/1704.00139

[16] T. Speicher, M. Ali, G. Venkatadri, F. N. Ribeiro, G. Arvanitakis, F. Benevenuto, K. P. Gummadi, P. Loiseau, and A. Mislove, "Potential for discrimination in online targeted advertising," in *ACM FAccT*, 2018.

[17] A. Allabban, J. Hollander, and J. Pines, “Gender, race and the presentation of acute coronary syndrome and serious cardiopulmonary diagnoses in ed patients with chest pain,” in *Emergency Medicine Journal*, vol. 34, 2017, pp. 653–658.

[18] A. K. Menon and R. C. Williamson, “The cost of fairness in binary classification,” ser. FAT* 2018, 2018.

[19] C. Dwork, N. Immorlica, A. T. Kalai, and M. Leiserson, “Decoupled classifiers for group-fair and efficient machine learning,” ser. ACM FAccT, 2018, pp. 119–133.

[20] V. Pareto, *Manuale di economia politica: con una introduzione alla scienza sociale*. Società editrice libraria, 1919, vol. 13.

[21] G. Debreu, “Valuation equilibrium and pareto optimum,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 40, no. 7, p. 588, 1954.

[22] A. Mas-Colell, M. D. Whinston, J. R. Green *et al.*, “Chapter 16: Equilibrium and its basic welfare properties,” *Microeconomic Theory*, vol. 1, 1995.

[23] P. Godfrey, R. Shipley, and J. Gryz, “Algorithms and analyses for maximal vector computation,” *The VLDB Journal*, vol. 16, no. 1, pp. 5–28, Jan. 2007. [Online]. Available: http://dx.doi.org/10.1007/s00778-006-0029-7

[24] A. Beutel, J. Chen, Z. Zhao, and E. H. Chi, “Data decisions and theoretical implications when adversarially learning fair representations,” *CoRR*, vol. abs/1707.00075, 2017.

[25] N. Martinez, M. Bertran, and G. Sapiro, “Minimax pareto fairness: A multi objective perspective,” 2020.

[26] N. Papernot, M. Abadi, Ú. Erlingsson, I. J. Goodfellow, and K. Talwar, “Semi-supervised knowledge transfer for deep learning from private training data,” *CoRR*, vol. abs/1610.05755, 2016. [Online]. Available: http://arxiv.org/abs/1610.05755

[27] B. Woodworth, S. Gunasekar, M. I. Ohannessian, and N. Srebro, “Learning non-discriminatory predictors,” in *COLT*, 2017.

[28] Z. Lipton, J. McAuley, and A. Chouldechova, “Does mitigating ml's impact disparity require treatment disparity?” in *NeurIPS 31*, 2018.

[29] N. Kilbertus, A. Gascon, M. Kusner, M. Veale, K. Gummadi, and A. Weller, “Blind justice: Fairness with encrypted sensitive attributes,” ser. ICML, 2018.

[30] E. E. Eban, M. Schain, A. Mackey, A. Gordon, R. A. Saurous, and G. Elidan, “Scalable Learning of Non-Decomposable Objectives,” *ArXiv e-prints*, Aug. 2016.

[31] S. Mannor, D. Peleg, and R. Rubinstein, “The cross entropy method for classification,” ser. ICML ’05, 2005.

[32] G. Pleiss, M. Raghavan, F. Wu, J. M. Kleinberg, and K. Q. Weinberger, “On fairness and calibration,” *CoRR*, vol. abs/1709.02012, 2017. [Online]. Available: http://arxiv.org/abs/1709.02012

[33] B. E. Simmons, A. Castaner, A. Campo, J. Ferlinz, M. Mar, and R. Cooper, “Coronary artery disease in blacks of lower socioeconomic status: Angiographic findings from the cook county hospital heart disease registry,” *American Heart Journal*, vol. 116, no. 1, Part 1, pp. 90–97, 1988. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0002870388902542

[34] E. Sato, F. Hatta, M. Levy-Neto, and S. Fernandes, “Demographic, clinical, and angiographic data of patients with takayasu arteritis in brazil,” *International Journal of Cardiology*, vol. 66, pp. S67–S70, 1998. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167527398001521

[35] G. L. Gierach, B. D. Johnson, C. N. B. Merz, S. F. Kelsey, V. Bittner, M. B. Olson, L. J. Shaw, S. Mankad, C. J. Pepine, S. E. Reis, W. J. Rogers, B. L. Sharaf, and G. Sopko, “Hypertension, menopause, and coronary artery disease risk in the women&#x2019;s ischemia syndrome evaluation (wise) study,” *Journal of the American College of Cardiology*, vol. 47, no. 3_Supplement, pp. S50–S58, 2006.

[36] A. M. Mohammad, H. H. Rashad, Q. S. Habeeb, B. H. Rashad, and S. Y. Saeed, “Demographic, clinical and angiographic profile of coronary artery disease in kurdistan region of Iraq,” *Am J Cardiovasc Dis*, vol. 11, no. 1, pp. 39–45, 2021.

[37] B. Prakash, A. Jaiswal, and M. M. Shah, “Demographic & angiographic profile of young patients aged 40 year & less undergoing coronary angiography in a tier ii city of eastern india,” *J Family Med Prim Care*, vol. 9, no. 10, pp. 5183–5187, Oct 2020.

[38] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, K. P. Gummadi, and A. Weller, “From parity to preference-based notions of fairness in classification,” 2017.

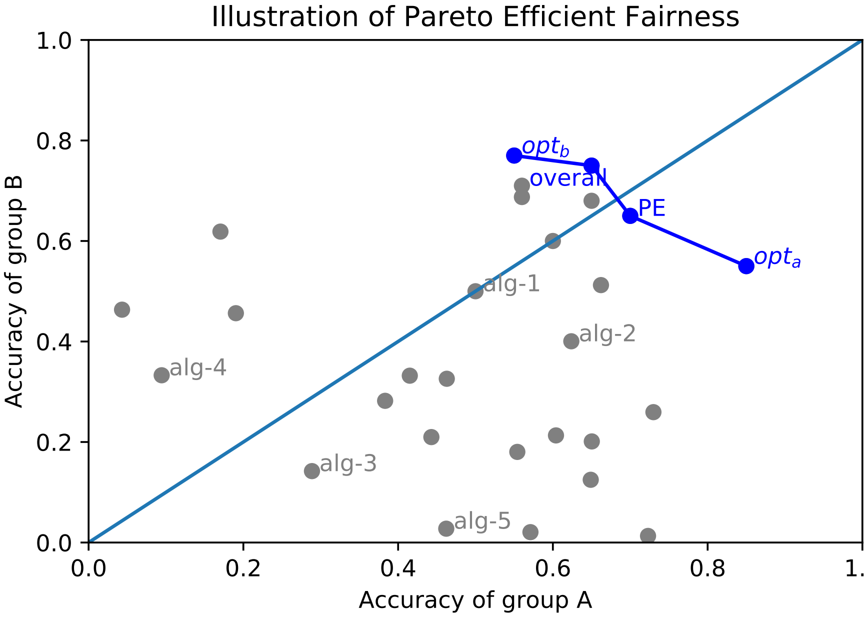


Figure 1: Illustration of Demographic Pareto Efficiency on synthetic data. Each point in the scatter plot corresponds to the group level accuracies of machine learning (ML) algorithms (alg-[1-5] indicated in grey) over groups A and B. The best performing ML algorithm with Demographic Parity yields accuracy metrics of $(0.60, 0.60)$ on groups a, b respectively. If accuracy for each of the groups is separately maximized, we would select points $opt_a = (0.83, 0.55)$, and $opt_b = (0.63, 0.77)$. Discovering all the Demographic Pareto Efficient classifiers gives us the Pareto front (dots in blue). Among these Demographic Pareto Efficient classifiers, we could choose $PE = (0.71, 0.63)$ (in blue and green), if our objective was to improve the accuracy metrics of both groups, with minimal deviation from optimal per-group accuracies (pareto loss).

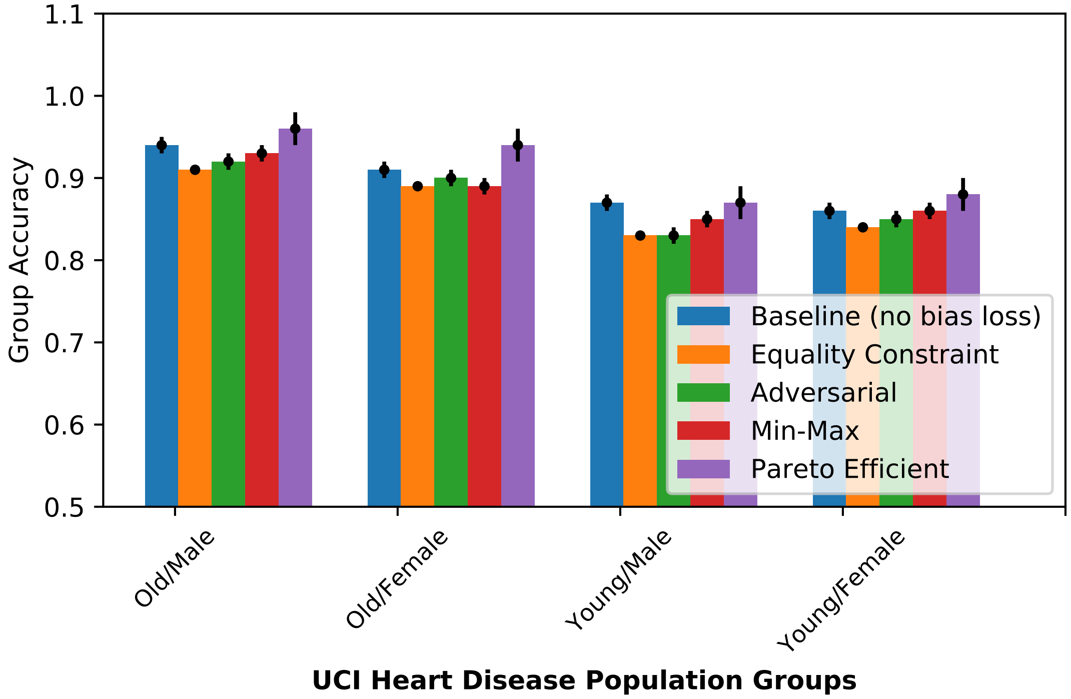


Figure 2: Group accuracy comparison showing we achieve Demographic Pareto Efficient group level accuracy for all groups in UCI Heart Disease dataset among constrained classifiers.

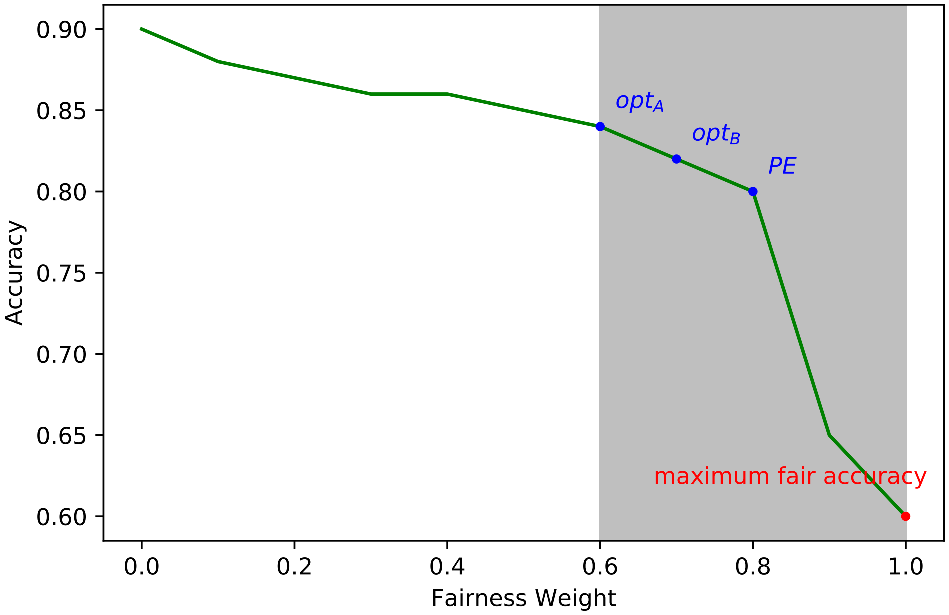
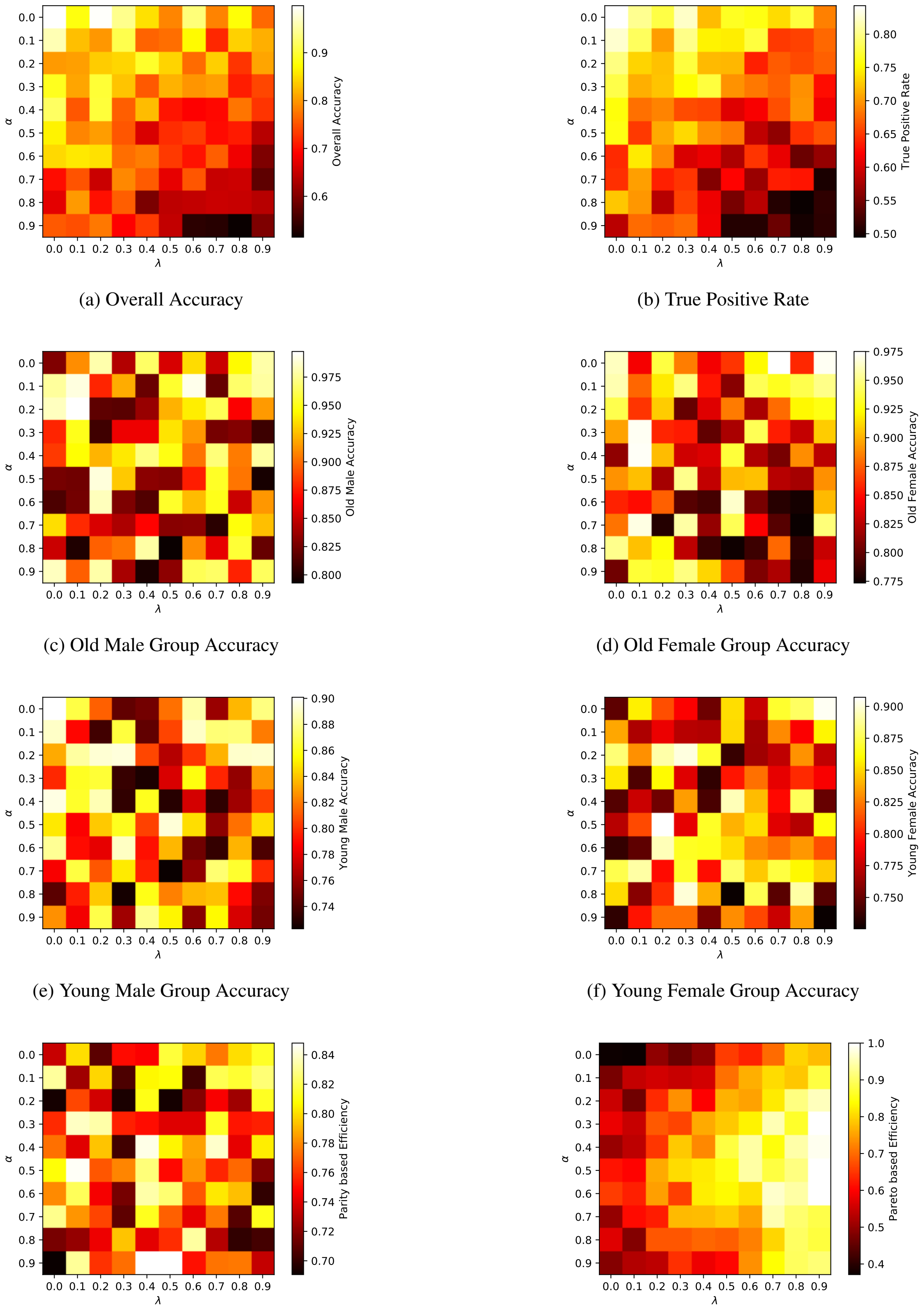


Figure 3: Relationship between the shape of the fairness frontier and the efficiency gain expected by using PEF in UCI Heart Disease dataset. y-axis denotes the maximum achievable overall accuracy for a given fairness weight (x-axis). A fairness weight of 1.0 does not permit deviation from the strict equality constraint, wherease a fairness weight of 0.0 is unconstrained and allows higher model performances. However, better accuracies are achievable by relaxing the strict equality constraint by a small amount (gray region) and using PEF.



(a) Overall Accuracy

(b) True Positive Rate

(c) Old Male Group Accuracy

(d) Old Female Group Accuracy

(e) Young Male Group Accuracy

(f) Young Female Group Accuracy

(g) Fairness Weight (1 - Parity Loss)

(h) Pareto Efficiency (1 - Pareto Loss)

Figure 4: Trade-offs between choosing parameters λ and α depends on the group-level versus overall measures chosen by the domain practitioner. Given the prior work that advocates for improving each of the demographic group's accuracy on the Pareto front, we chose our model to optimize Pareto Efficient Fairness (h).