Loss-Conditioned Utility-Fairness Boundary Modeling for Medical Imaging

Anonymous Authors

Anonymous Institution

Abstract. Medical image classifiers can achieve strong global performance while exhibiting different error profiles across sensitive groups. We propose a loss-conditioned framework for efficient utility-fairness boundary modeling in medical image classification. The method combines binary cross-entropy with a differentiable Equalized Odds loss and uses a two-dimensional preference vector to control their relative weights. Independently trained models define reference frontiers, while a single loss-conditioned backbone receives the same preference vector and is queried at both training and unseen preferences. We evaluate the framework on Harvard-GF-3300 glaucoma detection with gender as the sensitive attribute using ResNet-18, EfficientNet-B1, DenseNet-121, and ViT-Small. Conditioned models are trained with nine preference vectors and evaluated on denser aligned grids to assess interpolation. Across architectures, conditioned models approximate useful regions of independently trained ACC-EOD frontiers and enable dense aligned preference queries without additional optimization. These results support loss-conditioned learning as an efficient frontier-exploration tool rather than a single-point accuracy maximization method. Source code will be made publicly available upon acceptance.

Keywords: Fairness in medical imaging · Utility-fairness trade-off · Multi-objective optimization · Equalized odds · Loss-conditioned training · FiLM · Pareto frontier exploration

1 Introduction

Fairness-aware medical imaging should model the utility-fairness spectrum rather than select a single fairness-corrected model. Clinical deployment rarely has a universally optimal operating point: screening, diagnosis, and triage may require different compromises between diagnostic performance and subgroup error parity. Models with strong global discrimination can still exhibit unequal false positive or false negative behavior across sensitive groups. We therefore study fairness-aware learning as a multi-objective utility-fairness boundary modeling problem, where the goal is to characterize the attainable trade-off spectrum rather than optimize one fixed operating point.

We construct reference utility-fairness boundaries using a differentiable Equalized Odds loss. Equalized Odds is clinically aligned with medical image classification because it compares true positive and false positive behavior across sensitive

groups [4], but its thresholded metric form is not directly differentiable. We use a soft Equalized Odds surrogate based on group-conditional predicted positive probabilities and combine it with binary cross-entropy in a weighted bi-objective loss. Independently trained models over predefined loss preferences define reference frontiers, while thresholded Equalized Odds Difference (EOD) provides the corresponding evaluation metric.

Training one model per preference reveals the boundary but scales inefficiently: a nine-point grid across four architectures already requires $9 \times 4 = 36$ independently trained models before adding datasets, sensitive attributes, or denser grids. Inspired by You Only Train Once (YOTO) loss-conditioned training [3], we train a single Feature-wise Linear Modulation (FiLM)-conditioned model [12] that receives the utility-fairness preference vector as input and can be queried at training and unseen preferences. This enables us to test whether one conditioned model can approximate independently trained frontiers and interpolate to denser aligned preference grids without additional training.

We address one main research question: *Can a single FiLM-conditioned model, trained with a differentiable Equalized Odds loss, approximate and interpolate the utility-fairness boundary of independently trained weighted-loss baselines in medical imaging?*

The contributions of this work are threefold. First, we formulate fairness-aware medical image classification as multi-objective utility-fairness boundary modeling using a differentiable Equalized Odds loss. Second, we compare independently trained reference frontiers with FiLM-conditioned counterparts and reported Harvard-GF-3300 baselines in ACC-EOD space. Third, we test whether a single nine-condition model can approximate reference frontiers and expose denser intermediate operating points through aligned interpolation grids without additional training.

2 Related Work

Fairness in medical imaging has increasingly emphasized subgroup reliability alongside global diagnostic performance. Prior work has shown that medical imaging models can exhibit subgroup performance differences across gender, race, age, and acquisition context, even when global performance is strong [7, 9]. This has motivated fairness analyses across medical imaging, where subgroup area under the curve (AUC), sensitivity, specificity, and error rates are clinically relevant indicators of reliability [8]. Our work follows this direction but models attainable utility-fairness boundaries rather than selecting a single mitigated model.

Multi-objective fairness evaluation and loss-conditioned learning provide the basis for our boundary-modeling framework. When utility and fairness objectives conflict, Pareto fronts, non-dominated solutions, hypervolume, and frontier coverage summarize trade-off behavior more completely than a single scalar score [14]. YOTO trains one model over a distribution of loss parameters and conditions inference on the selected loss weights [3], while FiLM provides a

lightweight mechanism for injecting such conditioning into intermediate features [12]. The closest recent work uses YOTO for fairness-accuracy tradeoff auditing with statistical guarantees [13]; in contrast, we study end-to-end medical image classification with a differentiable Equalized Odds loss, independently trained reference frontiers, FiLM-conditioned approximation, and unseen-condition interpolation for efficient ACC-EOD frontier exploration.

3 Methodology

3.1 Problem Formulation

We formulate fairness-aware medical image classification as bi-objective optimization. Let $\mathcal{D} = \{(x_i, y_i, a_i)\}_{i=1}^{n}$, where x_i is an image, $y_i \in \{0, 1\}$ is the disease label, and $a_i \in \mathcal{A}$ is a sensitive attribute. A model f_θ outputs logit $s_i = f_\theta(x_i)$ and probability $p_i = \sigma(s_i)$. We optimize binary cross-entropy and a differentiable Equalized Odds loss,

$$\mathbf{L}(\theta) = [\mathcal{L}_{\mathrm{BCE}}(\theta), \mathcal{L}_{\mathrm{EO}}(\theta)]\,. \tag{1}$$

Preference vectors define the utility-fairness spectrum. For a grid size K, define

$$\mathcal{W}_K = \left\{ w_k = (1 - \lambda_k, \lambda_k)\,, \lambda_k = \frac{k}{K-1} : k = 0, \ldots, K-1 \right\}. \tag{2}$$

For $w_k = (w_{k,1}, w_{k,2})$, the linearly scalarized objective is

$$\mathcal{L}(\theta; w_k) = w_{k,1}\mathcal{L}_{\mathrm{BCE}}(\theta) + w_{k,2}\mathcal{L}_{\mathrm{EO}}(\theta). \tag{3}$$

We train models with $K_{\mathrm{train}} = 9$ preference vectors and query conditioned models at $K_{\mathrm{eval}} \in \{9, 25, 49, 81, 121\}$. These aligned refinements preserve endpoints and all training preferences because $K_{\mathrm{eval}} - 1$ is divisible by $K_{\mathrm{train}} - 1 = 8$, while adding unseen intermediate locations without selecting grids based on test performance.

Base models define reference frontiers, while conditioned models learn a shared frontier. For each $w_k \in \mathcal{W}_{K_{\mathrm{train}}}$, the base setting trains an independent model

$$\theta_k^\star = \arg\min_\theta \mathcal{L}(\theta; w_k). \tag{4}$$

The conditioned model instead receives both image and preference vector, $f_\theta(x, w)$, and is trained as

$$\theta_{\mathrm{cond}}^\star = \arg\min_\theta \mathbb{E}_{w \sim \mathrm{Unif}(\mathcal{W}_{K_{\mathrm{train}}})} \left[w_1 \mathcal{L}_{\mathrm{BCE}}(\theta; w) + w_2 \mathcal{L}_{\mathrm{EO}}(\theta; w) \right]. \tag{5}$$

where $\mathrm{Unif}(\mathcal{W}_{K_{\mathrm{train}}})$ denotes uniform sampling over the nine training preference vectors. At test time, base and conditioned models are compared in the ACC-EOD metric space and summarized using Pareto-frontier indicators.

3.2 Differentiable Equalized Odds Loss

The EO loss penalizes group-conditional prediction differences within each ground-truth class. For group a and label y, define

$$\mu_{a,y} = \frac{1}{|\mathcal{I}_{a,y}|} \sum_{i \in \mathcal{I}_{a,y}} p_i, \quad \mu_y = \frac{1}{|\mathcal{I}_y|} \sum_{i \in \mathcal{I}_y} p_i, \tag{6}$$

where $\mathcal{I}_{a,y} = \{i : a_i = a, y_i = y\}$ and $\mathcal{I}_y = \{i : y_i = y\}$. The differentiable EO loss is

$$\mathcal{L}_{\text{EO}} = \frac{1}{|\mathcal{S}|} \sum_{(a,y) \in \mathcal{S}} |\mu_{a,y} - \mu_y|^q, \quad q \in \{1, 2\}, \tag{7}$$

where $\mathcal{S}$ contains non-empty group-label pairs. We use $q = 1$ unless otherwise stated.

The soft EO loss is designed to align with thresholded EOD. For evaluation, EOD is computed from binary predictions as

$$\text{EOD} = \max\left(\max_{a,a'} |\text{TPR}_a - \text{TPR}_{a'}|, \max_{a,a'} |\text{FPR}_a - \text{FPR}_{a'}| \right). \tag{8}$$

The loss uses probabilities for differentiability, while the reported metric uses thresholded predictions for evaluation.

3.3 Loss-Conditioned FiLM Architectures

Conditioned models inject the utility-fairness preference through lightweight FiLM modulation. Given an intermediate activation tensor T_ℓ at layer ℓ, FiLM computes

$$\text{FiLM}_\ell(T_\ell, w) = \gamma_\ell(w) \odot T_\ell + \beta_\ell(w), \tag{9}$$

where a two-layer MLP maps the two-dimensional preference vector w to FiLM parameters. For a feature block with C channels, the MLP has hidden dimension 64 and outputs $2C$ channel-wise scale and shift parameters. The final FiLM projection is zero-initialized so that $\gamma_\ell(w) = 1$ and $\beta_\ell(w) = 0$, making FiLM start as an identity transformation. Base and conditioned models use the same ImageNet-pretrained backbone initialization and are fully fine-tuned. For conditioned models, FiLM MLP parameters use a fixed $10\times$ larger learning rate than the backbone to accelerate adaptation from identity initialization.

FiLM is inserted without changing the main diagnostic backbone. For ResNet-18, FiLM is applied inside residual blocks; for EfficientNet-B1, after feature stages; for DenseNet-121, to newly produced dense-layer features before concatenation; and for ViT-Small, to attention and MLP residual branches. All conditioned models use the same two-dimensional utility-fairness preference vector and are evaluated by sweeping this vector at test time.

3.4 Utility-Fairness Trade-off Control

A utility-fairness trade-off occurs when improving utility worsens, or fails to improve, fairness over part of the attainable model set. For utility $U(\theta)$ and fairness violation $F(\theta)$, this means $U(\theta_i) > U(\theta_j)$ but $F(\theta_i) > F(\theta_j)$ for some $\theta_i, \theta_j \in \Theta$. Prior work has shown that fairness constraints can impose measurable costs in binary classification and that different fairness criteria may be difficult to satisfy simultaneously [4, 10, 6]. However, we do not assume that a trade-off must appear in every setting; we empirically characterize the resulting frontier.

Equalized Odds can be viewed as restricting the feasible set of predictors. For tolerance ϵ, define

$$\Theta_\epsilon = \Big\{ \theta \in \Theta : \max_{a,a' \in \mathcal{A},\, y \in \{0,1\}} \Big| \Pr(\hat{Y}_\theta = 1 \mid A = a, Y = y) - \Pr(\hat{Y}_\theta = 1 \mid A = a', Y = y) \Big| \le \epsilon \Big\}. \tag{10}$$

The constrained optimum is

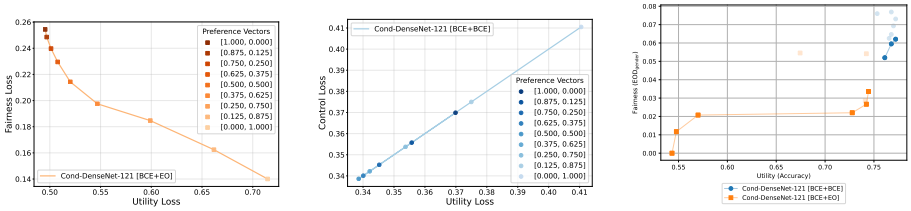
$$\theta^\star_\epsilon = \arg\max_{\theta \in \Theta_\epsilon} U(\theta). \tag{11}$$

As ϵ decreases, the feasible set becomes more restrictive, so the best achievable utility may change. The weighted BCE+EO objective searches different attainable regions, but the trade-off is established only by the resulting metric frontier.

An aligned-objective control tests whether condition-dependent FiLM modulation alone can create an apparent frontier. Replacing the EO term with a second BCE term gives

$$\mathcal{L}_{\text{ctrl}}(\theta; w_k) = w_{k,1}\mathcal{L}_{\text{BCE}}(\theta) + w_{k,2}\mathcal{L}_{\text{BCE}}(\theta) = \mathcal{L}_{\text{BCE}}(\theta), \tag{12}$$

because $w_{k,1} + w_{k,2} = 1$. Thus, the control still receives w_k through FiLM, but all conditions optimize the same effective objective. Fig. 1 compares conditioned DenseNet-121 under BCE+EO and BCE+BCE training. In Fig. 1a, BCE+EO



(a) BCE+EO train loss (b) BCE+BCE train loss (c) Test ACC-EOD

Fig. 1: **Aligned-objective control on conditioned DenseNet-121.** BCE+BCE isolates FiLM conditioning under an identical effective objective, while BCE+EO produces broader movement toward lower EOD.

forms an ordered loss-space trajectory, moving toward lower fairness loss as EO weight increases. In contrast, Fig. 1b shows irregular BCE+BCE scatter because all conditions optimize the same effective objective. At test time, Fig. 1c shows that BCE+BCE collapses to a narrow ACC-EOD region, whereas BCE+EO moves toward lower EOD, indicating that the observed boundary is driven by the utility-fairness objective rather than FiLM conditioning alone.

4 Experimental Analysis

4.1 Utility-Fairness Boundary Analysis

We evaluate whether one loss-conditioned model can approximate fairness boundaries that otherwise require independently trained models on Harvard-GF-3300 gender. For each of four backbones, nine independent base models define the reference frontier, while one conditioned model is trained on the same preferences and queried at aligned grids $K_{\text{eval}} \in \{9, 25, 49, 81, 121\}$. All runs use 20 epochs; conditioned runs include a two-epoch $w = (1.000, 0.000)$ warm-up, replacing a 9×20-epoch base frontier with one conditioned run. Base-conditioned pairs use matched hyperparameters and batch size 256; EMA is used only for Cond-ResNet-18 and Cond-DenseNet-121, with validation-selected settings fixed for test evaluation.

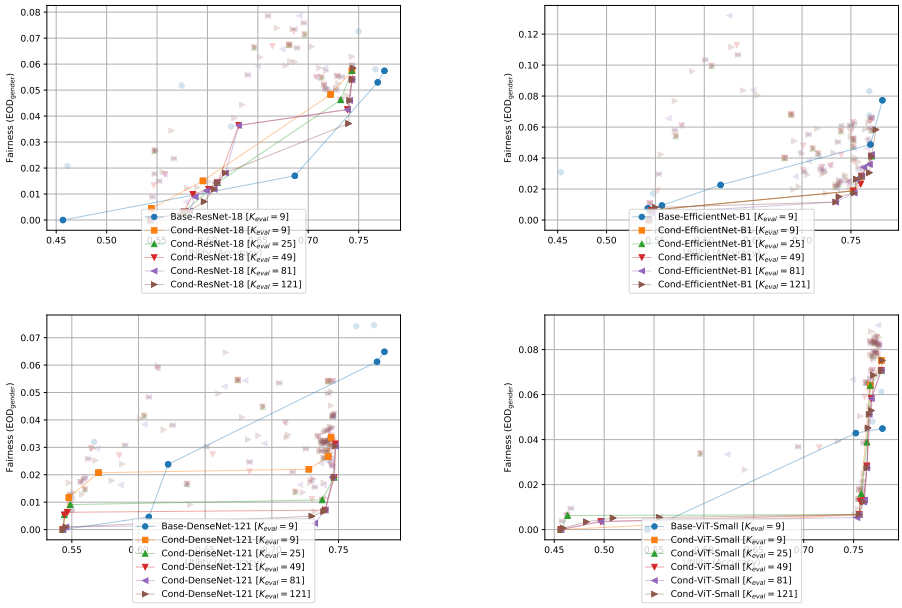


Fig. 2: **Experimental ACC-EOD boundaries on Harvard-GF-3300 gender.** Base and conditioned frontiers are shown; higher ACC and lower EOD are preferred. Lines indicate non-dominated frontiers computed using Fairical [14].

Table 1: **Pareto-frontier indicators on the test split.** K_{eval} denotes the number of queried preference vectors. NDS, R-ONVG, ONVGR, UD, AS, HV, and A summarize non-dominated frontier capacity, distribution, spread, hypervolume, and aggregate trade-off quality.

System	K_{eval}	NDS ↑	R-ONVG ↑	ONVGR ↑	UD ↑	AS ↑	HV ↑	A ↑
Base-ResNet-18	9	4	0.36	0.44	0.63	0.19	0.77	0.20
Cond-ResNet-18	9	5	0.45	0.83	0.65	0.13	0.74	0.29
Cond-ResNet-18	25	5	0.45	0.36	1.00	0.13	0.74	0.22
Cond-ResNet-18	49	8	0.73	0.31	0.68	0.13	0.74	0.23
Cond-ResNet-18	81	11	1.00	0.26	0.59	0.13	0.74	0.26
Cond-ResNet-18	121	11	1.00	0.18	0.59	0.13	0.74	0.24
Base-EfficientNet-B1	9	5	0.45	0.56	1.00	0.15	0.77	0.29
Cond-EfficientNet-B1	9	4	0.36	0.57	0.63	0.12	0.76	0.20
Cond-EfficientNet-B1	25	5	0.45	0.29	0.65	0.13	0.77	0.17
Cond-EfficientNet-B1	49	6	0.55	0.18	0.61	0.13	0.77	0.16
Cond-EfficientNet-B1	81	9	0.82	0.17	0.59	0.13	0.77	0.21
Cond-EfficientNet-B1	121	11	1.00	0.14	0.59	0.15	0.77	0.24
Base-DenseNet-121	9	5	0.71	0.62	0.65	0.15	0.77	0.32
Cond-DenseNet-121	9	6	0.86	0.67	0.66	0.12	0.74	0.36
Cond-DenseNet-121	25	5	0.71	0.21	0.54	0.11	0.74	0.19
Cond-DenseNet-121	49	6	0.86	0.13	0.48	0.12	0.75	0.19
Cond-DenseNet-121	81	6	0.86	0.08	0.66	0.12	0.75	0.19
Cond-DenseNet-121	121	7	1.00	0.07	0.59	0.12	0.75	0.20
Base-ViT-Small	9	3	0.21	0.60	1.00	0.14	0.77	0.23
Cond-ViT-Small	9	4	0.29	0.67	1.00	0.20	0.77	0.29
Cond-ViT-Small	25	7	0.50	0.44	0.65	0.20	0.77	0.23
Cond-ViT-Small	49	8	0.57	0.28	0.49	0.20	0.77	0.20
Cond-ViT-Small	81	10	0.71	0.22	0.45	0.20	0.78	0.21
Cond-ViT-Small	121	14	1.00	0.21	0.56	0.20	0.78	0.27

ACC-EOD plots use a fixed threshold of 0.5, unweighted BCE, and no class or positive-class adjustment; differences across operating points are therefore driven by BCE+EO preferences rather than threshold tuning or imbalance reweighting. Fig. 2 shows that conditioned models need not dominate every independent point to be useful: EfficientNet-B1, DenseNet-121, and ViT-Small move toward lower-EOD regions while retaining usable thresholded utility. Overall, differentiable BCE+EO produces non-degenerate reference frontiers, and a single FiLM-conditioned model approximates useful regions of these boundaries.

Pareto indicators quantify how interpolation changes frontier resolution. Table 1 shows that increasing K_{eval} generally discovers more non-dominated solutions (NDS) without retraining: NDS increases from 5 to 11 for Cond-ResNet-18, from 4 to 11 for Cond-EfficientNet-B1, and from 4 to 14 for Cond-ViT-Small. HV remains stable across dense queries, indicating that interpolation mainly refines the explored boundary rather than creating a new frontier.

Fig. 3 compares learned frontiers with reported Harvard-GF-3300 baselines. Reported baselines are shown as single points because preference-swept frontiers are unavailable. The selected conditioned ViT-Small point gives a low-disparity operating region, with lower EOD than the reported baselines, moderate ACC, closely matched sensitivity across gender groups (83.33% vs. 82.67%), identical specificity (66.67% vs. 66.67%), and small subgroup gaps (ΔBA = 0.33, ΔAUC = 0.28 percentage points).

Overall, the proposed method is best interpreted as fairness-boundary modeling rather than single-point fairness optimization. The comparative plot shows

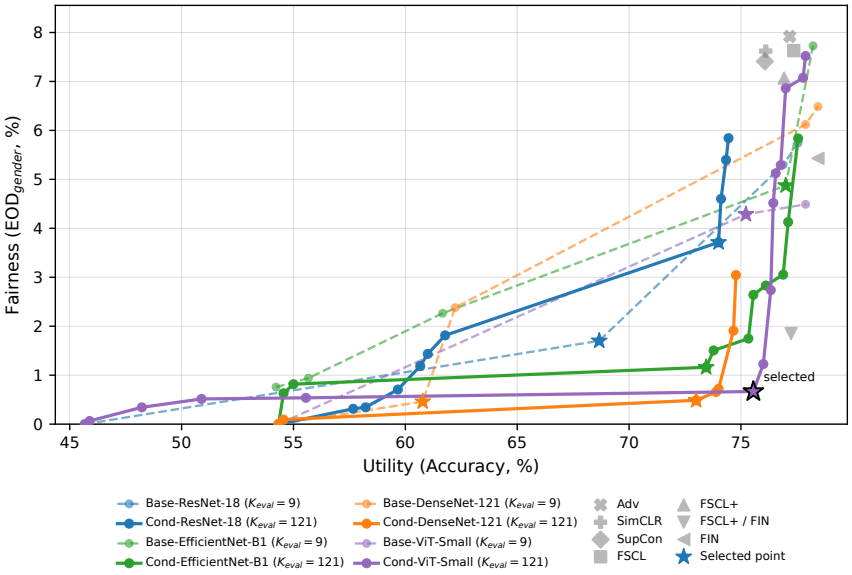


Fig. 3: **Utility-fairness comparison with reported Harvard-GF-3300 baselines.** Base and conditioned models are shown as ACC-EOD frontiers, while reported methods are single operating points: Adv [1], SimCLR [2], SupCon [5], FSCL/FSCL+ [11], FIN and FSCL+ / FIN [8]. Stars mark closest-to-ideal representative points; the outlined star marks the selected balanced point.

that no single operating point dominates all objectives: literature baselines may achieve higher ACC, while conditioned frontiers expose lower-disparity points. The main advantage is lower-cost frontier exploration: one loss-conditioned 20-epoch model exposes multiple utility-fairness operating points, including dense interpolated conditions, whereas a nine-point independent frontier requires nine separately trained models.

5 Conclusion

This study formulates fairness-aware medical image classification as modeling the utility-fairness boundary rather than single-point debiasing. Using a differentiable Equalized Odds loss, we construct reference frontiers and show that loss-conditioned FiLM models approximate useful regions of these frontiers with one trainable system. Querying aligned unseen preference grids enables higher-resolution boundary exploration without additional optimization. Overall, the framework provides a lower-cost alternative to exhaustive independent training for studying utility-fairness trade-offs, rather than a method aimed primarily at maximizing single-point accuracy. This may be especially valuable for larger backbones and foundation-style medical imaging models, where repeatedly training separate models for each utility-fairness preference can be computationally challenging.

References

1. Beutel, A., Chen, J., Zhao, Z., Chi, E.H.: Data decisions and theoretical implications when adversarially learning fair representations. arXiv preprint arXiv:1707.00075 (2017)
2. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
3. Dosovitskiy, A., Djolonga, J.: You only train once: Loss-conditional training of deep networks. In: International conference on learning representations (2019)
4. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29** (2016)
5. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in neural information processing systems **33**, 18661–18673 (2020)
6. Kleinberg, J., Mullainathan, S., Raghavan, M.: Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807 (2016)
7. Larrazabal, A.J., Nieto, N., Peterson, V., Milone, D.H., Ferrante, E.: Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proceedings of the National Academy of Sciences **117**(23), 12592–12594 (2020)
8. Luo, Y., Tian, Y., Shi, M., Pasquale, L.R., Shen, L.Q., Zebardast, N., Elze, T., Wang, M.: Harvard glaucoma fairness: a retinal nerve disease dataset for fairness learning and fair identity normalization. IEEE Transactions on Medical Imaging **43**(7), 2623–2633 (2024)
9. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM computing surveys (CSUR) **54**(6), 1–35 (2021)
10. Menon, A.K., Williamson, R.C.: The cost of fairness in binary classification. In: Conference on Fairness, accountability and transparency. pp. 107–118. PMLR (2018)
11. Park, S., Lee, J., Lee, P., Hwang, S., Kim, D., Byun, H.: Fair contrastive learning for facial attribute classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10389–10398 (2022)
12. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
13. Taufiq, M.F., Ton, J.F., Liu, Y.: Achievable fairness on your data with utility guarantees. Advances in Neural Information Processing Systems **37**, 140405–140450 (2024)
14. Özbulak, G., Jimenez-del Toro, O., Fatoretto, M., Berton, L., Anjos, A.: A multi-objective evaluation framework for analyzing utility-fairness trade-offs in machine learning systems. Machine Learning for Biomedical Imaging **3**(Special issue on FAIMI), 938–957 (2025). https://doi.org/10.59275/j.melba.2025-ab9a