

A FAIRNESS METRICS

In Table 2, we present the comprehensive list of fairness metrics taken from the literature along with their mathematical definitions and abbreviations. In our code-base, that we will release to the fairness community, all these metrics are provided and can be used for evaluation.

Fairness Criteria	Definition	Abbreviation
Group conditioned s-accuracy	$p(\hat{Y} = y Y = y, S = s)$	s-accuracy
s-True positive (Verma & Rubin (2018))	$ \{x \hat{g} = 1 \text{ for } (x, y = 1, S = s) \in X\} $ where $ \cdot $ refers to cardinality of a set	s-TP
s-False positive (Verma & Rubin (2018))	$ \{x \hat{g} = 1 \text{ for } (x, y = 0, S = s) \in X\} $ where $ \cdot $ refers to cardinality of a set	s-FP
s-False negative (Verma & Rubin (2018))	$ \{x \hat{g} = 0 \text{ for } (x, y = 1, S = s) \in X\} $ where $ \cdot $ refers to cardinality of a set	s-FN
s-True negative (Verma & Rubin (2018))	$ \{x \hat{g} = 0 \text{ for } (x, y = 0, S = s) \in X\} $ where $ \cdot $ refers to cardinality of a set	s-TN
s-True positive rate (Friedler et al. (2019)) = s-positive predictive value (s-PPV)	$p(\hat{Y} = 1 Y = 1, S = s)$	s-TPR
s-True negative rate	$p(\hat{Y} = 0 Y = 0, S = s)$	s-TNR
s-False positive rate	$p(\hat{Y} = 1 Y = 0, S = s)$ equivalent to $1 - \text{s-TNR}$	s-FPR
s-False negative rate	$p(\hat{Y} = 0 Y = 1, S = s)$ equivalent to $1 - \text{s-TPR}$	s-FNR
s-Balanced classification rate	$0.5 \times [p(\hat{Y} = 1 Y = 1, S = s) + p(\hat{Y} = 0 Y = 0, S = s)]$ equivalent to $0.5 \times (\text{s-TPR} + \text{s-TNR})$	s-BCR
Equality of odds (Hardt et al. (2016) & Beutel et al. (2017)) = Equalized odds (Hardt et al. (2016)) = conditional procedure accuracy equality (Berk et al. (2018)) = disparate mistreatment (Zafar et al. (2017b))	$p(\hat{Y} = \hat{y} Y = y) = p(\hat{Y} = \hat{y} Y = y, S = s)$ equivalent to $[p(\hat{Y} = 1 Y = 1, S = 1) = p(\hat{Y} = 1 Y = 1, S = 0) \text{ and } p(\hat{Y} = 1 Y = 0, S = 1) = p(\hat{Y} = 1 Y = 0, S = 0)]$ equivalent to $[1\text{-TPR} = 0\text{-TPR and } 0\text{-TNR} = 1\text{-TNR}]$	-
s-calibration+ (Friedler et al. (2019))	$p(Y = 1 \hat{Y} = 1, S = s)$	-
s-calibration- (Friedler et al. (2019))	$p(Y = 1 \hat{Y} = 0, S = s)$	-
Conditional use accuracy equality (Berk et al. (2018))	$[p(Y = 1 \hat{Y} = 1, S = 1) = p(Y = 1 \hat{Y} = 1, S = 0) \text{ and } p(Y = 0 \hat{Y} = 0, S = 1) = p(Y = 0 \hat{Y} = 0, S = 0)]$ equivalent to $[0\text{-calibration+} = 1\text{-calibration+ and } 0\text{-calibration-} = 1\text{-calibration-}]$	-
Calders and Verwer (Calders & Verwer (2010))	$1 - [p(\hat{Y} = 1 S = 1) - p(\hat{Y} = 1 S \neq 1)]$	CV
Demographic parity (Hardt et al. (2016) & Beutel et al. (2017)) = Group fairness (Dwork et al. (2012)) = statistical parity (Dwork et al. (2012)) = equal acceptance rate (Zhang et al. (2015))	$p(\hat{Y}) = p(\hat{Y} S)$ equivalent to 1-CV equivalent to $p(\hat{Y} = 1 S = 1) = p(\hat{Y} = 1 S \neq 1)$	DP
Disparate Impact (Feldman et al. (2015) & Zafar et al. (2017a))	$\frac{p(\hat{Y}=1 S \neq 1)}{p(\hat{Y}=1 S=1)}$	DI
Equality of opportunity with respect to y (Hardt et al. (2016))	$p(\hat{Y} = \hat{y} Y = y) = p(\hat{Y} = \hat{y} Y = y, S = s)$ Equality of odds is stronger than equality of opportunity	-
False positive error rate balance (Chouldechova (2017)) = predictive equality (Corbett-Davies et al. (2017))	$p(\hat{Y} = 1 Y = 0, S = 1) = p(\hat{Y} = 1 Y = 0, S = 0)$ equivalent to $p(\hat{Y} = 0 Y = 0, S = 1) = p(\hat{Y} = 0 Y = 0, S = 0)$ equivalent to $1\text{-TNR} = 0\text{-TNR}$ equivalent to $[\text{Equality of opportunity with respect to } y = 0]$	-
False negative error rate balance (Chouldechova (2017)) = equal opportunity (Kusner et al. (2017) & Hardt et al. (2016))	$p(\hat{Y} = 0 Y = 1, S = 1) = p(\hat{Y} = 0 Y = 1, S = 0)$ equivalent to $p(\hat{Y} = 1 Y = 1, S = 1) = p(\hat{Y} = 1 Y = 1, S = 0)$ equivalent to $1\text{-TPR} = 0\text{-TPR}$ equivalent to $[\text{Equality of opportunity with respect to } y = 1]$	-
Matthews correlation coefficient	$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$	MCC

Table 2: Fairness metrics. X, Y, S denote respectively the input sample, the ground truth label, and the sensitive attribute. p is the output probability of the model and \hat{Y} is the model’s prediction. For the metrics presented in this table, the sensitive attribute S takes binary values in $\{0, 1\}$.

B EXPERIMENTAL SETUP

Figure 5 shows samples of the dataset used in our experiments. We used 50000 images for the training set, 10000 images each for validation and test sets.

In all models, we used three fully connected layers for discriminator and classifier networks. In encoder network, we used three fully connected layers for all models except FFVAE, in which we used a convolutional encoder and decoder networks for increased training stability (Kim & Mnih, 2018). Leaky ReLU is used for all activation functions and Glorot (Glorot & Bengio, 2010) is used to initialize all weights. The models are trained using Adam optimizer with a learning rate of $1e-3$.



Figure 5: Sampled images from our dataset.

Models are trained for 500 epochs (with an early stopping of 5 epochs patience) to find the best model. The best model is separately found for each fairness metric based on the validation set.

Baseline MLP Setup: The baseline MLP model consists of an encoder for the input image and a classifier for eligibility prediction. Cross entropy loss is used for optimization.

LAFTR Setup: The model consists of an encoder, a classifier, and a discriminator. We used the codebase released by the authors of the original paper (Madras et al., 2018). Following the original code, we train the encoder and classifier together and train the discriminator in alternate steps. We used two discriminator iterations per encoder-classifier iteration and applied cross-entropy loss for optimization of both the classifier and discriminator. We used the default classification coefficient of 1.0 and used 5 values of adversarial coefficient $\gamma \in [0.1, 0.5, 1, 2, 3]$, as proposed in the original paper.

CFAIR Setup: The model consists of an encoder, a classifier, and two discriminators (one for each eligibility class label). We used the code provided by authors to run the experiments. We experimented with 5 values of adversarial coefficient $\gamma \in [0.1, 1, 10, 100, 1000]$, as proposed in the original paper. The binary loss (0-1 loss) in Eq.6 is NP-hard to optimize directly (Feldman et al., 2009; Ben-David et al., 2003), hence the model uses a convex relaxation of the binary loss, which is a weighted cross-entropy loss as shown below.

$$\mathcal{D}(\hat{Y} \neq y \mid Y = y) = \frac{\mathcal{D}(\hat{Y} \neq y, Y = y)}{\mathcal{D}(Y = y)} \leq \frac{\text{CE}_{\mathcal{D}_y}(\hat{Y} \| Y)}{\mathcal{D}(Y = y)} \quad (9)$$

FFVAE Setup: The model consists of a convolutional encoder, a convolutional decoder, a fully connected classifier, and a fully connected discriminator. We used the code provided by authors to run the experiments. We applied the adversarial coefficients $\gamma \in [0.1, 0.2, 0.4]$ and the alignment coefficients $\alpha \in [10, 100, 1000]$. We observed that the training of FFVAE becomes unstable for higher values of γ . This is due to the fact that the stability between *predictiveness* and *disentanglement* gets harder to achieve as they work against each other when the sensitive attribute and the eligibility are correlated. FFVAE model takes ELBO loss for the VAE and approximates the *disentanglement* term using the mean error difference between discriminator logits (Kim & Mnih, 2018). The model uses cross-entropy loss for the *predictiveness* term and the discriminator network.

In all the experiments the eligible and ineligible groups represent each 50% of the training data. We kept the widths of encoder, decoder, discriminator constant at 32, and the encoded latent representation size is 16 for all models. We experimented with two values of classifier widths 32, 64, and were unable to observe the trend which is recently emphasized by Sagawa et al. (2020) that increasing model capacities may lead to being unfair toward minorities while accuracy is getting better. However, this needs to be further investigated.

C EXPERIMENTS AND RESULTS

C.1 IMPACT OF REDUCING REPRESENTATION OF UNPRIVILEGED GROUP

We report the complete set of results for debiasing models of MLP, CFAIR, FFVAE, LAFTR-EqOdd, LAFTR-EqOpp1, LAFTR-EqOpp0, and LAFTR-DP, in Tables 4 to 10, corresponding to

Encoder	Classifier	Discriminator	FFVAE Encoder	FFVAE Decoder
Input $\in \mathbb{R}^{3072}$	Input $\in \mathbb{R}^{16}$	Input $\in \mathbb{R}^{16}$	Input: $32 \times 32 \times 3$ image	Input $\in \mathbb{R}^{16}$
FC. 32 LReLU	FC. 64/32 LReLU	FC. 32 LReLU	4×4 conv. 32 LReLU. stride 2	FC. 256 LReLU
FC. 32 LReLU	FC. 64/32 LReLU	FC. 32 LReLU	4×4 conv. 64 LReLU. stride 2	FC. $4 \times 4 \times 64$ LReLU
FC. 16 LReLU	FC. 2 LReLU	FC. 2 LReLU	4×4 conv. 64 LReLU. stride 2	4×4 upconv. 64 LReLU. stride 2
			FC. 256 LReLU	4×4 upconv. 32 LReLU. stride 2
			FC. 2×16	4×4 upconv. 3 LReLU. stride 2

Table 3: Architectures used for Baseline MLP, LAFTR, CFAIR, FFVAE models.

the experimental setup described in Section 5 of the main paper. Each pair in *clr-ratio* column indicate (b_e, b_o) , which is the ratio of images with blue background for (even=eligible, odd=ineligible) data. The pairs in all other columns show results respectively for (32, 64) eligibility classifier width. Figures 7 compare all models side-by-side.

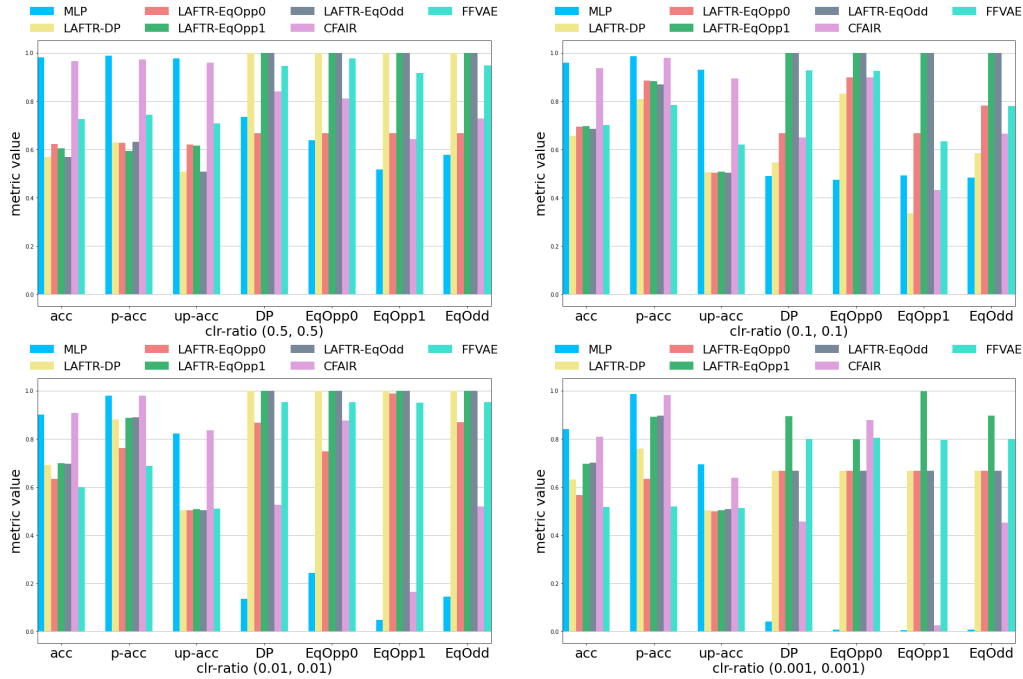


Figure 6: Comparing different models while decreasing minority representation.

<i>clr-ratio</i>	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd
(0.5, 0.5)	(0.97, 0.98)	(0.97, 0.99)	(0.97, 0.98)	(0.62, 0.73)	(0.34, 0.64)	(0.43, 0.52)	(0.39, 0.58)
(0.1, 0.1)	(0.96, 0.96)	(0.99, 0.99)	(0.92, 0.93)	(0.57, 0.49)	(0.4, 0.47)	(0.38, 0.49)	(0.39, 0.48)
(0.01, 0.01)	(0.91, 0.9)	(0.98, 0.98)	(0.83, 0.82)	(0.42, 0.13)	(0.4, 0.24)	(0.41, 0.05)	(0.41, 0.14)
(0.001, 0.001)	(0.86, 0.84)	(0.99, 0.99)	(0.74, 0.69)	(0.11, 0.04)	(0.19, 0.01)	(0.0, 0.0)	(0.1, 0.01)

Table 4: MLP results when decreasing minority representation, sensitive attribute: *bck*. The pairs in all other columns show results respectively for (32, 64) eligibility classifier width.

<i>clr-ratio</i>	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd
(0.5, 0.5)	(0.96, 0.96)	(0.97, 0.97)	(0.95, 0.96)	(0.82, 0.84)	(0.84, 0.81)	(0.76, 0.64)	(0.8, 0.73)
(0.1, 0.1)	(0.95, 0.94)	(0.98, 0.98)	(0.91, 0.89)	(0.68, 0.65)	(0.91, 0.9)	(0.45, 0.43)	(0.68, 0.67)
(0.01, 0.01)	(0.91, 0.91)	(0.98, 0.98)	(0.83, 0.84)	(0.49, 0.52)	(0.87, 0.88)	(0.35, 0.16)	(0.61, 0.52)
(0.001, 0.001)	(0.8, 0.81)	(0.98, 0.98)	(0.62, 0.64)	(0.48, 0.46)	(0.89, 0.88)	(0.03, 0.02)	(0.46, 0.45)

Table 5: CFAIR results when decreasing minority representation, sensitive attribute: *bck*, selected best result per attribute

<i>clr-ratio</i>	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd
(0.5, 0.5)	(0.72, 0.72)	(0.75, 0.74)	(0.7, 0.71)	(0.96, 0.95)	(0.98, 0.98)	(0.92, 0.92)	(0.95, 0.95)
(0.1, 0.1)	(0.7, 0.7)	(0.78, 0.78)	(0.61, 0.62)	(0.88, 0.93)	(0.92, 0.93)	(0.76, 0.63)	(0.84, 0.78)
(0.01, 0.01)	(0.6, 0.6)	(0.69, 0.69)	(0.51, 0.51)	(0.98, 0.95)	(0.88, 0.95)	(0.84, 0.95)	(0.86, 0.95)
(0.001, 0.001)	(0.52, 0.52)	(0.51, 0.52)	(0.52, 0.51)	(0.78, 0.8)	(0.78, 0.8)	(0.78, 0.8)	(0.78, 0.8)

Table 6: FFVAE results when decreasing minority representation, sensitive attribute:*bck*, selected best result per attribute

<i>clr-ratio</i>	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd
(0.5, 0.5)	(0.62, 0.57)	(0.63, 0.63)	(0.62, 0.51)	(1.0, 1.0)	(1.0, 1.0)	(1.0, 1.0)	(1.0, 1.0)
(0.1, 0.1)	(0.63, 0.69)	(0.76, 0.87)	(0.5, 0.5)	(0.71, 1.0)	(0.87, 1.0)	(0.68, 1.0)	(0.78, 1.0)
(0.01, 0.01)	(0.7, 0.7)	(0.89, 0.89)	(0.5, 0.5)	(0.77, 1.0)	(0.86, 1.0)	(0.67, 1.0)	(0.77, 1.0)
(0.001, 0.001)	(0.71, 0.71)	(0.89, 0.9)	(0.53, 0.51)	(0.73, 0.67)	(0.79, 0.67)	(0.67, 0.67)	(0.73, 0.67)

Table 7: LAFTR-EqOdd results when decreasing minority representation, sensitive attribute:*bck*, selected best result per attribute

<i>clr-ratio</i>	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd
(0.5, 0.5)	(0.62, 0.6)	(0.63, 0.59)	(0.61, 0.61)	(0.6, 1.0)	(0.62, 1.0)	(0.59, 1.0)	(0.6, 1.0)
(0.1, 0.1)	(0.7, 0.7)	(0.88, 0.88)	(0.51, 0.51)	(1.0, 1.0)	(1.0, 1.0)	(1.0, 1.0)	(1.0, 1.0)
(0.01, 0.01)	(0.7, 0.7)	(0.89, 0.89)	(0.5, 0.51)	(0.88, 1.0)	(0.76, 1.0)	(1.0, 1.0)	(0.88, 1.0)
(0.001, 0.001)	(0.7, 0.7)	(0.9, 0.89)	(0.5, 0.5)	(0.89, 0.89)	(0.94, 0.8)	(1.0, 1.0)	(0.97, 0.9)

Table 8: LAFTR-EqOpp1 results when decreasing minority representation, sensitive attribute:*bck*, selected best result per attribute

<i>clr-ratio</i>	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd
(0.5, 0.5)	(0.62, 0.62)	(0.63, 0.63)	(0.62, 0.62)	(0.67, 0.67)	(0.67, 0.67)	(0.67, 0.67)	(0.67, 0.67)
(0.1, 0.1)	(0.63, 0.69)	(0.75, 0.88)	(0.51, 0.5)	(0.5, 0.67)	(0.63, 0.9)	(0.37, 0.67)	(0.5, 0.79)
(0.01, 0.01)	(0.7, 0.63)	(0.89, 0.76)	(0.5, 0.5)	(0.89, 0.87)	(0.79, 0.75)	(1.0, 0.99)	(0.9, 0.87)
(0.001, 0.001)	(0.71, 0.56)	(0.9, 0.63)	(0.51, 0.5)	(0.69, 0.67)	(0.71, 0.67)	(0.67, 0.67)	(0.69, 0.67)

Table 9: LAFTR-EqOpp0 results when decreasing minority representation, sensitive attribute:*bck*, selected best result per attribute

<i>clr-ratio</i>	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd
(0.5, 0.5)	(0.62, 0.57)	(0.63, 0.63)	(0.62, 0.51)	(0.77, 1.0)	(0.87, 1.0)	(0.67, 1.0)	(0.77, 1.0)
(0.1, 0.1)	(0.63, 0.66)	(0.76, 0.81)	(0.5, 0.5)	(0.79, 0.55)	(0.89, 0.83)	(0.67, 0.33)	(0.78, 0.58)
(0.01, 0.01)	(0.7, 0.69)	(0.9, 0.88)	(0.5, 0.5)	(0.68, 1.0)	(0.69, 1.0)	(0.65, 1.0)	(0.67, 1.0)
(0.001, 0.001)	(0.7, 0.63)	(0.89, 0.76)	(0.5, 0.5)	(0.81, 0.67)	(0.93, 0.67)	(0.69, 0.67)	(0.81, 0.67)

Table 10: LAFTR-DP results when decreasing minority representation, sensitive attribute:*bck*, selected best result per attribute

C.2 IMPACT OF CORRELATION OF SENSITIVE ATTRIBUTE WITH ELIGIBILITY

We report the complete set of results for debiasing models of MLP, CFAIR, FFVAE, LAFTR-EqOdd, LAFTR-EqOpp1, LAFTR-EqOpp0, and LAFTR-DP, in Tables 11 to 17, corresponding to the experimental setup described in Section 5 of the main paper. Each pair in *clr-ratio* column indicate (b_e, b_o), which is the ratio of images with blue background for (even=qualified, odd=unqualified) data. The pairs in all other columns show results respectively for (32, 64) architectures. Figure 7 compare all models side-by-side.

C.3 IMPACT OF CORRELATION OF NON-SENSITIVE ATTRIBUTE WITH ELIGIBILITY

We report the complete set of results for debiasing models of MLP, CFAIR, FFVAE, LAFTR-EqOdd, LAFTR-EqOpp1, LAFTR-EqOpp0, and LAFTR-DP, in Tables 18 to 24, corresponding to the experimental setup described in Section 5 of the main paper. Each pair in *e-o-ratio* column indicate (*e-ratio*, *o-ratio*), which specifies the patterns on digits. The pairs in all other columns show results respectively for (32, 64) architectures. Figure 8 compare all models side-by-side.

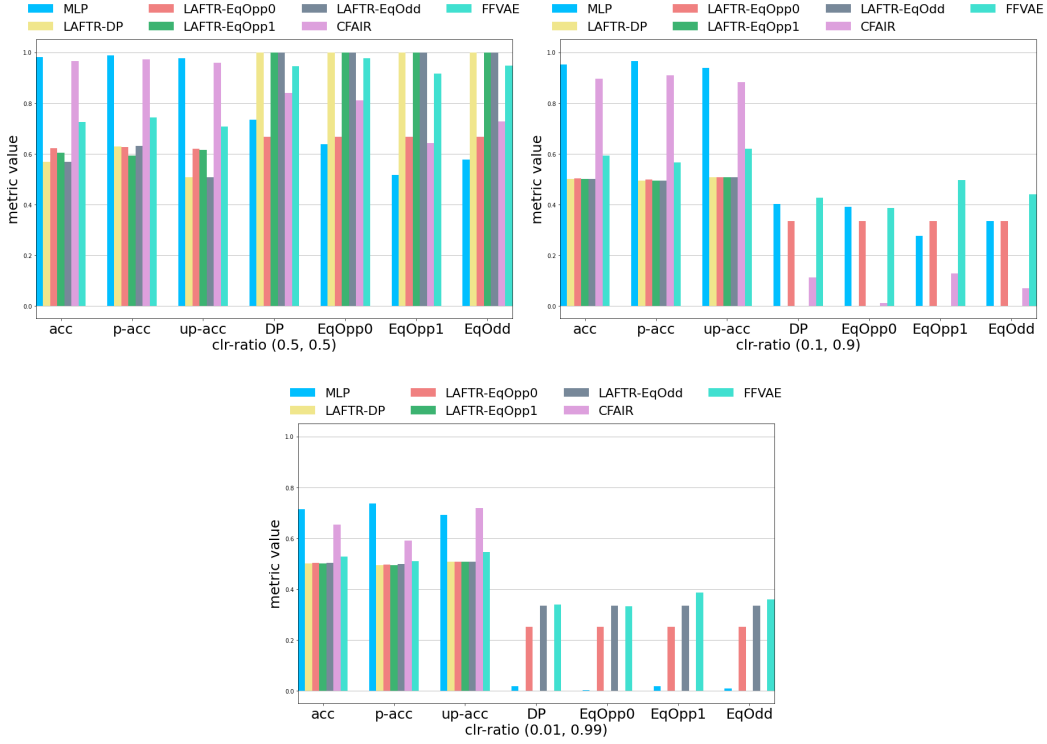


Figure 7: Comparing different models while shifting correlation of sensitive attribute (bck) and the eligibility.

$clr\text{-}ratio$	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd
(0.5, 0.5)	(0.97, 0.98)	(0.97, 0.99)	(0.97, 0.98)	(0.62, 0.73)	(0.34, 0.64)	(0.43, 0.52)	(0.39, 0.58)
(0.1, 0.9)	(0.94, 0.95)	(0.94, 0.96)	(0.93, 0.94)	(0.25, 0.4)	(0.16, 0.39)	(0.22, 0.28)	(0.19, 0.34)
(0.01, 0.99)	(0.74, 0.71)	(0.77, 0.74)	(0.72, 0.69)	(0.04, 0.02)	(0.0, 0.0)	(0.04, 0.02)	(0.02, 0.01)

Table 11: MLP results on correlation of sensitive attribute (bck) and eligibility

$clr\text{-}ratio$	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd
(0.5, 0.5)	(0.96, 0.96)	(0.97, 0.97)	(0.95, 0.96)	(0.82, 0.84)	(0.84, 0.81)	(0.76, 0.64)	(0.8, 0.73)
(0.1, 0.9)	(0.89, 0.9)	(0.9, 0.91)	(0.88, 0.88)	(0.24, 0.11)	(0.13, 0.01)	(0.11, 0.13)	(0.12, 0.07)
(0.01, 0.99)	(0.64, 0.66)	(0.58, 0.59)	(0.7, 0.72)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)

Table 12: CFAIR results on correlation of sensitive attribute (bck) and eligibility, selected best result per attribute

$clr\text{-}ratio$	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd
(0.5, 0.5)	(0.72, 0.72)	(0.75, 0.74)	(0.7, 0.71)	(0.96, 0.95)	(0.98, 0.98)	(0.92, 0.92)	(0.95, 0.95)
(0.1, 0.9)	(0.59, 0.59)	(0.56, 0.57)	(0.62, 0.62)	(0.45, 0.43)	(0.42, 0.39)	(0.44, 0.5)	(0.43, 0.45)
(0.01, 0.99)	(0.52, 0.53)	(0.5, 0.51)	(0.54, 0.55)	(0.35, 0.34)	(0.32, 0.33)	(0.37, 0.38)	(0.34, 0.35)

Table 13: FFVAE results on correlation of sensitive attribute (bck) and eligibility, selected best result per attribute

C.4 IMPACT OF SMALL FEATURES IN THE INPUT IMAGES

Comparing baseline model with debiasing models of MLP, CFAIR, FFVAE, LAFTR-EqOdd, LAFTR-EqOpp1, LAFTR-EqOpp0, and LAFTR-DP, when a small part of the image correlates with eligibility. Results are depicted in Figure 9 for a green visual component of only one pixel ($g_1 = 1$) and a five

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd
(0.5, 0.5)	(0.62, 0.57)	(0.63, 0.63)	(0.62, 0.51)	(1.0, 1.0)	(1.0, 1.0)	(1.0, 1.0)	(1.0, 1.0)
(0.1, 0.9)	(0.5, 0.5)	(0.49, 0.49)	(0.51, 0.51)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)
(0.01, 0.99)	(0.51, 0.51)	(0.5, 0.5)	(0.51, 0.51)	(0.33, 0.33)	(0.33, 0.33)	(0.33, 0.33)	(0.33, 0.33)

Table 14: LAFTR-EqOdd results on correlation of sensitive attribute (*bck*) and eligibility, selected best result per attribute

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd
(0.5, 0.5)	(0.62, 0.6)	(0.63, 0.59)	(0.61, 0.61)	(0.6, 1.0)	(0.62, 1.0)	(0.59, 1.0)	(0.6, 1.0)
(0.1, 0.9)	(0.51, 0.5)	(0.5, 0.49)	(0.51, 0.51)	(0.33, 0.0)	(0.33, 0.0)	(0.33, 0.0)	(0.33, 0.0)
(0.01, 0.99)	(0.5, 0.5)	(0.49, 0.49)	(0.51, 0.51)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)

Table 15: LAFTR-EqOpp1 results on correlation of sensitive attribute (*bck*) and eligibility, selected best result per attribute

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd
(0.5, 0.5)	(0.62, 0.62)	(0.63, 0.63)	(0.62, 0.62)	(0.67, 0.67)	(0.67, 0.67)	(0.67, 0.67)	(0.67, 0.67)
(0.1, 0.9)	(0.5, 0.51)	(0.49, 0.5)	(0.51, 0.51)	(0.0, 0.33)	(0.0, 0.33)	(0.0, 0.33)	(0.0, 0.33)
(0.01, 0.99)	(0.51, 0.51)	(0.5, 0.5)	(0.51, 0.51)	(0.33, 0.25)	(0.33, 0.25)	(0.33, 0.25)	(0.33, 0.25)

Table 16: LAFTR-EqOpp0 results on correlation of sensitive attribute (*bck*) and eligibility, selected best result per attribute

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd
(0.5, 0.5)	(0.62, 0.57)	(0.63, 0.63)	(0.62, 0.51)	(0.77, 1.0)	(0.87, 1.0)	(0.67, 1.0)	(0.77, 1.0)
(0.1, 0.9)	(0.5, 0.5)	(0.49, 0.49)	(0.51, 0.51)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)
(0.01, 0.99)	(0.5, 0.5)	(0.49, 0.49)	(0.51, 0.51)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)

Table 17: LAFTR-DP results on correlation of sensitive attribute (*bck*) and eligibility, selected best result per attribute

pixel width ($g_1 = 5$). The corresponding results are reported in Tables 25 to 31. The results presented in These tables are only for 64 architecture. See section 5 for experiment setups.

C.5 IMPACT OF SEED

In Figure 10 we illustrate the standard deviation of all models for all of the experiments described in Section 5 of the main paper for the 64 classifier architecture.

C.6 CORRELATION BETWEEN DATASET FEATURES AND MODEL’S PREDICTION.

We present Spearman Correlation plots for all of the experiments presented in Section 5 in Figures 11.

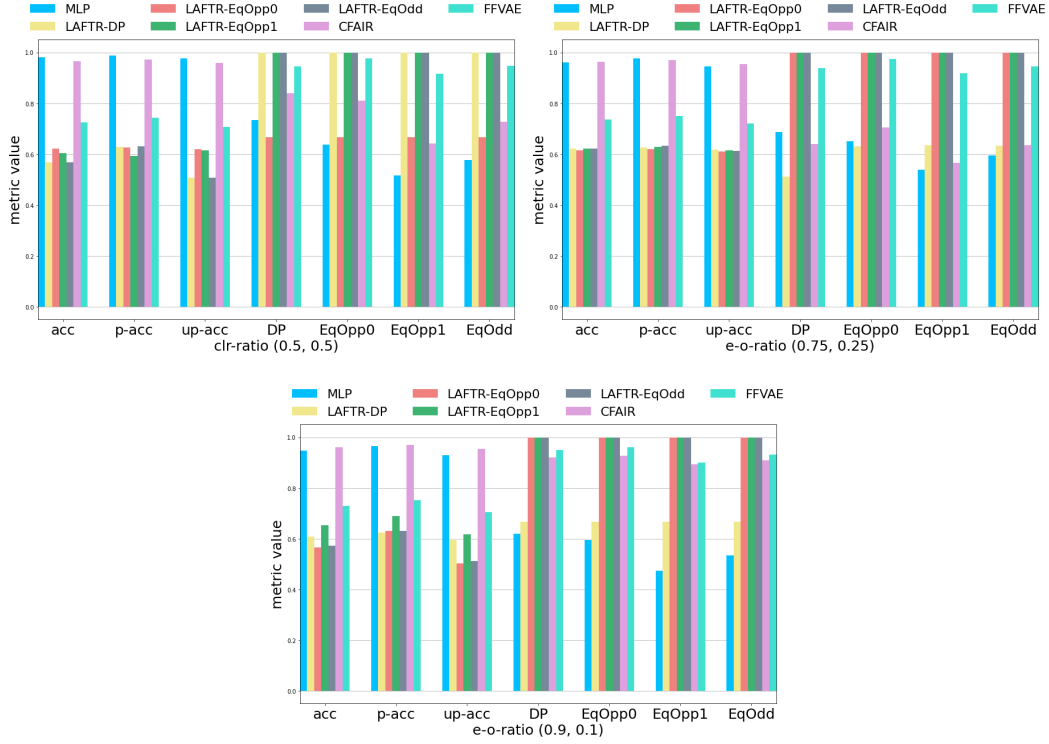


Figure 8: Comparing different models while shifting correlation of a non-sensitive attribute and the eligibility.

<i>e-o-ratio</i>	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd
(0.5, 0.5)	(0.97, 0.98)	(0.97, 0.99)	(0.97, 0.98)	(0.62, 0.73)	(0.34, 0.64)	(0.43, 0.52)	(0.39, 0.58)
(0.9, 0.1)	(0.96, 0.95)	(0.97, 0.96)	(0.96, 0.93)	(0.65, 0.62)	(0.49, 0.59)	(0.41, 0.47)	(0.45, 0.53)

Table 18: MLP results on correlation of non-sensitive attribute (digit lines) and eligibility

<i>e-o-ratio</i>	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd
(0.5, 0.5)	(0.96, 0.96)	(0.97, 0.97)	(0.96, 0.96)	(0.55, 0.9)	(0.59, 0.9)	(0.51, 0.87)	(0.55, 0.89)
(0.9, 0.1)	(0.96, 0.96)	(0.97, 0.97)	(0.95, 0.95)	(0.56, 0.92)	(0.78, 0.93)	(0.53, 0.89)	(0.66, 0.91)

Table 19: CFAIR results on correlation of non-sensitive attribute (digit lines) and eligibility, selected best result per attribute

<i>e-o-ratio</i>	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd
(0.5, 0.5)	(0.72, 0.72)	(0.74, 0.74)	(0.7, 0.71)	(0.95, 0.96)	(0.98, 0.98)	(0.92, 0.9)	(0.95, 0.94)
(0.9, 0.1)	(0.72, 0.73)	(0.75, 0.75)	(0.7, 0.71)	(0.94, 0.95)	(0.95, 0.96)	(0.9, 0.9)	(0.93, 0.93)

Table 20: FFVAE results on correlation of non-sensitive attribute (digit lines) and eligibility, selected best result per attribute

<i>e-o-ratio</i>	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd
(0.5, 0.5)	(0.59, 0.52)	(0.57, 0.52)	(0.62, 0.51)	(1.0, 1.0)	(1.0, 1.0)	(1.0, 1.0)	(1.0, 1.0)
(0.9, 0.1)	(0.67, 0.57)	(0.63, 0.63)	(0.7, 0.51)	(0.86, 1.0)	(0.98, 1.0)	(0.74, 1.0)	(0.86, 1.0)

Table 21: LAFTR-EqOdd results on correlation of non-sensitive attribute (digit lines) and eligibility, selected best result per attribute

<i>e-o-ratio</i>	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd
(0.5, 0.5)	(0.62, 0.61)	(0.63, 0.62)	(0.61, 0.61)	(0.77, 1.0)	(0.68, 1.0)	(0.86, 1.0)	(0.77, 1.0)
(0.9, 0.1)	(0.62, 0.66)	(0.63, 0.69)	(0.62, 0.62)	(0.77, 1.0)	(0.68, 1.0)	(0.85, 1.0)	(0.77, 1.0)

Table 22: LAFTR-EqOpp1 results on correlation of non-sensitive attribute (digit lines) and eligibility, selected best result per attribute

$e-o$ -ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd
(0.5, 0.5)	(0.56, 0.62)	(0.51, 0.63)	(0.61, 0.61)	(0.67, 0.68)	(0.67, 0.67)	(0.67, 0.69)	(0.67, 0.68)
(0.9, 0.1)	(0.69, 0.56)	(0.76, 0.63)	(0.61, 0.5)	(0.77, 1.0)	(0.68, 1.0)	(0.85, 1.0)	(0.77, 1.0)

Table 23: LAFTR-EqOpp0 results on correlation of non-sensitive attribute (digit lines) and eligibility, selected best result per attribute

$e-o$ -ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd
(0.5, 0.5)	(0.68, 0.56)	(0.63, 0.63)	(0.72, 0.5)	(0.8, 1.0)	(0.92, 1.0)	(0.68, 1.0)	(0.8, 1.0)
(0.9, 0.1)	(0.62, 0.6)	(0.63, 0.62)	(0.62, 0.59)	(0.67, 0.67)	(0.67, 0.67)	(0.67, 0.67)	(0.67, 0.67)

Table 24: LAFTR-DP results on correlation of non-sensitive attribute (digit lines) and eligibility, selected best result per attribute

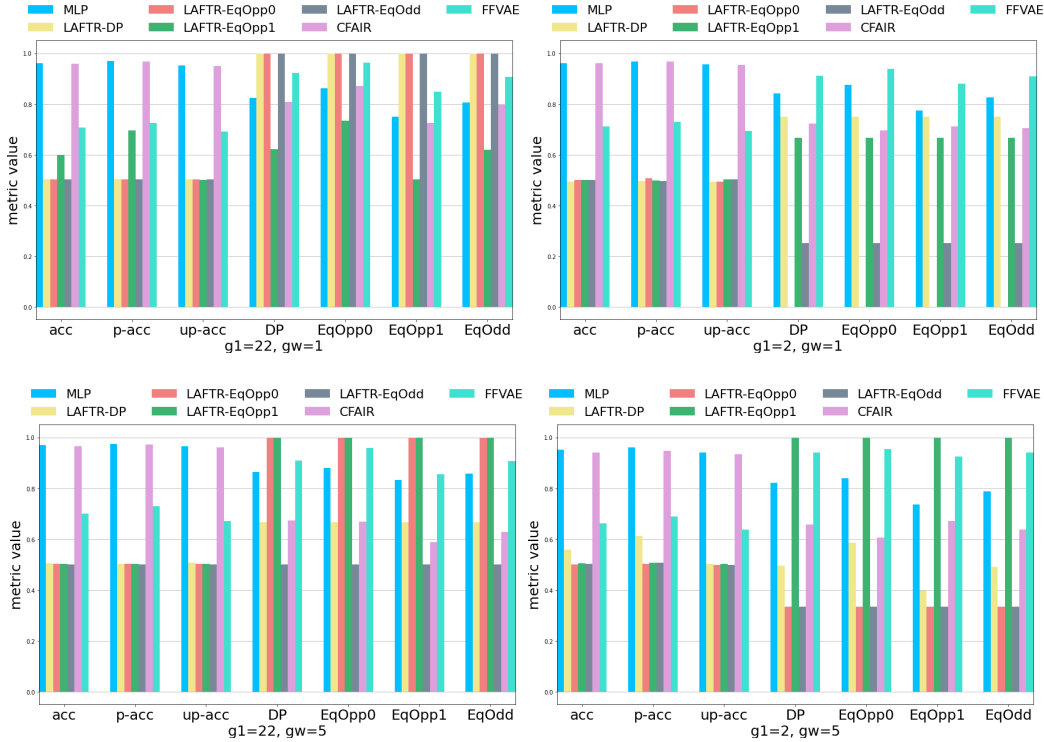


Figure 9: Impact of small visual components on different models' performance.

g_1, g_w	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd
(2, 1)	0.96	0.97	0.96	0.84	0.88	0.77	0.82
(22, 1)	0.96	0.97	0.95	0.82	0.86	0.75	0.8
(2, 5)	0.95	0.96	0.94	0.82	0.84	0.74	0.79
(22, 5)	0.96	0.97	0.96	0.87	0.88	0.83	0.85

Table 25: MLP results on impact of small visual components, sensitive attribute: *bck*

g_1, g_w	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd
(2, 1)	0.96	0.97	0.95	0.72	0.7	0.71	0.7
(22, 1)	0.96	0.97	0.95	0.81	0.87	0.72	0.79
(2, 5)	0.94	0.95	0.94	0.66	0.61	0.67	0.64
(22, 5)	0.96	0.97	0.96	0.67	0.67	0.59	0.63

Table 26: CFAIR results on impact of small visual components, sensitive attribute: *bck*, selected best result per attribute

g_1, g_w	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd
(2, 1)	0.71	0.73	0.69	0.91	0.94	0.88	0.91
(22, 1)	0.7	0.72	0.69	0.92	0.96	0.85	0.91
(2, 5)	0.67	0.69	0.64	0.94	0.95	0.93	0.94
(22, 5)	0.7	0.73	0.67	0.91	0.96	0.85	0.91

Table 27: FFVAE results on impact of small visual components, sensitive attribute:*bck*, selected best result per attribute

g_1, g_w	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd
(2, 1)	0.5	0.5	0.5	0.25	0.25	0.25	0.25
(22, 1)	0.5	0.5	0.5	1.0	1.0	1.0	1.0
(2, 5)	0.51	0.51	0.5	0.33	0.33	0.33	0.33
(22, 5)	0.5	0.5	0.5	0.5	0.5	0.5	0.5

Table 28: LAFTR-EqOdd sensitive attribute:*bck*, selected best result per attribute

g_1, g_w	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd
(2, 1)	0.5	0.5	0.5	0.67	0.67	0.67	0.67
(22, 1)	0.6	0.7	0.5	0.62	0.73	0.5	0.61
(2, 5)	0.51	0.51	0.5	1.0	1.0	1.0	1.0
(22, 5)	0.5	0.5	0.5	1.0	1.0	1.0	1.0

Table 29: LAFTR-EqOpp1 results on impact of small visual components, sensitive attribute:*bck*, selected best result per attribute

g_1, g_w	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd
(2, 1)	0.5	0.51	0.49	0.0	0.0	0.0	0.0
(22, 1)	0.5	0.5	0.5	1.0	1.0	1.0	1.0
(2, 5)	0.5	0.5	0.5	0.33	0.33	0.33	0.33
(22, 5)	0.5	0.5	0.5	1.0	1.0	1.0	1.0

Table 30: LAFTR-EqOpp0 results on impact of small visual components, sensitive attribute:*bck*, selected best result per attribute

g_1, g_w	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd
(2, 1)	0.49	0.5	0.49	0.75	0.75	0.75	0.75
(22, 1)	0.5	0.5	0.5	1.0	1.0	1.0	1.0
(2, 5)	0.55	0.61	0.5	0.5	0.58	0.4	0.49
(22, 5)	0.51	0.5	0.51	0.67	0.67	0.67	0.67

Table 31: LAFTR-DP results on impact of small visual components, sensitive attribute:*bck*, selected best result per attribute

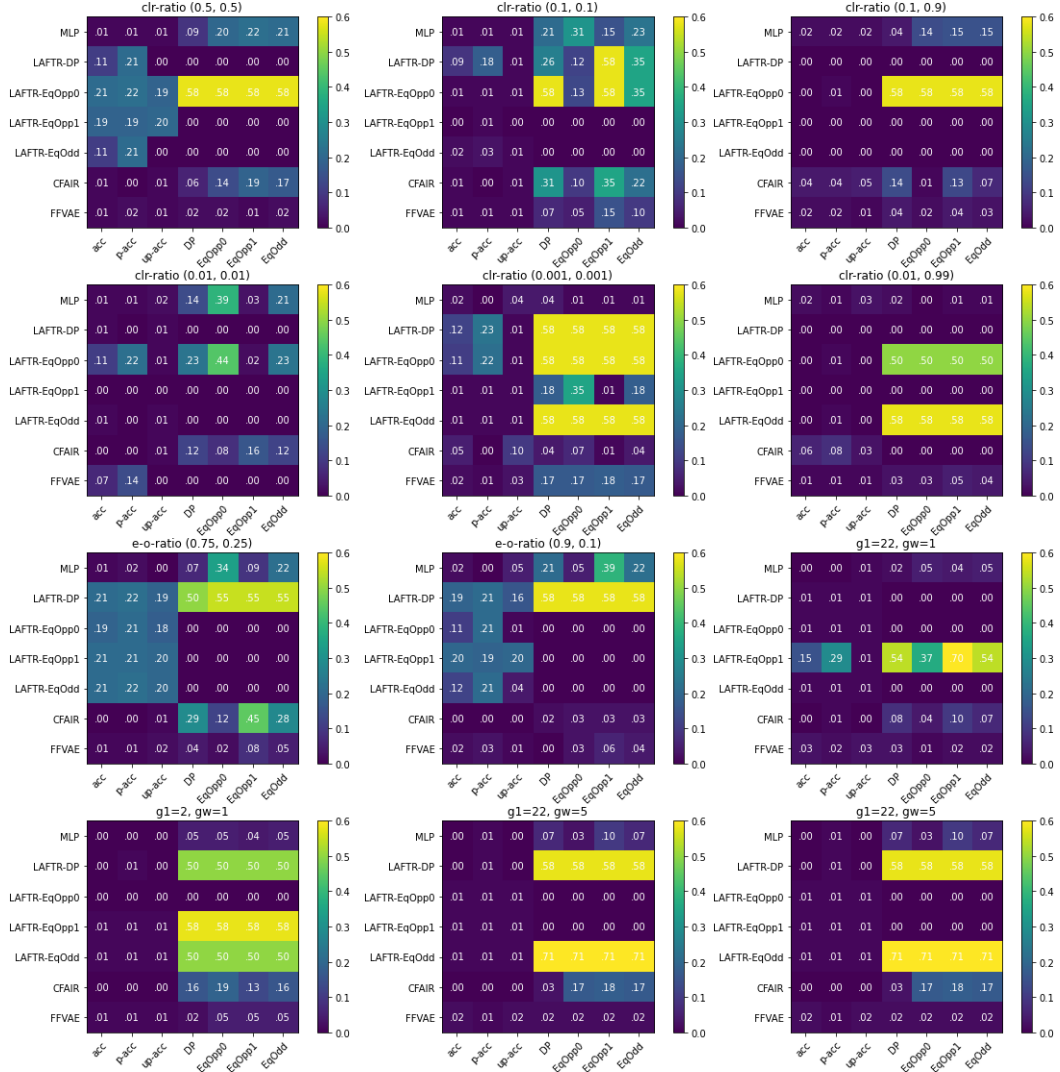


Figure 10: Standard deviation of different fairness metrics (x -axis) in different models (y -axis) for the 64 classifier architecture over three seeds. Each plot corresponds to a different experimental setup presented in Section 5.

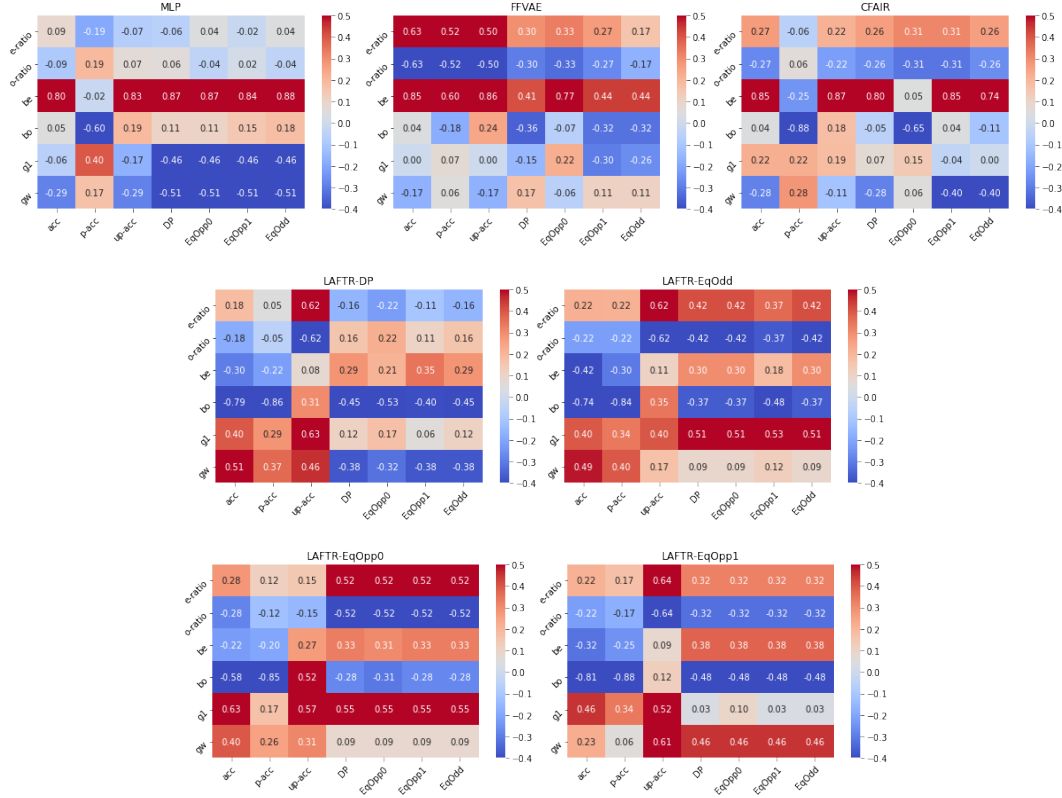


Figure 11: Spearman correlation plots for 64 classifier architecture. Each plot depicts correlation for one model and is measured over all of the experiments presented in Section 5. x -axis shows the fairness metric and the y -axis shows the dataset feature. Values can range in $[-1, 1]$ where the positive and negative indicating changes in the same or opposite directions and zero indicates no correlation. Almost all fairness models suffer from not mitigating the strong correlation between the overall accuracy (expected at the very least to be reasonably good) and the sensitive attribute background. These two variables are sometimes as strongly correlated as in the MLP model. Moreover, the absolute value of correlations between the blue and red background with accuracy in many models are not close enough which means the models act often biased.