A Paper Checklist

A.1 Potential negative societal impacts

While we report results using the best currently established fairness metrics, there is neither a universal fairness metric nor a universally accepted definition of fairness. Therefore, although we take steps towards proper benchmarking of bias mitigation algorithms, one should still be cautious about the chosen metrics before using them in real-world applications. Group parity metrics like DP, EqOpp0, EqOpp1, EqOdd can suffer from statistical limitations, as the true underlying protected group distributions might differ regardless of other unprotected features used for prediction, thus enforcing these metrics can lower accuracies and can harm the very groups designed to protect [47].

Hence before using any specific metric, the true data distribution should be studied well by designing suitable interventions. Real-world assessments and understanding consequences also help in achieving equitable metrics. Today with representation learning methods used in automatic decision-making applications [43, 44, 45, 46], more interpretable and explainable models are necessary to avoid any harmful consequences.

A.2 Limitations

The results presented in this work are obtained using binary sensitive attributes. It would be interesting to extend the present work to non-binary sensitive attributes in addition to observing the impact of bias-mitigation methods on multiple sensitive features. Due to the extensiveness of evaluating all these cases, we focused only on single binary attributes. Also, just mitigating the effect of sensitive features might not be enough, as there can be proxy features present, which are partially correlated with the sensitive features [47], it is aspiring to see fairness research that exploits the correlations between input features.

On a separate note, we have evaluated some of the recent promising bias-mitigation algorithms out of many proposed models. This field is expanding rapidly, and we could not evaluate all possible models. It would be interesting to evaluate other promising models in the proposed setups of this paper and try the proposed settings on more datasets to observe any potential change of performance due to difference in the modality of data.

A.3 Privacy and author consent

CI-MNIST : This data is an extension of the publicly available MNIST dataset [39], which does not contain any personal data. Yann LeCun and Corinna Cortes hold the copyright of MNIST dataset, which is a derivative work from original NIST datasets. MNIST dataset is made available under the terms of the Creative Commons Attribution-Share Alike 3.0 license (CC BY-SA 3.0).

Adult: The Adult dataset was originally extracted by Barry Becker from the 1994 Census bureau database and the data was first cited in [48]. It was donated by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics) and publicly released to the community on the UCI Data Repository [10]. It is licensed under Creative Commons Public Domain (CC0).

We do not put these datasets in our repository; instead, we provide code and guidelines on processing the original dataset to obtain the dataset variants used in our experiments.

A.4 Reproducibility

We have released our code at `https://github.com/charan223/FairDeepLearning`. We have provided the instructions to reproduce all experiments reported in the paper. Model details are provided in Section C. Dataset, architectural, and hyper-parameter details are presented in Section D. The detailed results on all experiments are presented in Section E. For each chart, we provide confidence-interval around the mean and run each experiment with three different seeds. We trained about 3000 models on 2 16GB NVIDIA Tesla V100 GPUs for 14 days.

B Fairness metrics

Table 1 presents all fairness metrics and their mathematical formulations that we use in our evaluations. 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. The provided tool by [36] is used to compute all of the metrics.

Table 1: X, Y, S denote the input, label, and the sensitive attribute. $\hat{Y}$ and p are the model's prediction and the output probability of the model. For all metrics, 1 indicates the perfect and 0 the lowest value.

Fairness Criteria	**Formulation**	**Short form**
Demographic Parity	$1 - \lvert p(\hat{Y}=1 \lvert S=\text{Protected}) - p(\hat{Y}=1 \lvert S=\text{Unprotected})\rvert$	DP
Equality of Opportunity (w.r.t $y=1$)	$1 - \lvert p(\hat{Y}=1 \lvert Y=1, S=\text{Unprotected}) - p(\hat{Y}=1 \lvert Y=1, S=\text{Protected})\rvert$	EqOpp1
Equality of Opportunity (w.r.t $y=0$)	$1 - \lvert p(\hat{Y}=1 \lvert Y=0, S=\text{Unprotected}) - p(\hat{Y}=1 \lvert Y=0, S=\text{Protected})\rvert$	EqOpp0
Equality of Odds	$0.5 \times [\text{EqOpp0} + \text{EqOpp1}]$	EqOdd
unprotected-accuracy	$p(\hat{Y}=y \lvert Y=y, S=\text{Unprotected})$	up-acc
protected-accuracy	$p(\hat{Y}=y \lvert Y=y, S=\text{Protected})$	p-acc
accuracy	$0.5 \times [\text{up-acc} + \text{p-acc}]$	acc

C Models

In this section, we describe in detail the models used in our evaluations.

Baseline Model. We use Mlp and Cnn as our baseline models, which given an input image x, predicts the probability p of the eligibility criteria. This probability is then transformed into a classification prediction $\hat{y}$. These models do not leverage any bias mitigation algorithm and are trained using a standard cross-entropy loss. It is meant to show how fair a baseline deep learning model would perform under different fairness criteria.

Learning Adversarially Fair and Transferable Representations (Laftr). Laftr [6] is an adversarial based bias mitigation algorithm within the scope of representation learning. In the supervised version of Laftr, given an input x, it first learns a latent encoded representation z that is passed to the discriminator to be debiased. The learned representation is then passed to a classifier to predict the task of interest y. The discriminator is trained by minimizing

$$\mathcal{L}_{fair}^{Laftr} = \mathbb{E}_{x,y,s\in\mathcal{D}}\mathcal{L}_{\mathcal{S}}(D(z,y),s) \tag{1}$$

where $\mathcal{L}_{\mathcal{S}}$ is the adversarial loss, and y is only passed in debiasing models aimed for *equality of odds* and *equality of opportunity* fairness metrics. The encoder and classifier are trained jointly by minimizing

$$\mathcal{L}_{Laftr} = \mathbb{E}_{x,y,s\in\mathcal{D}}\mathcal{L}_{\mathcal{Y}}(C(z),y) - \gamma\mathcal{L}_{fair}^{Laftr} \tag{2}$$

where $z = E(x)$ is the encoded feature, passed to both the classifier C and the discriminator D. The first term on the right side of the equation measures the classification loss (denoted as $\mathcal{L}_{cl}^{Laftr}$), and the second term (or the fairness objective) gets the adversarial gradients from the discriminator regarding the sensitive attribute s. Following the original paper, four variants of Laftr model are considered that represent the desired fairness criteria via $\mathcal{L}_{fair}^{Laftr}$. This includes:

(i) Laftr-DP in which the fairness objective is defined as

$$\mathcal{L}_{DP}^{Laftr} = 1 - \sum_{s\in\{0,1\}} \mathbb{E}_{x,s\in\mathcal{D}_s}\lvert D(z) - s\rvert \tag{3}$$

(ii) Laftr-EqOpp0 in which the fairness objective is considered as

$$\mathcal{L}_{EqOpp0}^{Laftr} = 1 - \sum_{s\in\{0,1\},y=0} \mathbb{E}_{x,s\in\mathcal{D}_s^y}\lvert D(z) - s\rvert \tag{4}$$

(iii) Laftr-EqOpp1 whose fairness objective $\mathcal{L}_{EqOpp1}^{Laftr}$ is obtained by replacing $y = 1$ in Eq. (4)

(iv) Laftr-EqOdd with the equality of odds fairness objective denoted as $\mathcal{L}_{EqOdd}^{Laftr}$ which is the sum of $\mathcal{L}_{EqOpp0}^{Laftr}$ and $\mathcal{L}_{EqOpp1}^{Laftr}$.

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 $\{0, 1\}$.

Fairness Criteria	**Definition**	**Abbreviation**
Group conditioned s-accuracy	$p(\hat{Y} = y \vert Y = y, S = s)$	s-accuracy
s-True positive ([37])	$\vert\{x \vert \hat{y} = 1 \text{ for } (x, y = 1, S = s) \in X\}\vert$ where $\vert \cdot \vert$ refers to cardinality of a set	s-TP
s-False positive ([37])	$\vert\{x \vert \hat{y} = 1 \text{ for } (x, y = 0, S = s) \in X\}\vert$ where $\vert \cdot \vert$ refers to cardinality of a set	s-FP
s-False negative ([37])	$\vert\{x \vert \hat{y} = 0 \text{ for } (x, y = 1, S = s) \in X\}\vert$ where $\vert \cdot \vert$ refers to cardinality of a set	s-FN
s-True negative ([37])	$\vert\{x \vert \hat{y} = 0 \text{ for } (x, y = 0, S = s) \in X\}\vert$ where $\vert \cdot \vert$ refers to cardinality of a set	s-TN
s-True positive rate ([36]) = s-positive predictive value (s-PPV)	$p(\hat{Y} = 1 \vert Y = 1, S = s)$	s-TPR
s-True negative rate	$p(\hat{Y} = 0 \vert Y = 0, S = s)$	s-TNR
s-False positive rate	$p(\hat{Y} = 1 \vert Y = 0, S = s)$ equivalent to $1 - s$-TNR	s-FPR
s-False negative rate	$p(\hat{Y} = 0 \vert Y = 1, S = s)$ equivalent to $1 - s$-TPR	s-FNR
s-Balanced classification rate	$0.5 \times [p(\hat{Y} = 1 \vert Y = 1, S = s) + p(\hat{Y} = 0 \vert Y = 0, S = s)]$ equivalent to $0.5\times$(s-TPR + s-TNR)	s-BCR
Equality of odds ([49] & [27]) = Equalized odds ([49])) = conditional procedure accuracy equality ([50]) = disparate mistreatment ([51])	$p(\hat{Y} = \hat{y} \vert Y = y) = p(\hat{Y} = \hat{y} \vert Y = y, S = s)$ equivalent to $\big[p(\hat{Y} = 1 \vert Y = 1, S = 1) = p(\hat{Y} = 1 \vert Y = 1, S = 0)$ and $p(\hat{Y} = 1 \vert Y = 0, S = 1) = p(\hat{Y} = 1 \vert Y = 0, S = 0)\big]$ equivalent to [1-TPR = 0-TPR and 0-TNR = 1-TNR]	-
s-calibration$+$ ([36])	$p(Y = 1 \vert \hat{Y} = 1, S = s)$	-
s-calibration$-$ ([36])	$p(Y = 1 \vert \hat{Y} = 0, S = s)$	-
Conditional use accuracy equality ([50])	$\big[p(Y = 1 \vert \hat{Y} = 1, S = 1) = p(Y = 1 \vert \hat{Y} = 1, S = 0)$ and $p(Y = 0 \vert \hat{Y} = 0, S = 1) = p(Y = 0 \vert \hat{Y} = 0, S = 0)\big]$ equivalent to [0-calibration$+$ = 1-calibration$+$ and 0-calibration$-$ = 1-calibration$-$]	-
Calders and Verwer ([52])	$1 - [p(\hat{Y} = 1 \vert S = 1) - p(\hat{Y} = 1 \vert S \neq 1)]$	CV
Demographic parity ([49] & [27]) = Group fairness ([17]) = statistical parity ([17]) = equal acceptance rate ([53])	$p(\hat{Y}) = p(\hat{Y} \vert S)$ equivalent to $1-$CV equivalent to $p(\hat{Y} = 1 \vert S = 1) = p(\hat{Y} = 1 \vert S \neq 1)$	DP
Disparate Impact ([54] & [40])	$\frac{p(\hat{Y}=1 \vert S\neq1)}{p(\hat{Y}=1 \vert S=1)}$	DI
Equality of opportunity with respect to y ([49])	$p(\hat{Y} = \hat{y} \vert Y = y) = p(\hat{Y} = \hat{y} \vert Y = y, S = s)$ Equality of odds is stronger than equality of opportunity	-
False positive error rate balance ([55]) = predictive equality([56])	$p(\hat{Y} = 1 \vert Y = 0, S = 1) = p(\hat{Y} = 1 \vert Y = 0, S = 0)$ equivalent to $p(\hat{Y} = 0 \vert Y = 0, S = 1) = p(\hat{Y} = 0 \vert Y = 0, S = 0)$ equivalent to 1-TNR = 0-TNR equivalent to [Equality of opportunity with respect to $y = 0$]	-
False negative error rate balance ([55]) = equal opportunity ([16] & [49])	$p(\hat{Y} = 0 \vert Y = 1, S = 1) = p(\hat{Y} = 0 \vert Y = 1, S = 0)$ equivalent to $p(\hat{Y} = 1 \vert Y = 1, S = 1) = p(\hat{Y} = 1 \vert Y = 1, S = 0)$ equivalent to 1-TPR = 0-TPR equivalent to [Equality of opportunity with respect to $y = 1$]	-
Matthews correlation coefficient	$\frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TP + FP})(\text{TP+FN})(\text{TN+FP})(\text{TN+FN})}}$	MCC

Conditional Learning of Fair Representations (Cfair). Proposed by [7], this model leverages two adversarial networks h_0 and h_1, predicting sensitive attribute s respectively for class labels $Y = 0$ and $Y = 1$. Cfair depends on an objective function called the balanced error rate (BER) [54, 57], which guarantees small joint error across demographic groups. The BER represents the sum of false positive rate and false negative rate. Therefore, it is equal to minimizing the below two conditional errors. $\text{BER}_{\mathcal{D}}(\hat{Y} \| Y)$ is defined as

$$\text{BER}_{\mathcal{D}}(\hat{Y} \| Y) \propto p(\hat{Y} = 1 | Y = 0) + p(\hat{Y} = 0 | Y = 1). \tag{5}$$

and $\text{BER}_{\mathcal{D}}(\hat{S} \| S)$ is defined similarly, where $\hat{S}$ is the predicted sensitive random variable. Cfair is optimized based on the following min-max formulation.

$$\mathcal{L}_{Cfair} = \min_{C,E} \max_{h_0,h_1} \left(\text{BER}_{\mathcal{D}}(C(E(X)) \| Y) - \gamma \mathcal{L}_{DP}^{Cfair} \right) \tag{6}$$

where

$$\mathcal{L}_{DP}^{Cfair} = \text{BER}_{\mathcal{D}^{y=0}} \left(h_0(E(X)) \| S \right) + \text{BER}_{\mathcal{D}^{y=1}} \left(h_1(E(X)) \| S \right) \tag{7}$$

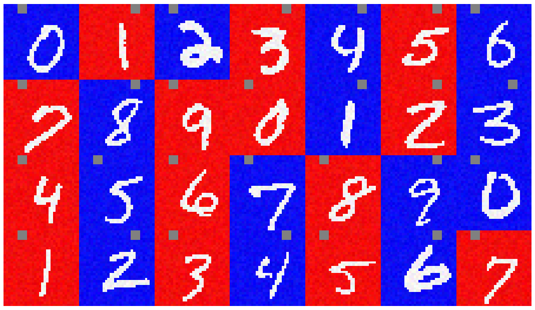


Figure 4: Sampled images from our dataset.

This approach proposes that using the balanced error rate along with the conditional alignment helps in achieving equalized odds across the groups without impacting demographic parity.

Cfair-EO is a variant of the Cfair model, which considers Cross-Entropy loss instead of BER loss for the classifier C to achieve equalized odds. In case of equal target class distribution, Cfair and Cfair-EO are the same, refer to Eq. (9) in the Appendix.

Flexibly Fair Representation Learning by Disentanglement (Ffvae). Inspired by FactorVAE [29], Ffvae [8] performs disentanglement by factorizing latent space. It learns a disentangled representation of the inputs, which is flexibly fair because it can be easily modified at test time to achieve demographic parity across various groups.

Given x, $s = (s_1, \ldots, s_N)$, $b = (b_1, \ldots, b_N)$, and z being respectively the input, the sensitive attribute, the sensitive latent, and non-sensitive latent, with N indicating the number of sensitive or non-sensitive features (depending on the dataset), Ffvae trains an encoder $q(z, b|x)$, a decoder $p(x|z, b)$, as well as an adversarial network. The latent representation is disentangled into sensitive b and non sensitive z latent attributes by encouraging both $\mathrm{MI}(b, z)$ and $\mathrm{MI}\left(b_i, s_j\right), \forall i \neq j$ to be low, where MI represents mutual information. Ffvae objective is defined as

$$\begin{aligned}\mathcal{L}_{\text{Ffvae}}(p,q) = \mathbb{E}_{q(z,b|x)}[\log p(x \mid z,b) + \alpha \log p(s \mid b)] - \gamma D_{KL}(q(z,b) \| q(z) \prod_j q\left(b_j\right)) \\ - D_{KL}(q(z,b \mid x) \| p(z,b))\end{aligned} \tag{8}$$

Eq.(8) has two terms, the first term consists of a reconstruction term (on left) and a *predictiveness* term $p(s \mid b)$, which aligns sensitive attributes to its respective sensitive latents, the second term is the *disentanglement* term which decorrelates the sensitive latent representation b from z using an adversarial network.

In addition to the models mentioned above, we also did experiments with [9] (based on the code released by authors), however, the model was very unstable on our dataset configurations even after extensive hyper-parameter search. We hypothesize that this is due to applying adversarial training directly to the class labels, which makes the model unstable, as indicated by the authors. Due to unstable results, we dropped this model from our evaluation.

D Experimental Setup

CI-MNIST dataset: Figure 4 shows samples of the dataset used in our experiments. Unless otherwise stated, we used 50000 images for the training set, 10000 images for each validation and test sets. In CI-MNIST experiments, the eligible and ineligible groups represent each 50% of the training data in both train and test sets. However, while the train set can be imbalanced with respect to sensitive attributes, the test set is always balanced. We initially pad the input image of size 28x28 on the top and sides to give a 32x32 image for the dataset creation. Blue and red colors with a 10% gaussian noise are used as background colors. For the small box, we used a 4x4 sized gray-colored box in the center of the top-left half or top-right half of the padded region of the image. We have experimented with multiple background colors, box colors and box sizes to understand the impact of colors, positions, sizes of the features on our models. Our motivation in choosing the current features is that the models can easily notice these features but find them difficult to remove. CI-MNIST currently supports multiple sensitive features (multiple background colors and positions of boxes). For more details refer to the released codebase.

Adult dataset: For training on the Adult dataset, we used the full Adult training set consisting of a total of 30,162 samples. We used 20% of the training set as the validation set. In the Adult dataset, the eligible and ineligible groups represent each 25%, 75% of the training data; hence the data is imbalanced with respect to target label with a skew towards ineligible lower-income class (<=50k). We use age as sensitive attribute and threshold it in the training sets (as indicated in Table 3) to create various *age-ratios*. Note that in all *age-ratios* the size of the training dataset does not change, and the eligible and ineligible groups remain at 25% and 75%. However, through thresholding of age as the sensitive attribute, we change the number of people in privileged and unprivileged groups. We find age-thresholds that closely resemble the dataset ratios used in CI-MNIST , in terms of the percentage of unprivileged individuals in eligible and ineligible groups.

We change the original test set to create a balanced version of it for testing, meaning it is balanced in terms of both sensitive attributes and target classes. Hence, the number of testing samples vary for each *age-ratio*. This is because we use age for the sensitive attribute, and when we change its threshold, the number of examples belonging to unprivileged and privileged group changes. We drop the minimum number of samples from the bigger subgroup of sensitive attribute and target class to make the dataset balanced. In Table 3, we mention the age thresholds used for the unprivileged group to achieve our desired *age-ratios* and also indicate the test-set size in each case. The remaining ages in either eligible and ineligible groups are considered privileged.

age-ratio	unprivileged age threshold	test set size
(0.5, 0.5)	25 <= age < 44	7336
(0.1, 0.1)	32 <= age < 36	1708
(0.01, 0.01)	71 <= age < 75	208
(0.66, 0.33)	38 <= age < 60	5480
(0.06, 0.36)	0 <= age < 30	908

Table 3: Thresholds of age and test set size used for various *age-ratios*. The test set is balanced in terms of both sensitive attribute and target class, while the train set is imbalanced and is of fixed size 30,162.

Architecture: In all models, we used three fully connected layers for discriminator and classifier networks. In the 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 [29]. Leaky ReLU is used for all activation functions, and Glorot [58] is used to initialize all weights. The models are trained using Adam optimizer with a learning rate of 1e-3. Models are trained for 500 epochs, with early stopping of 5 epochs patience on the validation set's loss to find the best model.

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.

Baseline Cnn Setup: The baseline Cnn model consists of an Cnn encoder for the input image and an Mlp 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 an adapted PyTorch version of the original codebase released by the authors of the original paper [6]. Following the original code's training method, we train the encoder, classifier 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 five values of adversarial coefficient $\gamma \in [0.1, 0.5, 1, 2, 4]$, 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 the authors to run the experiments. We experimented with five 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 [59, 60], hence the model uses a convex relaxation of the binary loss, which is a weighted cross-entropy loss as shown below.

$$\begin{aligned}\mathcal{D}(\widehat{Y} \neq y \mid Y = y) =& \frac{\mathcal{D}(\widehat{Y} \neq y, Y = y)}{\mathcal{D}(Y = y)} \\ \leq& \frac{\mathrm{CE}_{\mathcal{D}^y}(\widehat{Y} \| Y)}{\mathcal{D}(Y = y)}\end{aligned} \tag{9}$$

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 adversarial coefficient $\gamma \in [10, 50, 100]$ and the alignment coefficient $\alpha \in [10, 100, 1000]$. 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 [29]. The model uses cross-entropy loss for the *predictiveness* term and the discriminator network.

In CI-MNIST experiments, 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 [61] that increasing model capacities may lead to being unfair toward minorities while accuracy is getting better. However, this needs to be further investigated. Tables 4, 5 show the architectural details for CI-MNIST and Adult experiments.

Table 4: Architectures used for Baseline Mlp, Baseline Cnn, Laftr, Cfair, Ffvae models for CI-MNIST dataset.

Mlp Encoder	Mlp Classifier/Discriminator	Cnn Encoder
Input $\in \mathbb{R}^{3072}$	Input $\in \mathbb{R}^{16}$	Input $32 \times 32 \times 3$ image
FC. 32 LReLU	FC. 32 LReLU	4×4 conv. 32 LReLU. stride 2, padding 1
FC. 32 LReLU	FC. 32 LReLU	4×4 conv. 64 LReLU. stride 2, padding 1
FC. 16 LReLU	FC. 2 LReLU	4×4 conv. 64 LReLU. stride 2, padding 1
		4×4 conv. 256 LReLU. stride 1
		1×1 conv. 16 LReLU.

Ffvae Encoder	Ffvae Decoder
Input: $32 \times 32 \times 3$ image	Input $\in \mathbb{R}^{16}$
4×4 conv. 32 LReLU. stride 2, padding 1	FC. 128 LReLU
4×4 conv. 64 LReLU. stride 2, padding 1	FC. 1024 LReLU, Resize $64 \times 4 \times 4$
4×4 conv. 64 LReLU. stride 2, padding 1	4×4 upconv. 64 LReLU. stride 2, padding 1
Flatten 1024, FC. 128 LRELU	4×4 upconv. 32 LReLU. stride 2, padding 1
FC. 2×16	4×4 upconv. 3 LReLU. stride 2, padding 1

In Adult experiments, we kept the widths, latent representation sizes the same as their respective original papers.

Table 5: Architectures used for Baseline Mlp, Laftr, Cfair, Ffvae models for Adult dataset.

Mlp Encoder	Mlp Classifier	Ffvae Encoder	Ffvae Classifier/ Discriminator
Input $\in \mathbb{R}^{112}$	Input $\in \mathbb{R}^{16}$	Input $\in \mathbb{R}^{112}$	Input $\in \mathbb{R}^{60}$
FC. 32 LReLU	FC. 32 LReLU	FC. 200 LReLU	FC. 200 LReLU
FC. 32 LReLU	FC. 32 LReLU	FC. 60 LReLU	FC. 2 LReLU
FC. 16 LReLU	FC. 2 LReLU		

Laftr Encoder	Laftr Classifier/Discriminator	Cfair Encoder	Cfair Classifier/Discriminator
Input $\in \mathbb{R}^{112}$	Input $\in \mathbb{R}^{8}$	Input $\in \mathbb{R}^{112}$	Input $\in \mathbb{R}^{60}$
FC. 8 LReLU	FC. 2 LReLU	FC. 60 LReLU	FC. 0/50 LReLU
			FC. 2 LReLU

For sensitive information removal experiments in Section 5.1, sensitive features are predicted from latent representations from model-specific encoders. We use the same architecture as Mlp Classifier in Table 5 for these experiments.

Hyperparameter details:

Laftr. We use adversarial coefficient $\gamma \in [0.1, 0.5, 1, 2, 4]$ as hyperparameter as proposed in the original paper and we use two discriminator iterations per encoder-classifier iteration.

Cfair. We use adversarial coefficient $\gamma \in [0.1, 1, 10, 100, 1000]$ as hyperparameter as proposed in the original paper.

Ffvae. We use adversarial coefficient $\gamma \in [10, 50, 100]$ and the alignment coefficient $\alpha \in [10, 100, 1000]$ as hyperparameters as proposed in the original paper. We also use patience epochs 5 for early stopping in VAE training as a hyperparameter.

Other general hyperparameters considered for all the models include classifier, encoder, and discriminator widths, number of layers, and latent representation size with values mentioned in Tables 4 and 5. We take 5 epochs as stopping patience, use Adam as an optimizer with a learning rate of 1e-3. Please check our repository for a complete set of hyper-parameters and training setups.

E Experiments and Results

E.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 6 to 21, corresponding to the experimental setup described in Setting 1 of Section 4 in 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. Figures 5, 6 compare all models side-by-side. Note that to report results, we initially averaged metrics over three seeds, then for each metric, the best value over the fairness coefficients of the models is reported on the test set.

Table 6: Mlp results when decreasing minority representation for Adult dataset, sensitive attribute:age, selected best result per attribute

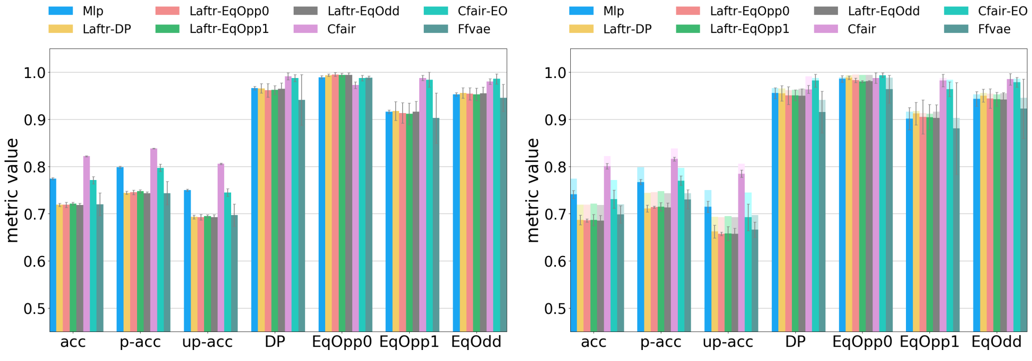
(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.78	0.8	0.75	0.97	0.99	0.92	0.96	0.66
(0.1, 0.1)	0.74	0.77	0.71	0.96	0.99	0.9	0.95	0.51
(0.01, 0.01)	0.74	0.81	0.66	0.91	0.97	0.77	0.87	0.54

Table 7: Cfair results when decreasing minority representation for Adult dataset, sensitive attribute:age, selected best result per attribute

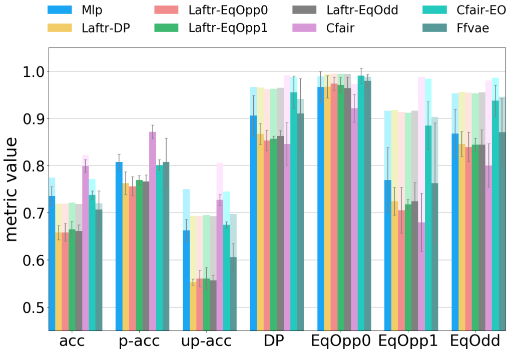
(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.82	0.84	0.81	0.99	0.97	0.99	0.98	0.54
(0.1, 0.1)	0.8	0.82	0.79	0.96	0.99	0.98	0.98	0.51
(0.01, 0.01)	0.8	0.87	0.73	0.85	0.92	0.68	0.8	0.54

Table 8: Cfair-EO results when decreasing minority representation for Adult dataset, sensitive attribute:age, selected best result per attribute

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.78	0.8	0.75	0.99	0.99	0.98	0.98	0.52
(0.1, 0.1)	0.73	0.77	0.69	0.98	0.99	0.96	0.97	0.51
(0.01, 0.01)	0.74	0.8	0.67	0.96	0.99	0.88	0.94	0.54



(a) *age-ratio* $(0.5, 0.5)$ (b) *age-ratio* $(0.1, 0.1)$



(c) *age-ratio* $(0.01, 0.01)$

Figure 5: Comparing different models while decreasing minority representation for Adult dataset. In sub-figures 5(b) and 5(c) the pale colors show the decrease in performance compared to the balanced case in 5(a).

Table 9: Ffvae results when decreasing minority representation for Adult dataset, sensitive attribute:age, selected best result per attribute

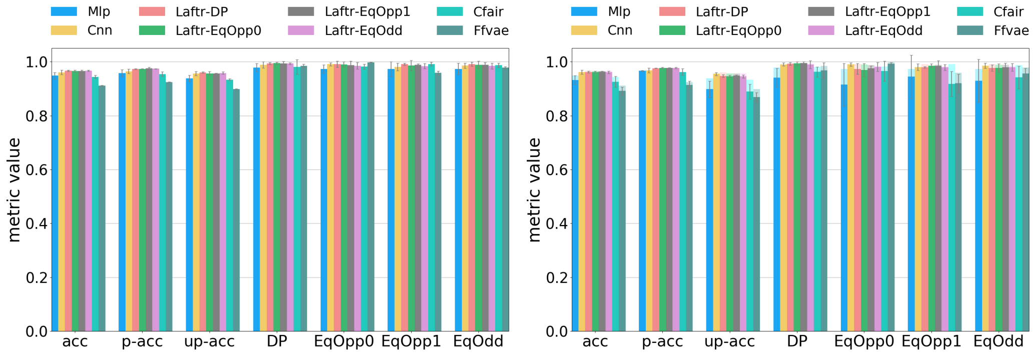
(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.72	0.74	0.7	0.94	0.99	0.9	0.95	0.93
(0.1, 0.1)	0.7	0.73	0.67	0.92	0.96	0.88	0.92	0.98
(0.01, 0.01)	0.71	0.81	0.61	0.91	0.98	0.76	0.87	0.92

Table 10: Laftr-EqOdd results when decreasing minority representation for Adult dataset, sensitive attribute:age, selected best result per attribute

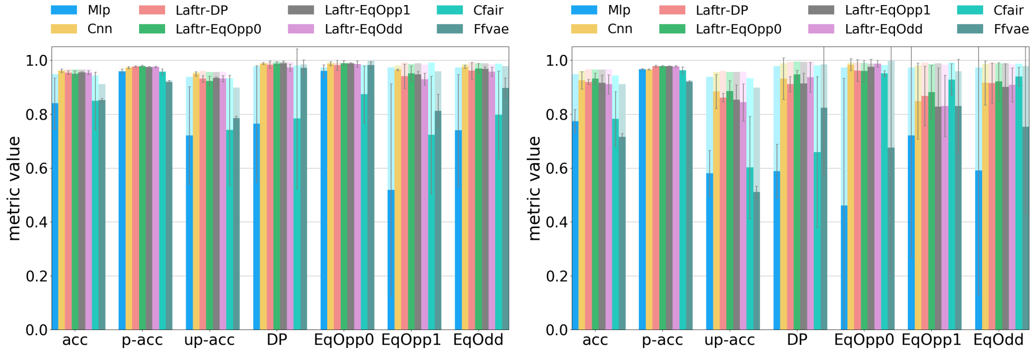
(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.71	0.74	0.69	0.96	0.99	0.92	0.96	0.52
(0.1, 0.1)	0.69	0.71	0.66	0.95	0.98	0.9	0.94	0.52
(0.01, 0.01)	0.67	0.77	0.56	0.86	0.96	0.72	0.84	0.52

Table 11: Laftr-EqOpp1 results when decreasing minority representation for Adult dataset, sensitive attribute:age, selected best result per attribute

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.72	0.75	0.7	0.96	0.99	0.91	0.95	0.53
(0.1, 0.1)	0.69	0.72	0.66	0.95	0.98	0.9	0.94	0.52
(0.01, 0.01)	0.67	0.77	0.56	0.86	0.97	0.72	0.84	0.52



(a) *clr-ratio* $(0.5, 0.5)$ (b) *clr-ratio* $(0.1, 0.1)$



(c) *clr-ratio* $(0.01, 0.01)$ (d) *clr-ratio* $(0.001, 0.001)$

Figure 6: Comparing different models while decreasing minority representation for CI-MNIST dataset. In sub-figures 6(b), 6(c), and 6(d) the pale colors show the decrease in performance compared to the balanced case in 6(a).

Table 12: Laftr-EqOpp0 results when decreasing minority representation for Adult dataset, sensitive attribute:age, selected best result per attribute

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.72	0.75	0.69	0.96	1.0	0.91	0.96	0.52
(0.1, 0.1)	0.69	0.71	0.66	0.95	0.98	0.91	0.95	0.52
(0.01, 0.01)	0.66	0.76	0.56	0.85	0.97	0.71	0.84	0.52

Table 13: Laftr-DP results when decreasing minority representation for Adult dataset, sensitive attribute:age, selected best result per attribute

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.71	0.74	0.69	0.97	0.99	0.92	0.96	0.60
(0.1, 0.1)	0.69	0.71	0.66	0.96	0.99	0.91	0.95	0.52
(0.01, 0.01)	0.66	0.76	0.55	0.87	0.97	0.72	0.84	0.55

Table 14: Mlp results when decreasing minority representation for CI-MNIST dataset, sensitive attribute:bck, selected best result per attribute

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.95	0.96	0.94	0.98	0.97	0.97	0.97	0.91
(0.1, 0.1)	0.94	0.97	0.9	0.94	0.91	0.94	0.93	0.99
(0.01, 0.01)	0.84	0.96	0.72	0.76	0.96	0.52	0.74	0.9
(0.001, 0.001)	0.77	0.97	0.58	0.59	0.46	0.72	0.59	0.67

Table 15: Cnn results when decreasing minority representation for CI-MNIST dataset, sensitive attribute:bck, selected best result per attribute

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.96	0.96	0.96	0.99	0.99	0.98	0.98	0.56
(0.1, 0.1)	0.96	0.97	0.95	0.99	0.99	0.98	0.98	0.5
(0.01, 0.01)	0.96	0.97	0.95	0.99	0.99	0.96	0.97	0.5
(0.001, 0.001)	0.93	0.97	0.88	0.93	0.98	0.85	0.92	0.5

Table 16: Cfair results when decreasing minority representation for CI-MNIST dataset, sensitive attribute:bck, selected best result per attribute

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.94	0.95	0.93	0.98	0.98	0.99	0.98	1.0
(0.1, 0.1)	0.93	0.96	0.89	0.96	0.97	0.92	0.95	1.0
(0.01, 0.01)	0.85	0.96	0.74	0.78	0.87	0.72	0.79	1.0
(0.001, 0.001)	0.78	0.96	0.6	0.66	0.95	0.93	0.94	1.0

Table 17: Ffvae results when decreasing minority representation for CI-MNIST dataset, sensitive attribute:bck, selected best result per attribute

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.91	0.92	0.9	0.98	1.0	0.96	0.98	1.0
(0.1, 0.1)	0.89	0.91	0.87	0.97	0.99	0.92	0.96	1.0
(0.01, 0.01)	0.85	0.92	0.79	0.97	0.98	0.81	0.9	1.0
(0.001, 0.001)	0.72	0.92	0.51	0.82	0.68	0.83	0.76	1.0

Table 18: Laftr-EqOdd results when decreasing minority representation for CI-MNIST dataset, sensitive attribute:bck, selected best result per attribute

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.96	0.97	0.96	0.99	0.99	0.98	0.98	1.0
(0.1, 0.1)	0.96	0.98	0.95	0.99	0.98	0.98	0.98	0.96
(0.01, 0.01)	0.96	0.98	0.93	0.97	0.99	0.93	0.96	0.8
(0.001, 0.001)	0.91	0.98	0.84	0.94	0.99	0.83	0.91	0.8

Table 19: Laftr-EqOpp1 results when decreasing minority representation for CI-MNIST dataset, sensitive attribute:bck, selected best result per attribute

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.97	0.98	0.96	0.99	0.99	0.99	0.99	0.99
(0.1, 0.1)	0.96	0.98	0.95	0.99	0.98	0.98	0.98	0.95
(0.01, 0.01)	0.95	0.97	0.93	0.99	0.99	0.95	0.97	0.85
(0.001, 0.001)	0.92	0.98	0.85	0.91	0.98	0.83	0.91	0.73

Table 20: Laftr-EqOpp0 results when decreasing minority representation for CI-MNIST dataset, sensitive attribute:bck, selected best result per attribute

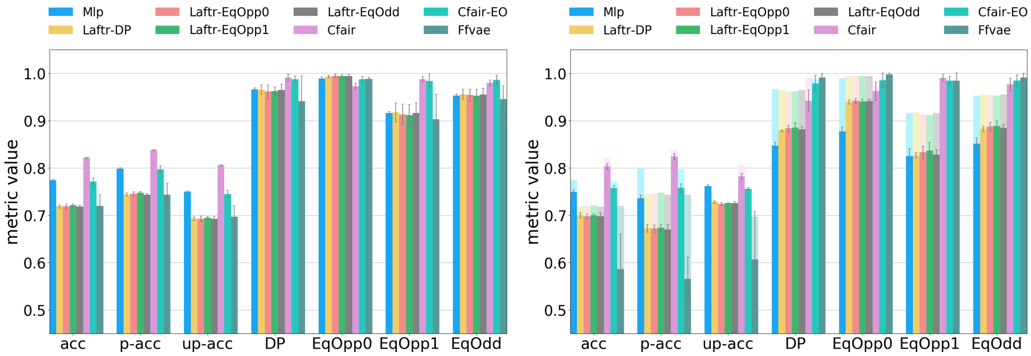
clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.96	0.97	0.96	0.99	0.99	0.99	0.99	0.99
(0.1, 0.1)	0.96	0.98	0.95	0.99	0.97	0.99	0.98	0.98
(0.01, 0.01)	0.95	0.98	0.92	0.99	0.99	0.95	0.97	0.78
(0.001, 0.001)	0.94	0.98	0.89	0.95	0.96	0.88	0.92	0.67

Table 21: Laftr-DP results when decreasing minority representation for CI-MNIST dataset, sensitive attribute:bck, selected best result per attribute

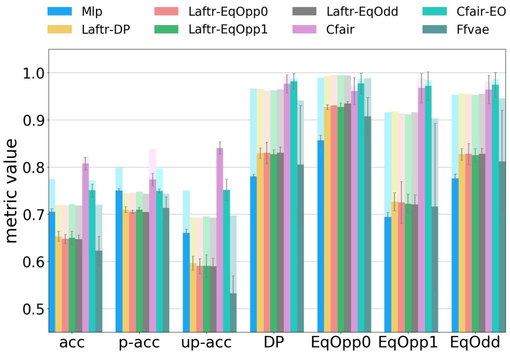
clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.96	0.97	0.96	0.99	0.99	0.99	0.99	1.0
(0.1, 0.1)	0.96	0.98	0.95	0.99	0.97	0.98	0.97	1.0
(0.01, 0.01)	0.96	0.98	0.93	0.98	0.98	0.94	0.96	0.86
(0.001, 0.001)	0.92	0.98	0.86	0.91	0.96	0.87	0.92	0.69

E.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 22 to 37, corresponding to Setting 2 in Section 4 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. Figures 7, 8 compare all models side-by-side.



(a) *age-ratio* $(0.5, 0.5)$ (b) *age-ratio* $(0.66, 0.33)$

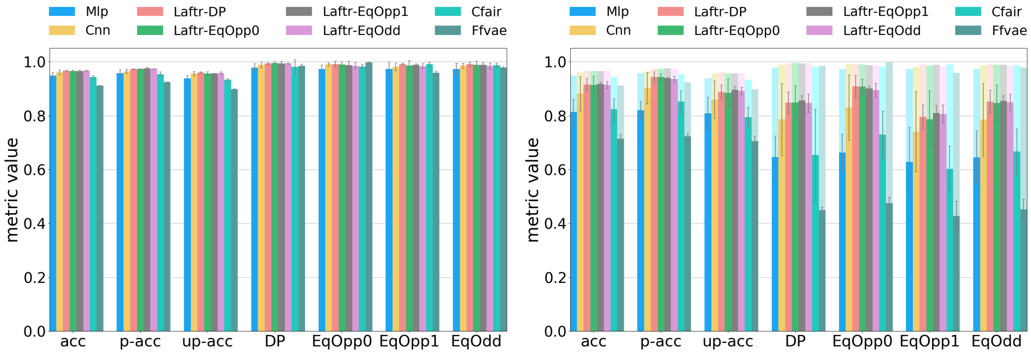


(c) *age-ratio* $(0.06, 0.36)$

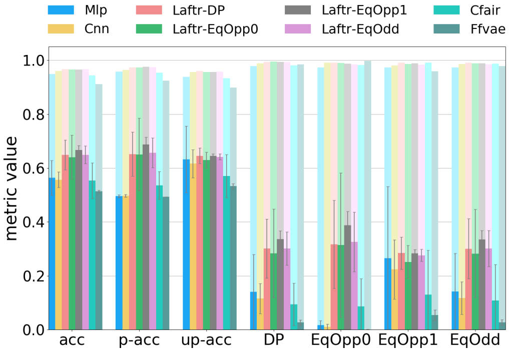
Figure 7: Comparing different models while shifting correlation of sensitive attribute (age) with the eligibility for Adult dataset. In sub-figures 7(b) and 7(c) the pale colors show the decrease in performance compared to the balanced case in 7(a).

Table 22: Mlp results on correlation of sensitive attribute (age) and eligibility for Adult dataset, selected best result per attribute

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.78	0.8	0.75	0.97	0.99	0.92	0.96	0.66
(0.66, 0.33)	0.75	0.74	0.76	0.85	0.88	0.83	0.85	0.65
(0.06, 0.36)	0.71	0.75	0.66	0.78	0.86	0.69	0.77	0.65



(a) *clr-ratio* $(0.5, 0.5)$ (b) *clr-ratio* $(0.1, 0.9)$



(c) *clr-ratio* $(0.01, 0.99)$

Figure 8: Comparing different models while shifting correlation of sensitive attribute (bck) and the eligibility for CI-MNIST dataset. In sub-figures 8(b) and 8(c) the pale colors show the decrease in performance compared to the balanced case in 8(a).

Table 23: Cfair results on correlation of sensitive attribute (age) and eligibility for Adult dataset, selected best result per attribute

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.82	0.84	0.81	0.99	0.97	0.99	0.98	0.54
(0.66, 0.33)	0.8	0.82	0.78	0.94	0.96	0.99	0.97	0.64
(0.06, 0.36)	0.8	0.77	0.84	0.98	0.96	0.97	0.96	0.54

Table 24: Cfair-EO results on correlation of sensitive attribute (age) and eligibility for Adult dataset, selected best result per attribute

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.78	0.8	0.75	0.99	0.99	0.98	0.98	0.52
(0.66, 0.33)	0.76	0.76	0.76	0.98	0.99	0.98	0.98	0.54
(0.06, 0.36)	0.75	0.75	0.75	0.98	0.98	0.97	0.97	0.53

Table 25: Ffvae results on correlation of sensitive attribute (age) and eligibility for Adult dataset, selected best result per attribute

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.72	0.74	0.7	0.94	0.99	0.9	0.95	0.93
(0.66, 0.33)	0.59	0.57	0.61	0.99	1.0	0.98	0.99	0.92
(0.06, 0.36)	0.62	0.71	0.53	0.81	0.91	0.72	0.81	0.98

Table 26: Laftr-EqOdd results on correlation of sensitive attribute (age) and eligibility for Adult dataset, selected best result per attribute

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.71	0.74	0.69	0.96	0.99	0.92	0.96	0.52
(0.66, 0.33)	0.7	0.67	0.73	0.88	0.94	0.83	0.89	0.54
(0.06, 0.36)	0.65	0.7	0.59	0.83	0.93	0.72	0.82	0.47

Table 27: Laftr-EqOpp1 results on correlation of sensitive attribute (age) and eligibility for Adult dataset, selected best result per attribute

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.72	0.75	0.7	0.96	0.99	0.91	0.95	0.53
(0.66, 0.33)	0.7	0.67	0.73	0.89	0.94	0.84	0.89	0.55
(0.06, 0.36)	0.65	0.71	0.59	0.83	0.93	0.72	0.82	0.47

Table 28: Laftr-EqOpp0 results on correlation of sensitive attribute (age) and eligibility for Adult dataset, selected best result per attribute

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.72	0.75	0.69	0.96	1.0	0.91	0.96	0.52
(0.66, 0.33)	0.7	0.67	0.72	0.88	0.94	0.83	0.89	0.55
(0.06, 0.36)	0.65	0.71	0.59	0.83	0.93	0.73	0.83	0.48

Table 29: Laftr-DP results on correlation of sensitive attribute (age) and eligibility for Adult dataset, selected best result per attribute

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.71	0.74	0.69	0.97	0.99	0.92	0.96	0.6
(0.66, 0.33)	0.7	0.67	0.73	0.88	0.94	0.83	0.89	0.57
(0.06, 0.36)	0.66	0.71	0.6	0.83	0.93	0.73	0.83	0.55

Table 30: Mlp results on correlation of sensitive attribute (bck) and eligibility for CI-MNIST dataset, selected best result per attribute

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.95	0.96	0.94	0.98	0.97	0.97	0.97	0.91
(0.1, 0.9)	0.81	0.82	0.81	0.65	0.66	0.63	0.65	0.96
(0.01, 0.99)	0.56	0.5	0.63	0.14	0.02	0.27	0.15	0.98

Table 31: Cnn results on correlation of sensitive attribute (bck) and eligibility for CI-MNIST dataset, selected best result per attribute

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.96	0.96	0.96	0.99	0.99	0.98	0.98	0.56
(0.1, 0.9)	0.88	0.9	0.86	0.79	0.83	0.74	0.78	0.85
(0.01, 0.99)	0.56	0.5	0.62	0.12	0.01	0.22	0.12	1.0

Table 32: Cfair results on correlation of sensitive attribute (bck) and eligibility for CI-MNIST dataset, selected best result per attribute

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.94	0.95	0.93	0.98	0.98	0.99	0.98	1.0
(0.1, 0.9)	0.82	0.85	0.79	0.65	0.73	0.6	0.67	1.0
(0.01, 0.99)	0.55	0.54	0.57	0.09	0.09	0.13	0.11	1.0

Table 33: Ffvae results on correlation of sensitive attribute (bck) and eligibility for CI-MNIST dataset, selected best result per attribute

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.91	0.92	0.9	0.98	1.0	0.96	0.98	1.0
(0.1, 0.9)	0.71	0.72	0.71	0.45	0.48	0.43	0.45	1.0
(0.01, 0.99)	0.51	0.49	0.53	0.03	0.0	0.05	0.03	1.0

Table 34: Laftr-EqOdd results on correlation of sensitive attribute (bck) and eligibility for CI-MNIST dataset, selected best result per attribute

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.96	0.97	0.96	0.99	0.99	0.98	0.98	1.0
(0.1, 0.9)	0.92	0.94	0.89	0.85	0.89	0.81	0.85	0.98
(0.01, 0.99)	0.65	0.66	0.64	0.3	0.33	0.28	0.31	0.97

Table 35: Laftr-EqOpp1 results on correlation of sensitive attribute (bck) and eligibility for CI-MNIST dataset, selected best result per attribute

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.97	0.98	0.96	0.99	0.99	0.99	0.99	0.99
(0.1, 0.9)	0.92	0.94	0.9	0.86	0.9	0.81	0.85	0.97
(0.01, 0.99)	0.67	0.69	0.65	0.34	0.39	0.28	0.34	0.97

Table 36: Laftr-EqOpp0 results on correlation of sensitive attribute (bck) and eligibility for CI-MNIST dataset, selected best result per attribute

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.96	0.97	0.96	0.99	0.99	0.99	0.99	0.99
(0.1, 0.9)	0.91	0.94	0.88	0.85	0.91	0.79	0.85	0.94
(0.01, 0.99)	0.64	0.65	0.63	0.28	0.31	0.25	0.28	0.96

Table 37: Laftr-DP results on correlation of sensitive attribute (bck) and eligibility for CI-MNIST dataset, selected best result per attribute

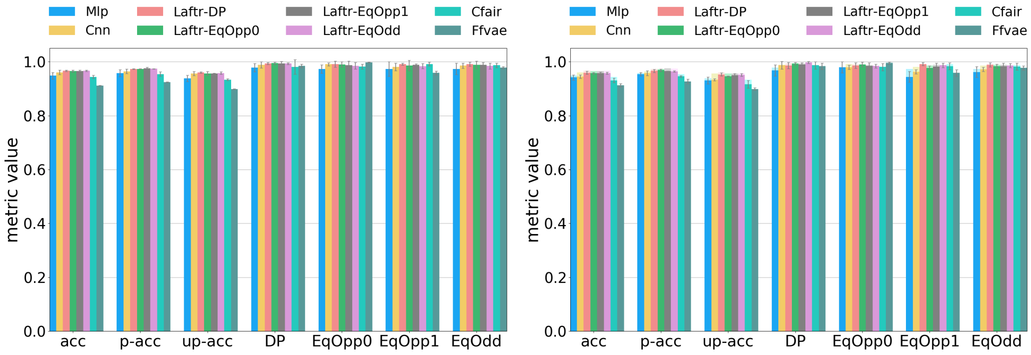
clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.96	0.97	0.96	0.99	0.99	0.99	0.99	1.0
(0.1, 0.9)	0.92	0.94	0.89	0.85	0.91	0.8	0.85	1.0
(0.01, 0.99)	0.65	0.65	0.65	0.3	0.32	0.28	0.3	0.99

E.3 Impact of correlation of non-sensitive attribute with eligibility

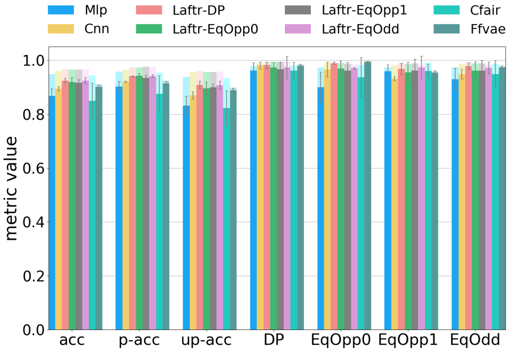
We report the complete set of results for debiasing models of Mlp, Cnn, Cfair, Ffvae, Laftr-EqOdd, Laftr-EqOpp1, Laftr-EqOpp0, and Laftr-DP, in Tables 38 to 45, corresponding to the experimental setup described in Setting 3 of Secton 4 in the main paper. Each pair in *pos-ratio* column indicate (l_e, l_o), which specifies the ratio of images with box on left side for (even=eligible, odd=ineligible). Figure 9 compare all models side-by-side.

Table 38: Mlp results on correlation of non-sensitive attribute (pos) and eligibility for CI-MNIST dataset, selected best result per attribute

pos	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.95	0.96	0.94	0.98	0.97	0.97	0.97	0.91
(0.75, 0.25)	0.94	0.95	0.93	0.97	0.98	0.94	0.96	0.98
(0.9, 0.1)	0.86	0.9	0.83	0.96	0.9	0.96	0.93	1.0



(a) *pos-ratio* $(0.5, 0.5)$ (b) *pos-ratio* $(0.75, 0.25)$



(c) *pos-ratio* $(0.9, 0.1)$

Figure 9: Comparing different models while shifting correlation of a non-sensitive attribute and the eligibility for CI-MNIST dataset. In sub-figures 9(b) and 9(c) the pale colors show the decrease in performance compared to the balanced case in 9(a).

Table 39: Cnn results on correlation of non-sensitive attribute (pos) and eligibility for CI-MNIST dataset, selected best result per attribute

pos	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.96	0.96	0.96	0.99	0.99	0.98	0.98	0.56
(0.75, 0.25)	0.95	0.96	0.93	0.99	0.98	0.96	0.97	0.57
(0.9, 0.1)	0.9	0.92	0.87	0.98	0.97	0.93	0.95	0.63

Table 40: Cfair results on correlation of non-sensitive attribute (pos) and eligibility for CI-MNIST dataset, selected best result per attribute

pos	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.94	0.95	0.93	0.98	0.98	0.99	0.98	1.0
(0.75, 0.25)	0.94	0.95	0.92	0.99	0.98	0.98	0.98	1.0
(0.9, 0.1)	0.85	0.88	0.82	0.96	0.94	0.96	0.95	1.0

Table 41: Ffvae results on correlation of non-sensitive attribute (pos) and eligibility for CI-MNIST dataset, selected best result per attribute

pos	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.91	0.92	0.9	0.98	1.0	0.96	0.98	1.0
(0.75, 0.25)	0.92	0.93	0.9	0.98	1.0	0.96	0.98	1.0
(0.9, 0.1)	0.9	0.91	0.89	0.98	0.99	0.95	0.97	1.0

Table 42: Laftr-EqOdd results on correlation of non-sensitive attribute (pos) and eligibility for CI-MNIST dataset, selected best result per attribute

pos	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.96	0.97	0.96	0.99	0.99	0.98	0.98	1.0
(0.75, 0.25)	0.95	0.96	0.95	1.0	0.98	0.99	0.98	1.0
(0.9, 0.1)	0.93	0.94	0.91	0.97	0.97	0.97	0.97	1.0

Table 43: Laftr-EqOpp1 results on correlation of non-sensitive attribute (pos) and eligibility for CI-MNIST dataset, selected best result per attribute

pos	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.97	0.98	0.96	0.99	0.99	0.99	0.99	0.99
(0.75, 0.25)	0.96	0.97	0.95	0.99	0.98	0.98	0.98	0.99
(0.9, 0.1)	0.92	0.93	0.9	0.97	0.96	0.96	0.96	0.99

Table 44: Laftr-EqOpp0 results on correlation of non-sensitive attribute (pos) and eligibility for CI-MNIST dataset, selected best result per attribute

pos	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.96	0.97	0.96	0.99	0.99	0.99	0.99	0.99
(0.75, 0.25)	0.96	0.97	0.95	0.99	0.99	0.98	0.98	0.99
(0.9, 0.1)	0.92	0.94	0.9	0.97	0.97	0.95	0.96	0.99

Table 45: Laftr-DP results on correlation of non-sensitive attribute (pos) and eligibility for CI-MNIST dataset, selected best result per attribute

pos	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.96	0.97	0.96	0.99	0.99	0.99	0.99	1.0
(0.75, 0.25)	0.96	0.97	0.95	0.99	0.99	0.99	0.99	1.0
(0.9, 0.1)	0.93	0.94	0.91	0.98	0.99	0.97	0.98	1.0

E.4 Impact of position and small features in the input images

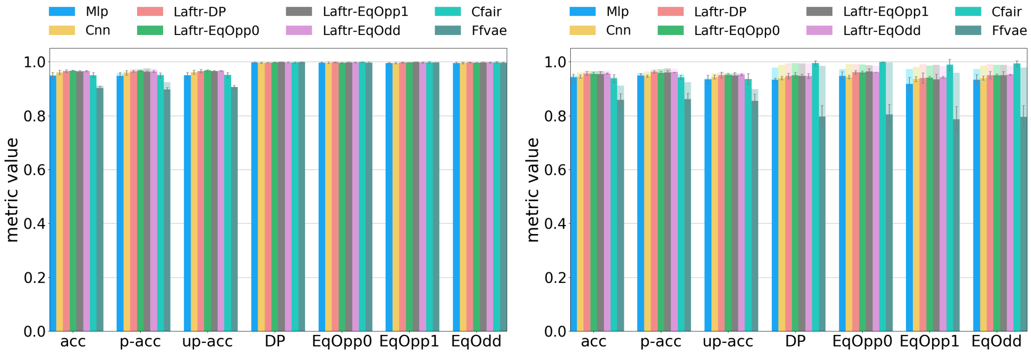
Comparing baseline model with debiasing models of Mlp, Cfair, Ffvae, Laftr-EqOdd, Laftr-EqOpp1, Laftr-EqOpp0, and Laftr-DP, when position and a small feature of the image correlates with eligibility. Results are depicted in Figure in Tables 46 to 53, corresponding to the experimental setup described in Setting 4 of Section 4 in the main paper. Each pair in *pos-ratio* column indicate (l_e, l_o), which specifies the ratio of images with box on left side for (even=eligible, odd=ineligible). Figure 10 compare all models side-by-side.

Table 46: Mlp results on correlation of sensitive attribute (pos) and eligibility for CI-MNIST dataset, selected best result per attribute

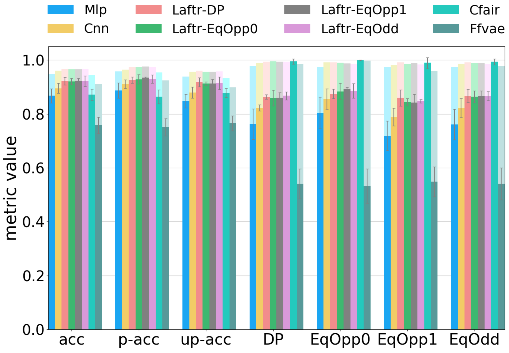
pos-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.95	0.95	0.95	1.0	1.0	0.99	0.99	0.51
(0.75, 0.25)	0.94	0.95	0.93	0.93	0.95	0.92	0.94	0.59
(0.9, 0.1)	0.87	0.89	0.85	0.76	0.8	0.72	0.76	0.67

Table 47: Cnn results on correlation of sensitive attribute (pos) and eligibility for CI-MNIST dataset, selected best result per attribute

pos-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.96	0.96	0.96	1.0	1.0	0.99	0.99	0.56
(0.75, 0.25)	0.94	0.95	0.94	0.94	0.94	0.94	0.94	0.79
(0.9, 0.1)	0.9	0.91	0.88	0.82	0.86	0.79	0.82	0.88



(a) *pos-ratio* $(0.5, 0.5)$ (b) *pos-ratio* $(0.75, 0.25)$



(c) *pos-ratio* $(0.9, 0.1)$

Figure 10: Impact of position and small visual components on different models' performance for CI-MNIST dataset. In sub-figures 10(b) and 10(c) the pale colors show the decrease in performance compared to the balanced case in 10(a).

Table 48: Cfair results on correlation of sensitive attribute (*pos*) and eligibility for CI-MNIST dataset, selected best result per attribute

pos-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.95	0.95	0.95	1.0	1.0	1.0	1.0	0.82
(0.75, 0.25)	0.94	0.94	0.94	0.99	1.0	0.99	0.99	0.88
(0.9, 0.1)	0.87	0.86	0.88	0.99	1.0	0.99	0.99	0.92

Table 49: Ffvae results on correlation of sensitive attribute (*pos*) and eligibility for CI-MNIST dataset, selected best result per attribute

pos-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.91	0.9	0.91	1.0	1.0	1.0	1.0	1.0
(0.75, 0.25)	0.86	0.86	0.86	0.8	0.81	0.79	0.8	1.0
(0.9, 0.1)	0.76	0.75	0.77	0.54	0.53	0.55	0.54	1.0

Table 50: Laftr-EqOdd results on correlation of sensitive attribute (*pos*) and eligibility for CI-MNIST dataset, selected best result per attribute

pos-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.96	0.96	0.97	1.0	1.0	1.0	1.0	0.99
(0.75, 0.25)	0.95	0.96	0.95	0.95	0.96	0.94	0.95	0.64
(0.9, 0.1)	0.92	0.93	0.91	0.87	0.89	0.85	0.87	0.72

Table 51: Laftr-EqOpp1 results on correlation of sensitive attribute (pos) and eligibility for CI-MNIST dataset, selected best result per attribute

pos-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.96	0.96	0.96	1.0	1.0	1.0	1.0	0.93
(0.75, 0.25)	0.95	0.96	0.95	0.95	0.96	0.93	0.95	0.65
(0.9, 0.1)	0.92	0.93	0.91	0.86	0.89	0.84	0.86	0.75

Table 52: Laftr-EqOpp0 results on correlation of sensitive attribute (pos) and eligibility for CI-MNIST dataset, selected best result per attribute

pos	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.97	0.97	0.97	1.0	1.0	1.0	1.0	0.74
(0.75, 0.25)	0.95	0.96	0.95	0.95	0.96	0.94	0.95	0.61
(0.9, 0.1)	0.92	0.93	0.91	0.86	0.88	0.84	0.86	0.74

Table 53: Laftr-DP results on correlation of sensitive attribute (pos) and eligibility for CI-MNIST dataset, selected best result per attribute

pos-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.96	0.96	0.97	1.0	1.0	1.0	1.0	1.0
(0.75, 0.25)	0.95	0.96	0.95	0.95	0.96	0.94	0.95	0.95
(0.9, 0.1)	0.93	0.93	0.92	0.86	0.87	0.86	0.86	0.97

E.5 Impact of seed

In Figures 11 and 12 we illustrate the standard deviation of all models for all of the experiments of Adult and CI-MNIST datasets described in Section 4 of the main paper.

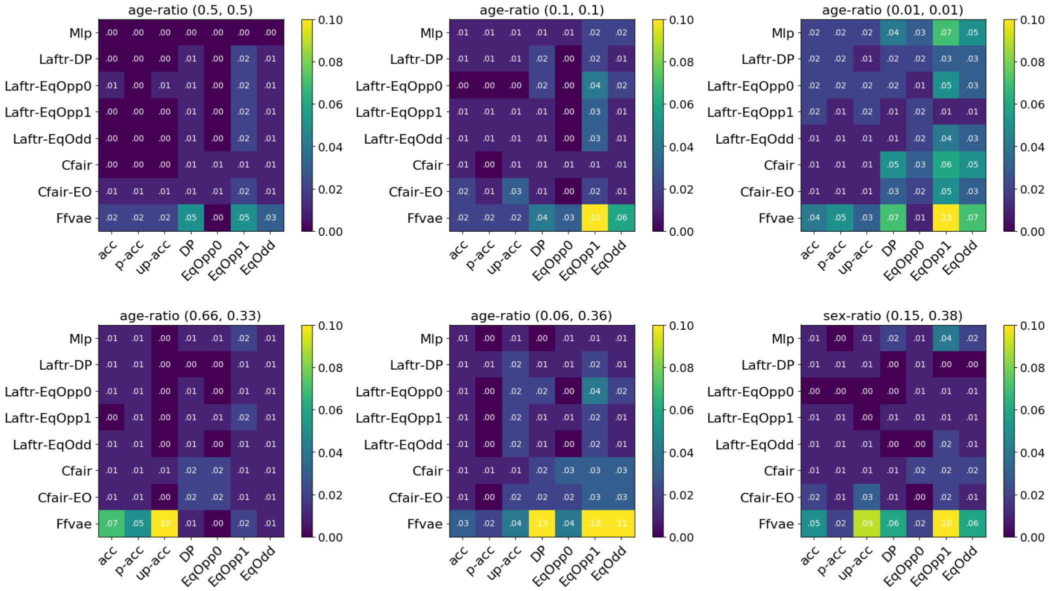


Figure 11: Standard deviation of different fairness metrics (x-axis) in different models (y-axis) over three seeds for Adult dataset. Each plot corresponds to a different experimental setup presented in Section 4.

E.6 Correlation between dataset features and model's prediction.

In Figure 13 we present Spearman Correlation plots for each dataset and each setting of the experiments presented in Section 4. Please check Section 5.1 of the main paper for the corresponding section.

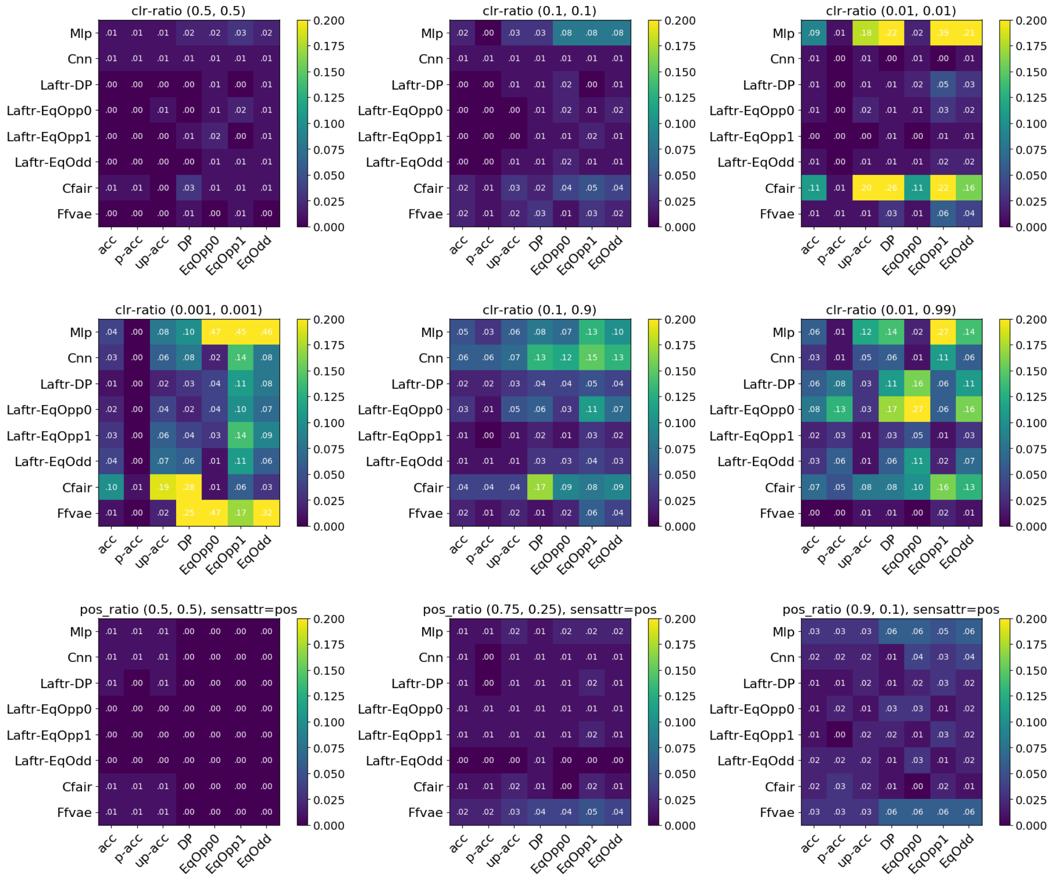


Figure 12: Standard deviation of different fairness metrics (x-axis) in different models (y-axis) over three seeds for CI-MNIST dataset. Each plot corresponds to a different experimental setup presented in Section 4.

E.7 Impact of small population bias

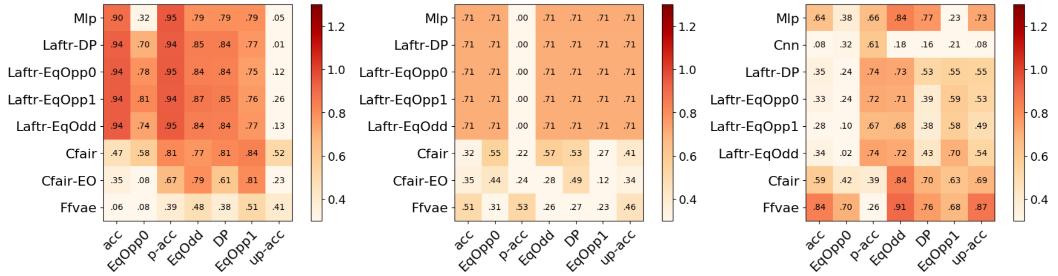
Table 54 presents results for Cnn and Table 55 for Laftr-EqOpp0, where clr-ratios is kept at (0.001, 0.001) but the total dataset size has changed from x to $10x$, $100x$ and $1000x$. Refer to Small percentage of the unprivileged group part in Section 5.2 of the main text.

Table 54: Cnn results for measuring whether the bias is due to small ratio or small number of samples.

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.001, 0.001) x	0.93	0.97	0.88	0.93	0.98	0.85	0.92	0.5
(0.001, 0.001) 10x	0.96	0.98	0.95	0.99	0.98	0.96	0.97	0.51
(0.001, 0.001) 100x	0.9	0.98	0.82	0.95	0.9	0.79	0.84	0.52
(0.001, 0.001) 1000x	0.9	0.98	0.82	0.95	0.89	0.78	0.83	0.52

Table 55: Laftr-EqOpp0 results for measuring whether the bias is due to small ratio or small number of samples.

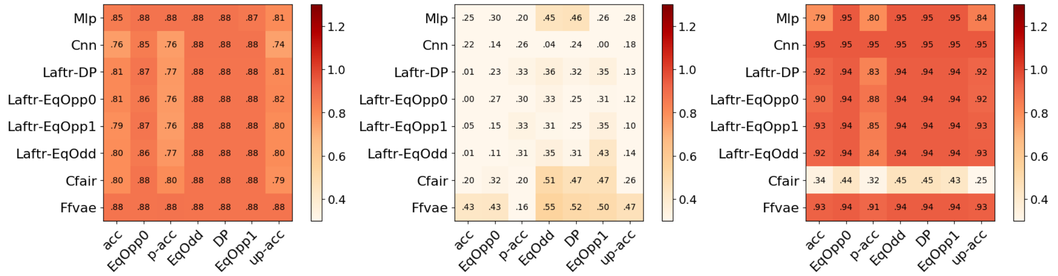
clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.001, 0.001) x	0.94	0.98	0.89	0.95	0.96	0.88	0.92	0.67
(0.001, 0.001) 10x	0.91	0.98	0.83	0.94	0.81	0.9	0.85	0.5
(0.001, 0.001) 100x	0.94	0.98	0.89	0.99	0.91	0.9	0.91	0.54
(0.001, 0.001) 1000x	0.96	0.99	0.93	0.99	0.95	0.93	0.94	0.54



(a) Adult, Setting 1, showing correlation for *age-ratio* attribute with fairness metrics

(b) Adult, Setting 2, showing correlation for *age-ratio* attribute with fairness metrics

(c) CI-MNIST, Setting 1, showing correlation for *clr-ratio* attribute with fairness metrics



(d) CI-MNIST, Setting 2, showing correlation for *clr-ratio* attribute with fairness metrics

(e) CI-MNIST, Setting 3, showing correlation for *pos-ratio* attribute with fairness metrics

(f) CI-MNIST, Setting 4, showing correlation for *pos-ratio* attribute with fairness metrics

Figure 13: Each plot depicts correlation of one dataset attribute with fairness metrics for one setting and one dataset in Section 4. On the Adult dataset, we depict the correlation of *age-ratio* with fairness metrics as this attribute has been the sensitive feature that is changed in the experiments. On CI-MNIST , in Settings 1 and 2, we depict *clr-ratio*, and in Settings 3 and 4, we show *pos-ratio*, hence showing only the feature that is changed from the balanced case. Note that contrary to other cases, in Setting 3 *pos-ratio* is not the sensitive attribute, and background is the sensitive attribute. We plot the absolute Spearman correlation metric, where we use absolute difference of the dataset attribute from the balanced case (0.5) as input to the Spearman function. This is because numbers such as 1 and 0 have a similar meaning as they are equally away from the balanced case. Finally, we report absolute averaged correlation values over all cases. Values range in $[0, 1]$, where one indicates maximum correlation. Almost all bias-mitigation models suffer from not mitigating the strong correlation between the overall accuracy and the sensitive attribute.

E.8 Merging bias-mitigation algorithms.

Tables 56 and 57 show results for merging Cfair and Ffvae and Tables 58 and 59 show results for merging Laftr and Ffvae.

To merge Ffvae with Laftr we added to Ffvae objective in Eq.(8), the $\mathcal{L}_{DP}^{Laftr}$ term in Eq.(3), yielding

$$\mathcal{L}_{DP}^{Ffvae-Laftr} = \mathcal{L}_{\text{Ffvae}}(p, q) - \eta \mathcal{L}_{DP}^{Laftr} \tag{10}$$

Similarly, to merge Ffvae with Cfair we added to Ffvae objective in Eq.(8), the $\mathcal{L}_{DP}^{Cfair}$ term in Eq.(7), yielding

$$\mathcal{L}_{DP}^{Ffvae-Cfair} = \mathcal{L}_{\text{Ffvae}}(p, q) - \eta \mathcal{L}_{DP}^{Cfair} \tag{11}$$

where η is a hyper-parameter, balancing the two losses. In both cases, the added loss ($\mathcal{L}_{DP}^{Laftr}$ or $\mathcal{L}_{DP}^{Cfair}$) is applied to non-sensitive latent z of Ffvae model. Please check Section 5.1 of the main paper for the discussion on the obtained results.

Table 56: Merged Ffvae and Cfair results when decreasing minority representation for Adult dataset, sensitive attribute:age. Added $\mathcal{L}_{DP}^{Cfair}$ to Eq.(8). Compare with Ffvae Table 9 and Cfair Table 7 results.

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.77	0.81	0.73	0.99	0.99	0.97	0.98	0.92
(0.1, 0.1)	0.75	0.78	0.72	0.99	1.0	0.99	0.99	0.98
(0.01, 0.01)	0.72	0.83	0.62	0.94	1.0	0.85	0.93	0.79

Table 57: Merged Ffvae and Cfair results on correlation of sensitive attribute (age) and eligibility for Adult dataset. Added $\mathcal{L}_{DP}^{Cfair}$ to Eq.(8). Compare with Ffvae Table 25 and Cfair Table 23 results.

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.77	0.81	0.73	0.99	0.99	0.97	0.98	0.92
(0.66, 0.33)	0.74	0.73	0.75	1.0	1.0	1.0	1.0	0.92
(0.06, 0.36)	0.7	0.76	0.64	0.98	0.99	0.96	0.97	0.97

Table 58: Merged Ffvae and Laftr-DP results when decreasing minority representation for Adult dataset, sensitive attribute:age. Added $\mathcal{L}_{DP}^{Laftr}$ to Eq.(8). Compare with Ffvae Table 9 and Laftr-DP Table 13 results.

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.66	0.71	0.61	0.99	1.0	0.98	0.99	0.93
(0.1, 0.1)	0.6	0.67	0.54	1.0	1.0	1.0	1.0	0.98
(0.01, 0.01)	0.6	0.69	0.52	1.0	1.0	1.0	1.0	0.92

Table 59: Merged Ffvae and Laftr-DP results on correlation of sensitive attribute (age) and eligibility for Adult dataset. Added $\mathcal{L}_{DP}^{Laftr}$ to Eq.(8). Compare with Ffvae Table 25 and Laftr-DP Table 29 results.

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.66	0.71	0.61	0.99	1.0	0.98	0.99	0.93
(0.66, 0.33)	0.49	0.5	0.49	1.0	1.0	1.0	1.0	0.93
(0.06, 0.36)	0.6	0.69	0.51	0.97	0.99	0.95	0.97	0.98