SUPPLEMENTARY MATERIAL

A LIMITATIONS

While the framework proposed shows evidence that models can hardly discriminate against samples with low uncertainty in the sensitive attribute prediction, our method relies on the prediction of missing sensitive information. Inferring sensitive information can raise ethical concerns and face legal restrictions, especially when individuals do not choose to disclose their sensitive information. The line of methods relying upon proxy attributes or inferring sensitive attributes faces this limitation (Diana et al., 2022; Awasthi et al., 2021; Coston et al., 2019). Furthermore, the inference of the sensitive attributes (using our proposed method or others) should not be used for a purpose different from bias assessment mitigation. Moreover, we showed that with a variant of our method (*Ours (uncertain)*), it is possible to train a fairer model without fairness constraints but using samples with high uncertainty in the sensitive attribute predictions. This variant can operate without (predicted) sensitive information, thereby alleviating potential legal or ethical concerns associated with the prediction of sensitive information. We show in Appendix D the impact of the uncertainty threshold on the fairness of a model trained without fairness constraints. Another limitation of our method is the evaluation focuses mostly on one fairness-enhancing algorithm (i.e., Exponentiated Gradient). It will be interesting to explore if our hypothesis applies to pre-processing and post-processing techniques, and with different fairness-enhancing algorithms. Finally, our assumption that the true sensitive attributes are available in the test dataset for fairness evaluation might not be true in several practical scenarios. This might require evaluation using proxy-sensitive attributes. These proxies are likely noisy and might require evaluations using bias assessment methods that effectively quantify fairness violation w.r.t to true sensitive attribute (Chen et al., 2019; Awasthi et al., 2021).

B FAIRNESS METRICS

In this work, we consider three popular group fairness metrics: demographic parity, equalized odds, and equal opportunity. These metrics aim to equalize different models' performances across different demographic groups. Samples belong to the same demographic group if they share the same demographic information, A, e.g., gender and race.

B.1 DEMOGRAPHIC PARITY

Also known as statistical parity, this measure requires that the positive prediction ($f(X) = 1$) of the model be the same regardless of the demographic group to which the user belongs (Dwork et al., 2012). More formally the classifier f achieves demographic parity if $P(f(X) = 1|A = a) = P(f(X) = 1)$ In other words, the outcome of the model should be independent of sensitive attributes. In practice, this metric is measured as follows:

$$\Delta_{\text{DP}}(f) = \left| \mathop{\mathbb{E}}_{x|A=0}[\mathbb{I}\{f(x) = 1\}] - \mathop{\mathbb{E}}_{x|A=1}[\mathbb{I}\{f(x) = 1\}] \right| \tag{6}$$

Where $\mathbb{I}(\cdot)$ is the indicator function.

B.2 EQUALIZED ODDS

This metric enforces the True Positive Rate (TPR) and False Positive Rate (FPR) for different demographic groups to be the same $P(f(X) = 1|A = 0, Y = y) = P(f(X) = 1|A = 1, Y = y), : ; \forall y \in \{0, 1\}$. The metric is measured as follows:

$$\Delta_{\text{EOD}} = \alpha_0 + \alpha_1 \tag{7}$$

Where,

$$\alpha_j = \left| \mathop{\mathbb{E}}_{x|A=0,Y=j}[\mathbb{I}\{f(x) = 1\}] - \mathop{\mathbb{E}}_{x|A=1,Y=j}[\mathbb{I}\{f(x) = 1\}] \right| \tag{8}$$

Dataset	Size	#features	Domain	Sensitive attribute
Adult Income	48,842	15	Finance	Gender
Compas	6,000	11	Criminal justice	Race
LSAC	20,798	12	Education	Gender
New Adult	1.6M	10	Finance	Gender
CelebA	202,600	40	Image	Gender

Table 5: Datasets

B.3 Equal Opportunity

In certain situations, bias in positive outcomes can be more harmful. Therefore, Equal Opportunity metric enforces the same TPR across demographic (Hardt et al., 2016) and is measured using α_1 (Eq. 8).

C Datasets

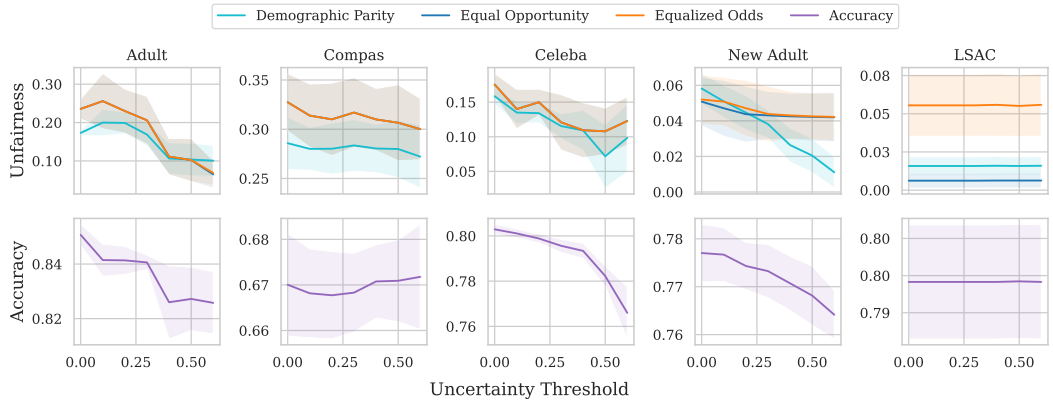
In the Adult dataset, the goal is to predict if an individual's annual income is greater or less than \$50k per year. We also considered the recent version of the Adult dataset (New Adult) for the year 2018 across different states in US (Ding et al., 2021). The goal in the Compas data is to predict whether a defendant will recidivate within two years. The LSAC dataset contains admission records to law school. The task is to predict whether a candidate would pass the bar exam. The CelebA[4] dataset contains 40 facial attributes of humaine annotated images. We consider the task of predicting *attractiveness* using gender as a sensitive attribute. We randomly sample 20% of the data to train the sensitive attribute classifier ($\mathcal{D}_2$). For all the datasets, all the features are used to train the attribute classifier except for the target variable. Table 5 provides more details about each dataset and sensitive attributes used.

D Using the Uncertainty of Sensitive Attributes to Train Fair Models without Fairness Constraints.

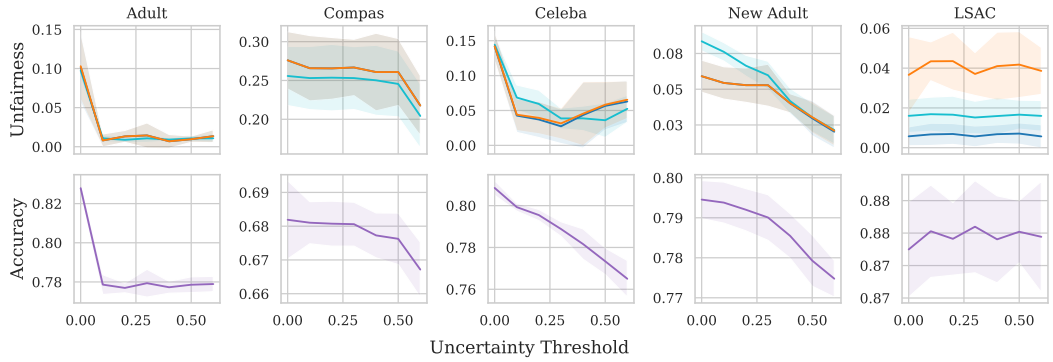
Results in Table 4 show that a model trained without fairness constraints tends to be fairer when the average uncertainty in predicting the sensitive attribute is high. This provides the intuition that training a model on samples with uncertain sensitive attributes can yield fairer outcomes. In this set of experiments, we trained different classifiers (Logistic Region and Random Forest) without fairness constraints but using training data with uncertain sensitive attributes. For different uncertainty thresholds $H \in \{0.0, 0.1, ..., 0.6\}$, we prune out samples whose uncertainty is lower than H and train a model without fairness constraints using the remaining training data, i.e., $\{(x, y) \in D_1 | u_x \geq H\}$. Where u_x is the estimated uncertainty provided by our method. In particular, when $H = 0$ all the data points are used for the training, and in other cases samples with uncertainty lower than H are removed, and the model is trained on samples with more uncertain sensitive attributes. For each uncertainty threshold, we train the model seven times with different seeds and report fairness and accuracy in the testing set, which contains sensitive attributes.

Figures 3(a) and 3(b) show the fairness and accuracy of the Logistic Regression and Random Forest classifiers, respectively. In the figures, each column represents the results on each dataset, and the first and the second rows provide the plots for fairness and accuracy respectively. Across different datasets, the results show that unfairness decreases as the uncertainty threshold increases. We observe that the improvement in fairness is consistent for different fairness metrics considered, i.e., demographic parity, equal opportunity, and equalized odd. We also observe a decrease in the accuracy, which is justified by reduced dataset size and consequently the tradeoff with fairness. On the LSAC dataset, fairness and accuracy remain almost constant as the average uncertainty in predicting the

[4] http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html



(a) Logistic Regression



(b) Random Forest

Figure 3: Training classifiers without fairness constraints using samples with high uncertainty of sensitive attribute predictions. For each uncertainty threshold H, the model is trained on samples with uncertainty $\geq H$. The training is done seven times and the average fairness (first row) and accuracy (second row) are reported. Shaded represents the standard deviation.

sensitive attribute on this dataset is 0.66, i.e., most of the samples already have the highest uncertainty. However, this method incurs a higher drop in accuracy and does not necessarily guarantee that an adversary cannot reconstruct the sensitive attributes from the trained model (Ferry et al., 2023).

E Additional results.

Table 6 shows the comparison with other baselines on the CelebA dataset with logistic regression as the base classifier. In the main paper, we provided a comparison with other existing methods using Logistic Regression as the base classifier. We performed experiments with a more complex non-linear model in order to analyze its impact on the performance of different methods. We considered a Multi-Layer Perceptron (MLP) with one hidden layer of 32 units and with Relu as the activation function for all the baselines. On the Adult dataset, Table 7, shows that when using a more complex model, our method (Ours (certain)) still provides Pareto dominants points in terms of fairness and accuracy compared to other baselines while we observed an improvement in the accuracy of other methods due to the increased model capacity.

F Comparison with other baselines

To assess the effect of the attribute classifier over the performances of downstream classifiers with fairness constraints w.r.t the proxy, we considered different methods of obtaining the missing sensitive attributes as baselines:

Method	Accuracy	Δ_{DP}	Δ_{EOP}	Δ_{EOD}
Vanilla (without fairness)	0.803 ± 0.002	0.176 ± 0.010	0.183 ± 0.015	0.183 ± 0.015
Vanilla (with fairness)	0.782 ± 0.001	0.008 ± 0.005	0.018 ± 0.014	0.017 ± 0.014
FairDA	0.802 ± 0.002	0.155 ± 0.010	0.165 ± 0.018	0.165 ± 0.018
ARL	$\mathbf{0.803} \pm 0.002$	0.157 ± 0.010	0.157 ± 0.010	0.166 ± 0.016
CVarDRO	0.781 ± 0.002	0.155 ± 0.010	0.162 ± 0.016	0.162 ± 0.016
KSMOTE	0.773 ± 0.008	0.020 ± 0.067	0.110 ± 0.082	0.144 ± 0.068
DRO	0.796 ± 0.006	0.142 ± 0.020	0.152 ± 0.020	0.129 ± 0.028
Ours (uncertain)	0.782 ± 0.004	0.071 ± 0.046	0.107 ± 0.033	0.107 ± 0.033
Ours (certain)	0.793 ± 0.000	$\mathbf{0.003} \pm 0.000$	$\mathbf{0.001} \pm 0.001$	$\mathbf{0.007} \pm 0.002$

Table 6: Comparison with different baselines on the CelebA dataset. *Ours (uncertain)* represents the variant of our approach where the model is trained without fairness constraints but using samples with higher uncertainty in the sensitive attribute predictions. And *Ours (certain)* the variant where only samples with reliable sensitive attributes are used to train the label classifier with fairness constraints using the exponentiated gradient.

Method	Accuracy	Δ_{DP}	Δ_{EOP}	Δ_{EOD}
Vanilla (without fairness)	0.853 ± 0.004	0.183 ± 0.019	0.100 ± 0.025	0.102 ± 0.023
Vanilla (with fairness)	0.801 ± 0.009	0.006 ± 0.004	0.049 ± 0.011	0.017 ± 0.007
FairRF	$\mathbf{0.853} \pm 0.002$	0.164 ± 0.009	0.077 ± 0.026	0.091 ± 0.013
FairDA	0.813 ± 0.014	0.118 ± 0.023	0.091 ± 0.050	0.099 ± 0.037
ARL	0.851 ± 0.003	0.166 ± 0.015	0.087 ± 0.019	0.090 ± 0.016
CVarDRO	0.850 ± 0.003	0.183 ± 0.018	0.095 ± 0.027	0.101 ± 0.026
KSMOTE	0.814 ± 0.020	0.201 ± 0.055	0.120 ± 0.021	0.130 ± 0.023
DRO	0.837 ± 0.016	0.232 ± 0.057	0.110 ± 0.057	0.140 ± 0.045
Ours (uncertain)	0.801 ± 0.027	0.110 ± 0.022	0.067 ± 0.027	0.059 ± 0.024
Ours (certain)	0.818 ± 0.004	$\mathbf{0.009} \pm 0.008$	$\mathbf{0.028} \pm 0.020$	$\mathbf{0.027} \pm 0.017$

Table 7: Comparison with different baselines on the Adult dataset using an MLP with a hidden layer of 34 units as base classifier.

- **Ground truth sensitive attribute**. We considered the case where the sensitive attribute is fully available and trained the model with fairness constraints w.r.t the ground truth. This represents the ideal situation where all the assumptions about the availability of demographic information are satisfied. This baseline is expected to achieve the best trade-offs.
- **Proxy-KNN**. Here the missing sensitive attributes are handled by data imputation using the k-nearest neighbors (KNN) of samples with missing sensitive attributes.
- **Proxy-DNN**. For this baseline, an MLP is trained on $\mathcal{D}_2$ to predict the sensitive attributes without uncertainty awareness. The network architecture used and the hyperparameter is the same as for the student in our model.

For fairness-enhancing mechanisms, we considered the Fairlean (Bird et al., 2020) implementation of the exponentiated gradient (Agarwal et al., 2018) and adversarial debiasing (Zhang et al., 2018) (Section 3). For the exponentiated gradient, we used various base classifiers including logistic regression, random forest, and gradient-boosted trees. Random forest was initialized with a maximum depth of 5 and minimum samples leaf 10, and default parameters were used for logistic regression without hyperparameter tuning. The same models and hyperparameters were used across all the datasets. Adversarial debiasing works for demographic parity and equalized odds. The architecture of the classifier, the adversary, as well as other hyperparameters used is the same as recommended by the original paper (Zhang et al., 2018). We evaluate the fairness-accuracy trade-off of every baseline by analyzing the accuracy achieved in different fairness regimes, i.e., by varying the parameter

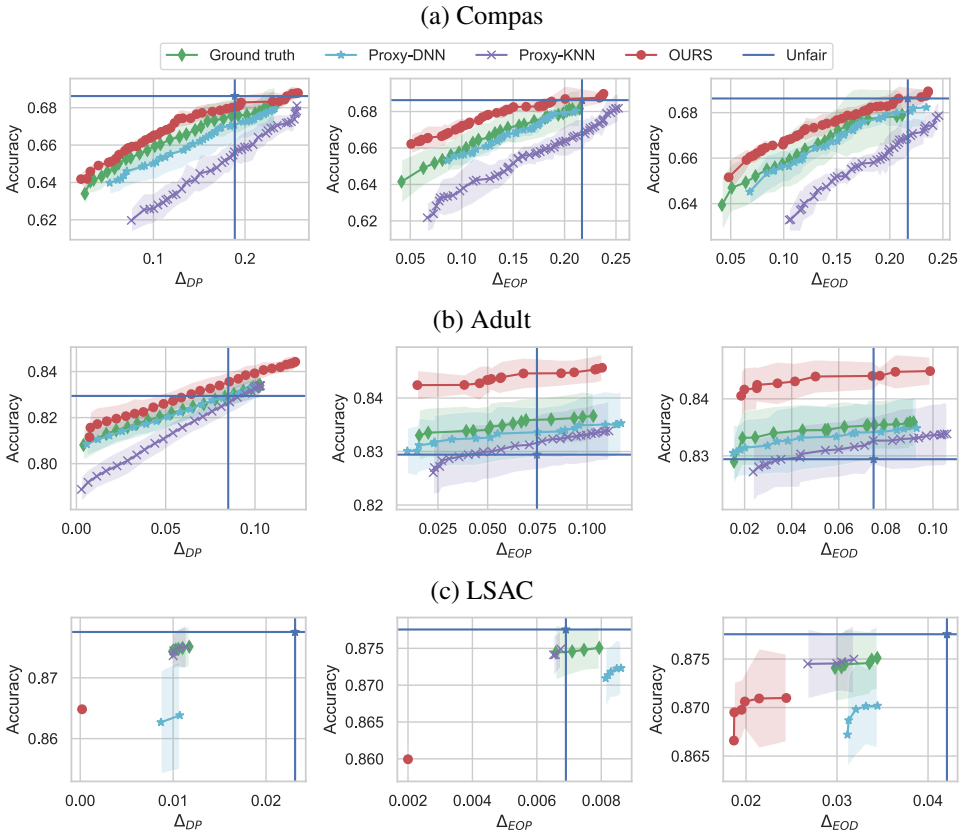


Figure 4: Accuracy-fairness trade-offs for various fairness metrics (Δ_{DP}, Δ_{EOP}, Δ_{EOD}) and proxy sensitive attributes. Top-left is the best (highest accuracy with the lowest unfairness). Curves are created by sweeping a range of fairness coefficients λ, taking the median of 7 runs per λ, and computing the Pareto front. The fairness mechanism used is the exponentiated gradient with Random Forests as the base classifier. Shaded in the figure are the standard deviations.

$\epsilon \in [0, 1]$ controlling the balance between fairness and accuracy. For a value of ϵ close to 0, the label classifier is enforced to achieve higher accuracy while for a value close to 1 it is encouraged to achieve lower unfairness. For each value of ϵ, we trained each baseline 7 times on a random subset of $\mathcal{D}_1$ (70%) using their predicted sensitive attributes, and the accuracy and fairness are measured on the remaining subset (30%), where we assumed that the joint distribution (X, Y, A) is available for fairness evaluation. The results are averaged and the Pareto front is computed.

Figure 4 shows the Pareto front of the exponentiated gradient method on the Adult, Compas, and LSAC datasets using Random Forests as the base classifier. The figure shows the trade-off between fairness and accuracy for the different methods of inferring the missing sensitive attributes. From the results, we observe on all datasets and across all fairness metrics that data imputation can be an effective strategy for handling missing sensitive attributes, i.e., this fairness mechanism can efficiently improve the fairness of the model with respect to the true sensitive attributes although fairness constraints were enforced on proxy-sensitive attributes. However, we observe a difference in the fairness-accuracy trade-off for each attribute classifier. Overall, the KNN-based attribute classifier has the worst fairness-accuracy trade-off on all datasets and fairness metrics. This shows that assigning sensitive attributes based on the nearest neighbors does not produce sensitive attributes useful for achieving a trade-off close to the ground truth. While the DNN-based attribute classifier produces a better trade-off but is still suboptimal compared to the ground truth sensitive attributes. We observed similar results with different baseline models such as logistic regression and gradient-booted trees and for adversarial debiasing as the fairness mechanism. In contrast, we see that our method consistently achieves a better trade-off on all datasets and across all the fairness metrics considered. Similar results are obtained on the exponentiated gradient with logistic regression and gradient-boosted trees as base classifiers and with adversarial debiasing (see Section 6). The choice of the uncertainty threshold depends on the level of bias in the dataset, i.e. the level of information about the sensitive attribute encoded in the feature space.

Figure 6 depicts the Pareto front of various baselines on the CelebA dataset. It shows that models trained with imputed sensitive attributes via KNN consistently achieve comparable tradeoffs to models trained with fairness constraints based on the true sensitive attribute. This could be explained by the fact that gender clusters are perfectly defined in the latent space. We observed that KNN-based imputation achieved 95% accuracy the assigning the right gender. Conversely, the figures illustrate that our method outperforms baselines using both ground truth-sensitive attributes and imputation, yielding a greater number of Pareto-dominant points. This highlights the advantages of applying fairness constraints to samples with reliable sensitive attributes. Furthermore, Figure 8 (c) shows decreasing the uncertainty threshold further improves fairness while preserving the accuracy. We note the CelebA dataset can raise ethical concerns and is used only for evaluation purposes. For instance, the task of predicting the attractiveness of a photo using other facial attributes as sensitive attributes can still harm individuals even if the model predicting attractiveness is not *biased*.

F.1 Exponentiated gradient with different baseline classifiers

Figure 5 and 7 show fairness-accuracy trade-offs achieved by the exponentiated gradient with logistic regression, and gradient-boosted trees, respectively. Similar to the results presented in the main paper, our method achieves better fairness-accuracy trade-offs.

Figure 8 shows the accuracy-fairness trade-off Exponentiated gradient using gradient-boosted trees as the base classifier for various uncertainty thresholds, the true sensitive attributes, and the predicted sensitive attributes with DNN. The results obtained are similar to random forests as the base classifier. The smaller uncertainty threshold produced the best trade-off in a high-bias regime such as the Adult dataset. While on datasets that do not encode much information about the sensitive attributes (most samples have high uncertainty) such as the New Adult and Compas datasets, the accuracy decreases as the uncertainty threshold reduces while fairness is improved or maintained. On the LSAC dataset (Figure 8(d)), we observe that increasing the uncertainty threshold results in a much higher drop in accuracy. This is explained by the fact that the average uncertainty is very high (0.66) and using a smaller threshold prunes out most of the data.

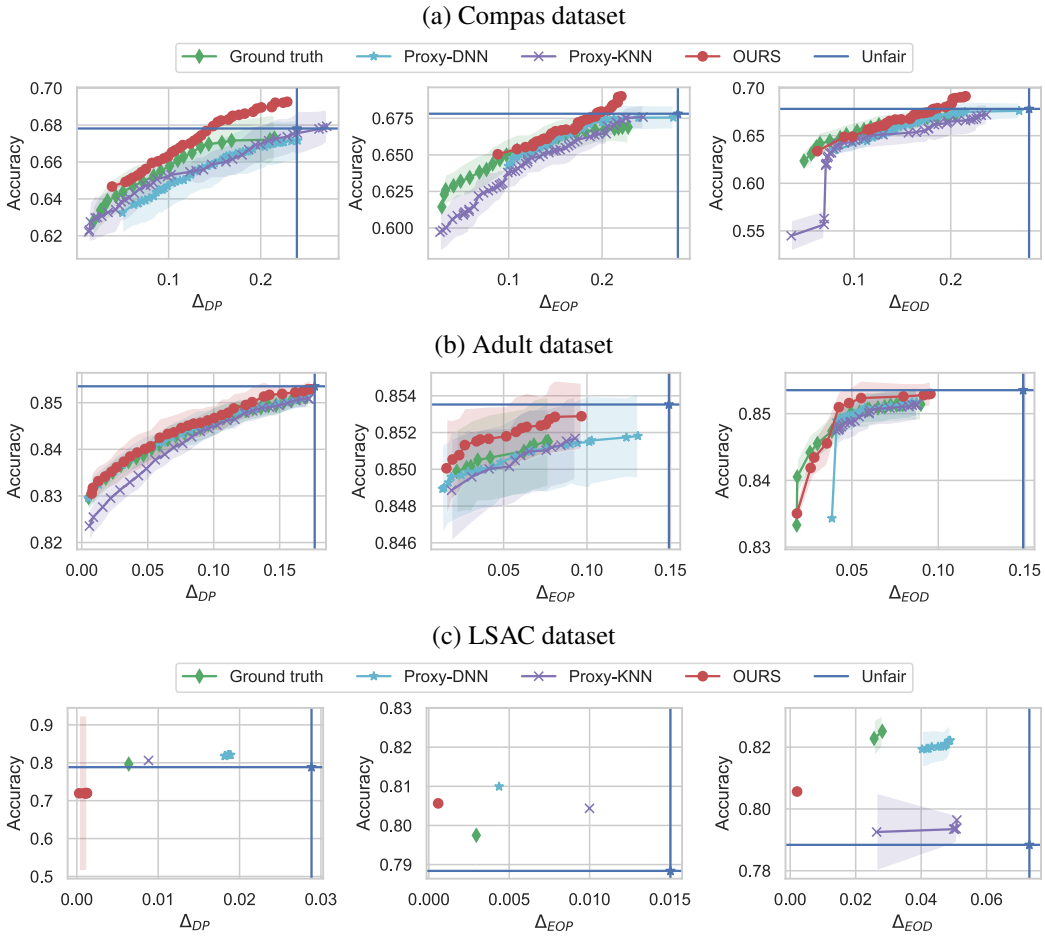


Figure 5: Accuracy-fairness trade-offs for various fairness metrics (Δ_{DP}, Δ_{EOP}, Δ_{EOD}) and proxy-sensitive attributes. Top-left is the best (Highest accuracy with the lowest unfairness). The fairness mechanism is the Exponentiated gradient with logistic regression as the base classifier on the Compas (a), Adult (b), and LSAC (c) datasets. Shaded in the figure is the standard deviation.

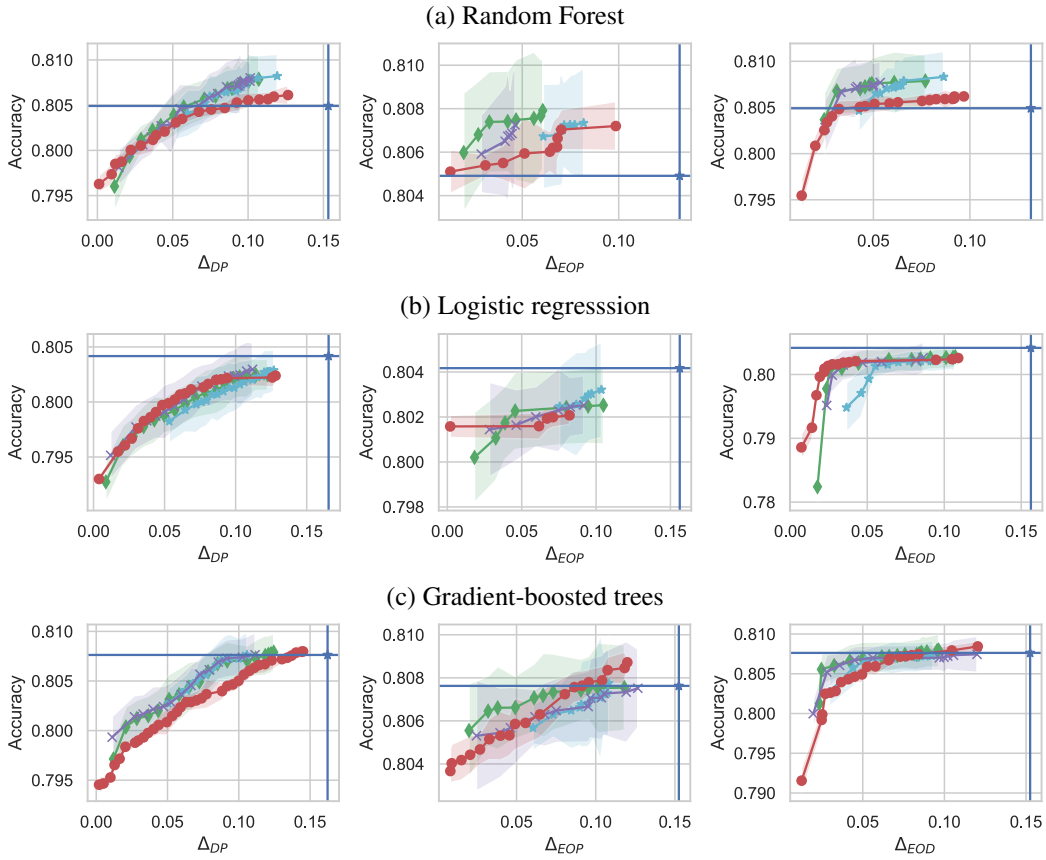


Figure 6: Exponentiated gradient with different base classifiers on the CelebA dataset: (a) random Forest, (b) logistic regression, and (c) gradient-boosted trees.

F.2 Experiments with adversarial debaising

Figure 10 shows the trade-offs for adversarial debiasing. Our methods achieve a better trade-off on the Adult datasets while for the Compas dataset, the ground-truth sensitive achieves a better trade-off. It is worth noting that adversarial debiasing is unstable to train.

G Uncertainty Estimation of Different Demographic Groups

In this paper, we showed that when the dataset does not encode enough information about the sensitive attributes, the attribute classifier suffers on average from greater uncertainty in the predictions of sensitive attributes. This encourages a choice of a higher uncertainty threshold to keep enough samples in order to maintain the accuracy, i.e. to prune out only the most uncertain samples. Figure 11 shows that the gap between demographic groups can increase as a smaller uncertainty threshold is used. This is explained by the fact that the model is more confident about samples from well-represented groups than samples from under-represented groups. While this gap between demographic groups can increase, our results show there are still enough samples from the disadvantaged group with reliable sensitive attributes. Thus, tuning the uncertainty threshold can result in a model that achieves a better trade-off between accuracy and various fairness metrics, by enforcing fairness constraints on samples with highly reliable sensitive attributes. Note that for the LSAC dataset, we observed the same trend. The average uncertainty is 0.66 and the minimum uncertainty is 0.62. We also observed that group representation remains consistent (35% difference) when using the average uncertainty value.

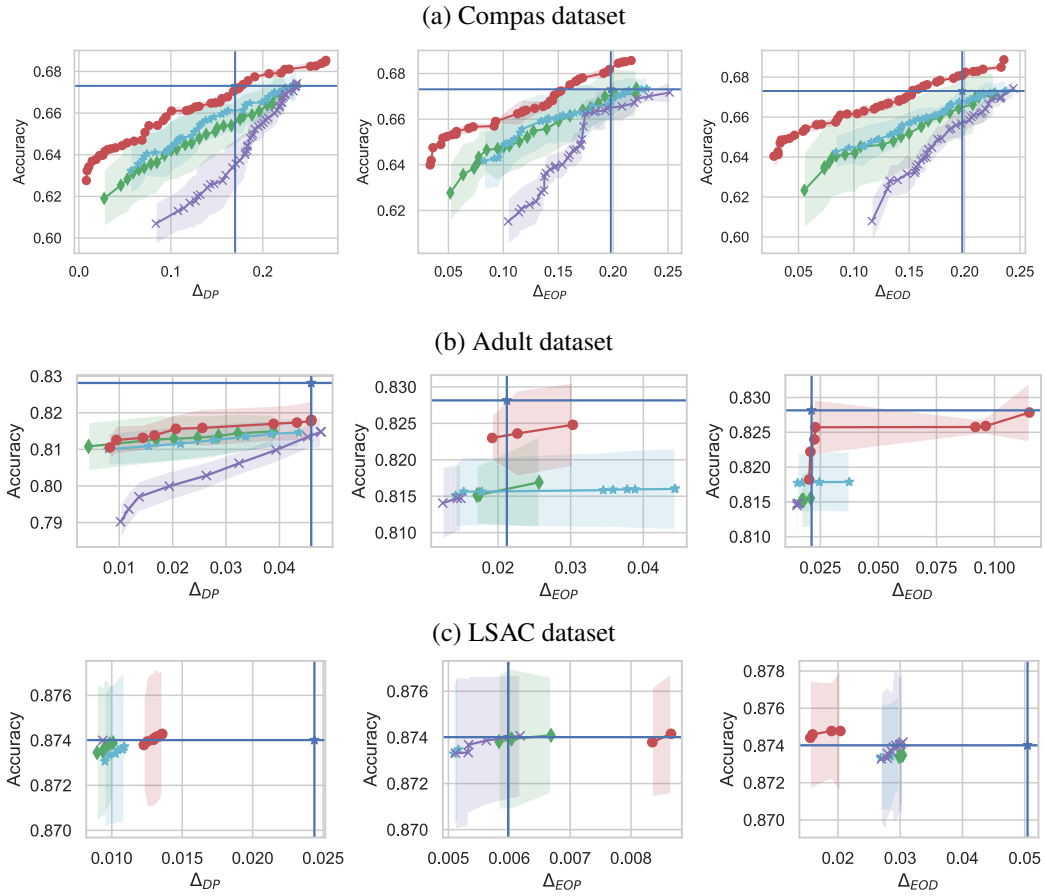


Figure 7: Accuracy-fairness tradeoffs for various fairness metrics (Δ_{DP}, Δ_{EOP}, Δ_{EOD}) and proxy sensitive attributes. The fairness mechanism used is the Exponentiated gradient with gradient-boosted trees as the base classifier on the Compas (a) and the Adult (b) datasets. Shaded in the figure is the standard deviation.

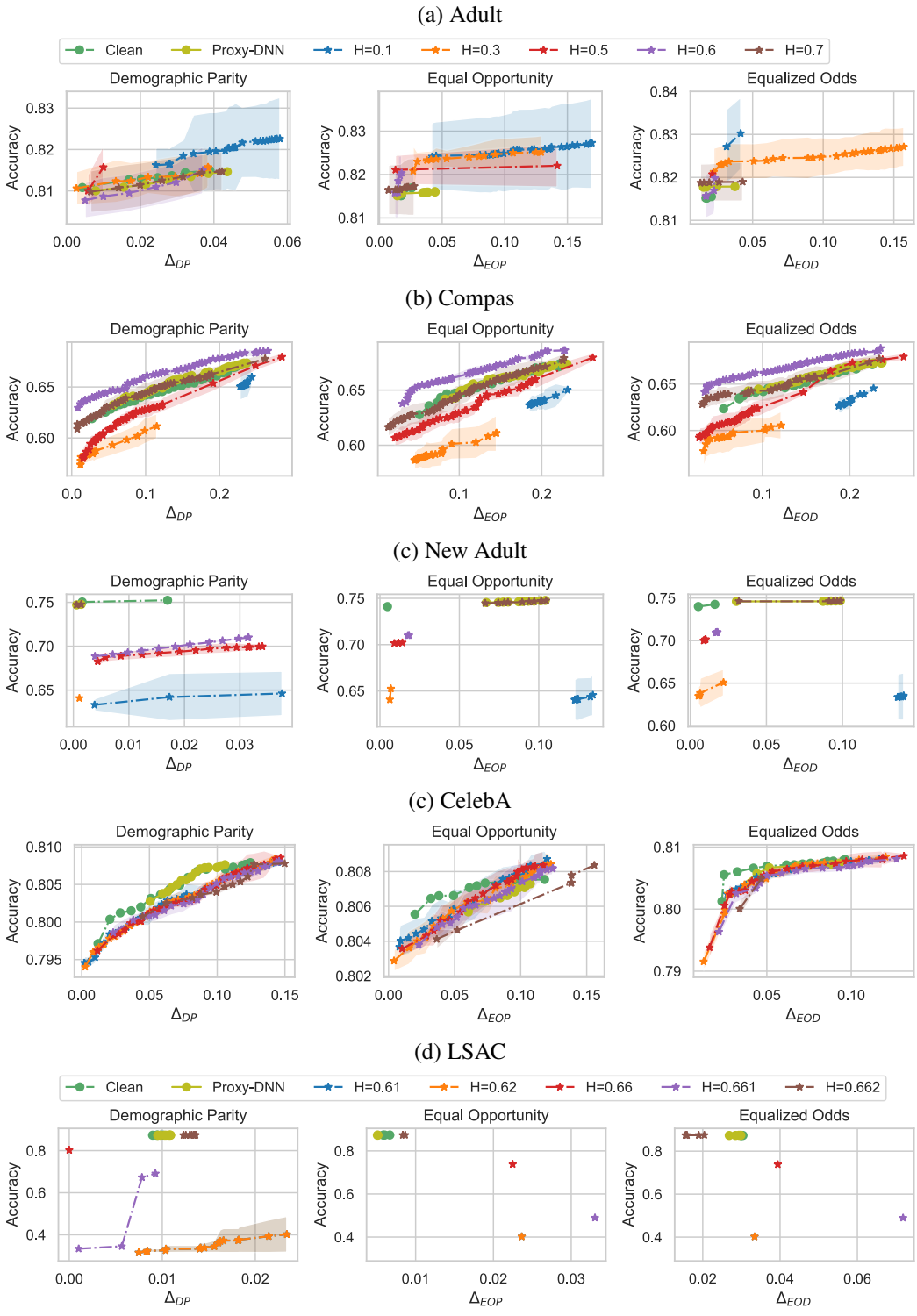


Figure 8: Exponentiated gradient with gradient-boosted trees as the base classifier. The impact of the uncertainty threshold H on the fairness-accuracy trade-off for the (a) Adult, (b) Compas, (c) New Adult, (c) CelebA dataset, and (d) LSAC datasets.

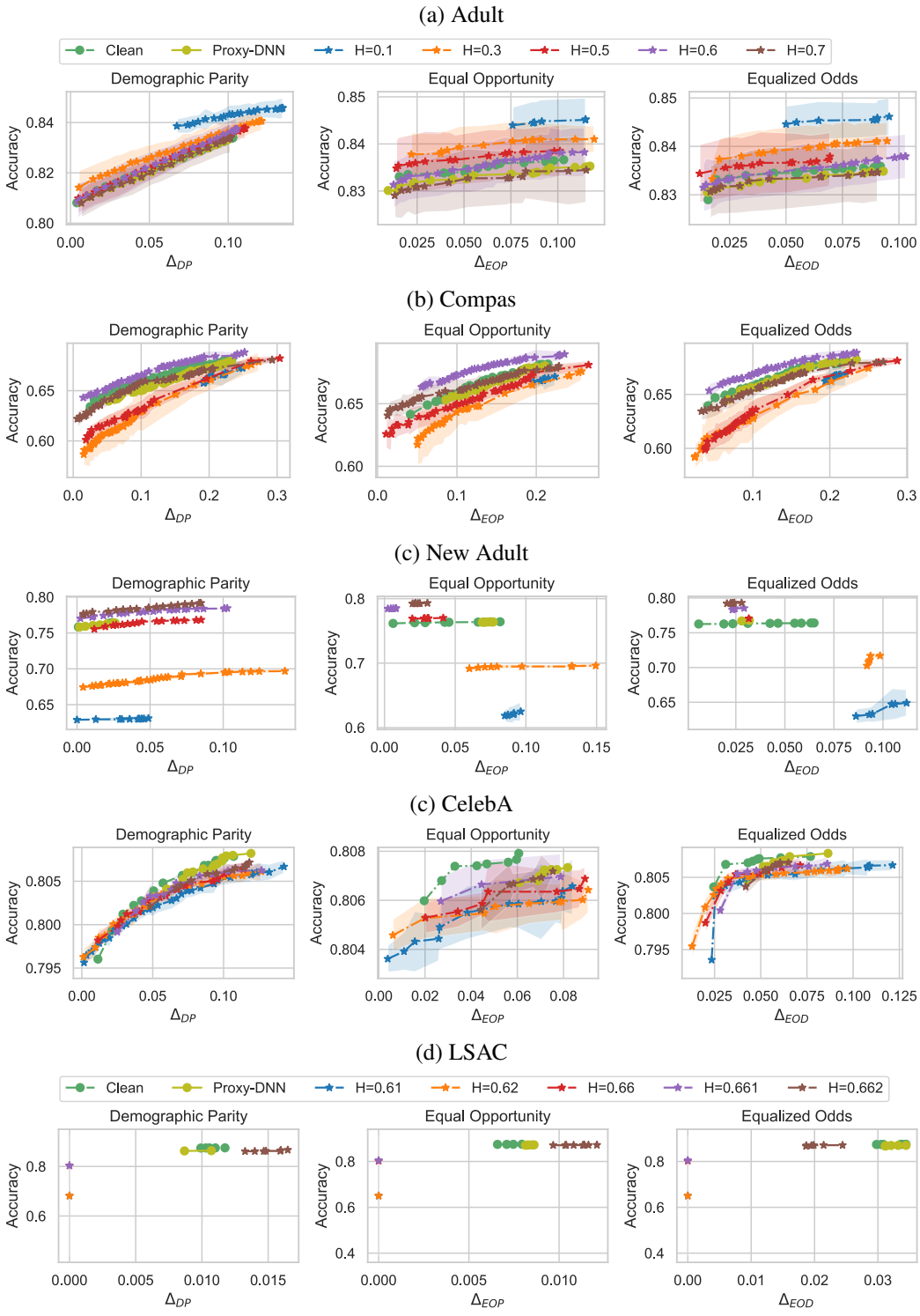


Figure 9: The impact of the uncertainty threshold H on the fairness-accuracy trade-off. For the exponentiated gradient with Random Forest as the base classifier for the (a) Adult, (b) Compas, (c) New Adult, (c) CelebA dataset, and (d) LSAC datasets.

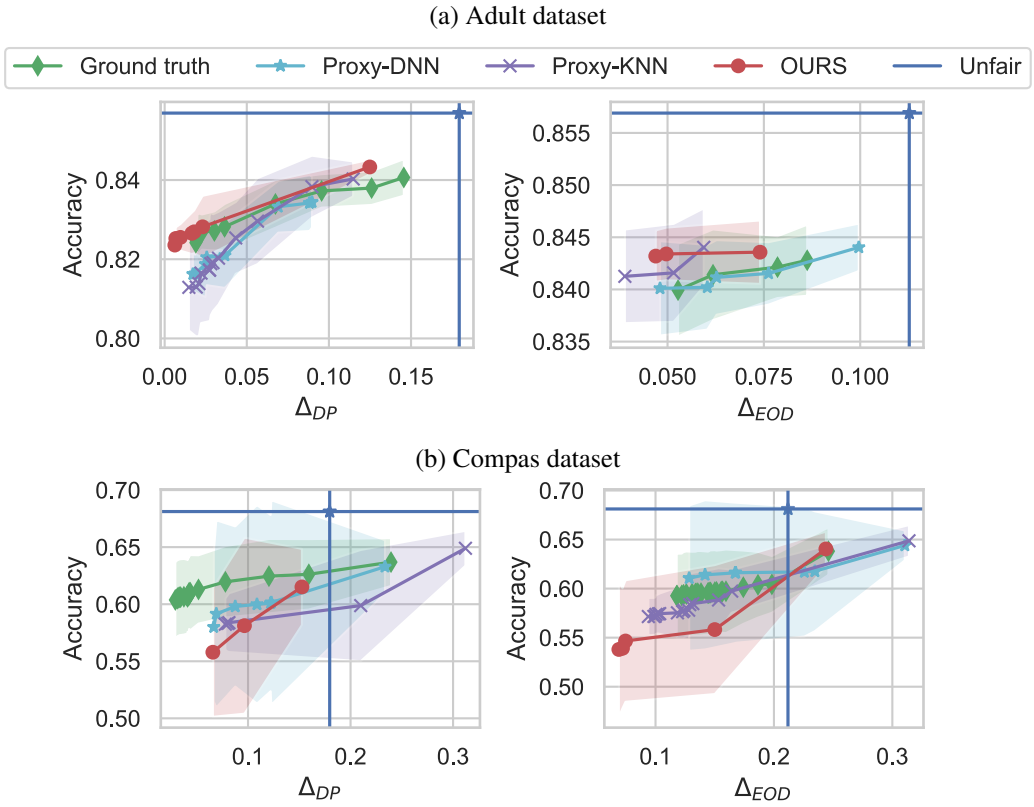


Figure 10: Adversarial debiasing. Accuracy-fairness trade-offs for various fairness metrics (Δ_{DP}, Δ_{EOP}) and proxy-sensitive attributes.

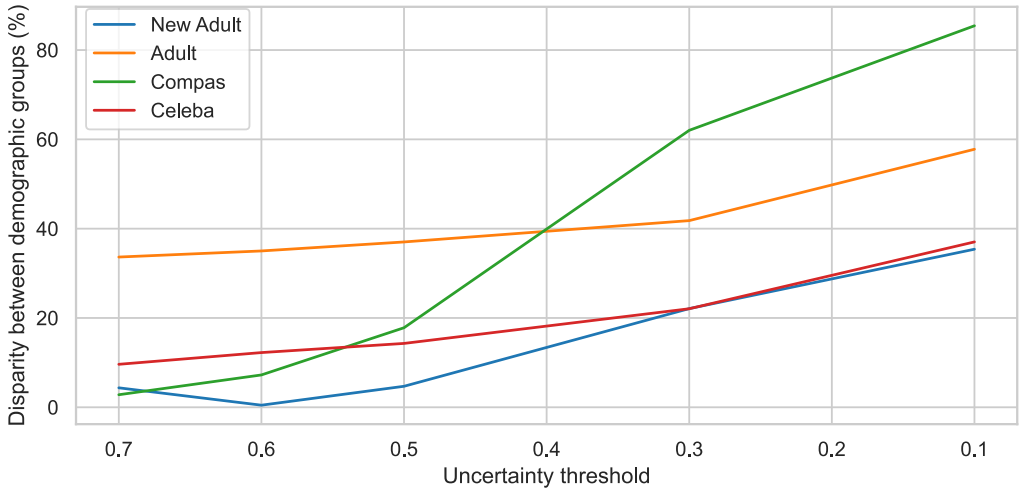


Figure 11: Demographic group representation in each dataset for different uncertainty thresholds. The gap between groups increases as the threshold becomes smaller. The plot reveals there are samples from the minority group that exhibit lower uncertainty in the prediction of their sensitive attributes.