Appendix A. Additional Details on Experiment Setup

As we directly borrow the experiment setup from Han et al. (2023), we redirect the reader to their work and the FFB benchmark code [2] for details on the underlying setup. In this section, we briefly mention the datasets and algorithms used in the benchmark, and the new additions and changes we made to their setup.

A.1. Datasets

We use 7 different tabular datasets for our experiments. This includes the Adult dataset (Becker and Kohavi, 1996), COMPAS dataset (Larson et al., 2016), German dataset (Hofmann, 1994), Bank Marketing dataset (Moro et al., 2014), KDD Census dataset (cen, 2000), and ACS dataset with tasks Income and Employment (Ding et al., 2021). We use the sensitive attribute *Race* for all datasets, except the Bank Marketing dataset and the German dataset, where we use *Age* as the sensitive attribute.

A.2. Bias Mitigation Algorithms

We use 7 different bias mitigation algorithms in our setup. This includes DiffDP, DiffEOdd, DiffEOpp, PRemover (Kamishima et al., 2012), HSIC (Baharlouei et al.; Gretton et al., 2005; Li et al., 2022), AdvDebias (Adel et al., 2019; Beutel et al., 2017; Edwards and Storkey, 2016; Louppe et al., 2017; Zhang et al., 2018), and LAFTR (Madras et al., 2018).

A.3. Hyperparameters

We use the Adam optimizer, with no weight decay and a step learning rate scheduler for training. We train the model for 150 epochs and record the fairness and accuracy scores at the final epoch.

We use three different values of the control parameter for each algorithm, as defined in Table 3.

Algorithm	**Control Hyperparameter**
DiffDP	0.2, 1.0, 1.8
DiffEOdd	0.2, 1.0, 1.8
DiffEOdd	0.2, 1.0, 1.8
PRemover	0.05, 0.25, 0.45
HSIC	50, 250, 450
AdvDebias	0.2, 1.0, 1.8
LAFTR	0.1, 0.5, 4.0

Table 3: Control hyperparameters.

We use seven different hyperparameter settings for each dataset, as defined in Table 4.

[2] https://github.com/ahxt/fair_fairness_benchmark

Adult and Bank Marketing			COMPAS and German		
Batch Size	**Learning Rate**	**MLP Layers**	**Batch Size**	**Learning Rate**	**MLP Layers**
1024	0.01	512,256	32	0.01	512,256
1024	0.01	64	32	0.01	64
1024	0.01	512,256,256,64	32	0.01	512,256,256,64
128	0.01	512,256	8	0.01	512,256
128	0.001	512,256	8	0.001	512,256
4096	0.01	512,256	128	0.01	512,256
4096	0.1	512,256	128	0.1	512,256
KDD and ACS					
Batch Size	**Learning Rate**	**MLP Layers**			
4096	0.01	512,256			
4096	0.01	64			
4096	0.01	512,256,256,64			
512	0.01	512,256			
512	0.001	512,256			
8192	0.01	512,256			
8192	0.1	512,256			

Table 4: Hyperparameters.

Appendix B. Additional Results for Trends Under Changing Hyperparameters

We present additional results for comparing trends under different hyperparameters in the Adult dataset for fairness definitions of equalized odds (Figure 5) and equal opportunity (Figure 6). We also present additional results for comparing trends in other datasets like Bank Marketing dataset (Figure 7), COMPAS dataset (Figure 8), German dataset (Figure 9), KDDCensus dataset (Figure 10), ACS-Income dataset (Figure 11) and ACS-Employment dataset (Figure 12).

Appendix C. Additional Results for Changing Trends Across Datasets

We present additional results for comparing trends across multiple datasets, under fairness definition as equalized odds (Figure 13) and equal opportunity (Figure 14). Similar to the observations in the main paper, we find distinct trends across different datasets and no clear single bias mitigation algorithm that excels across all datasets.

Appendix D. Additional Results at Pareto Front

We present additional results on the pareto front for various algorithms and datasets, under fairness definition as equalized odds (Figure 15) and equal opportunity (Figure 16). Similar to the trends seen in the main paper, we find many different algorithms provide competitive tradeoffs when allowed to perform appropriate hyperparameter optimization.

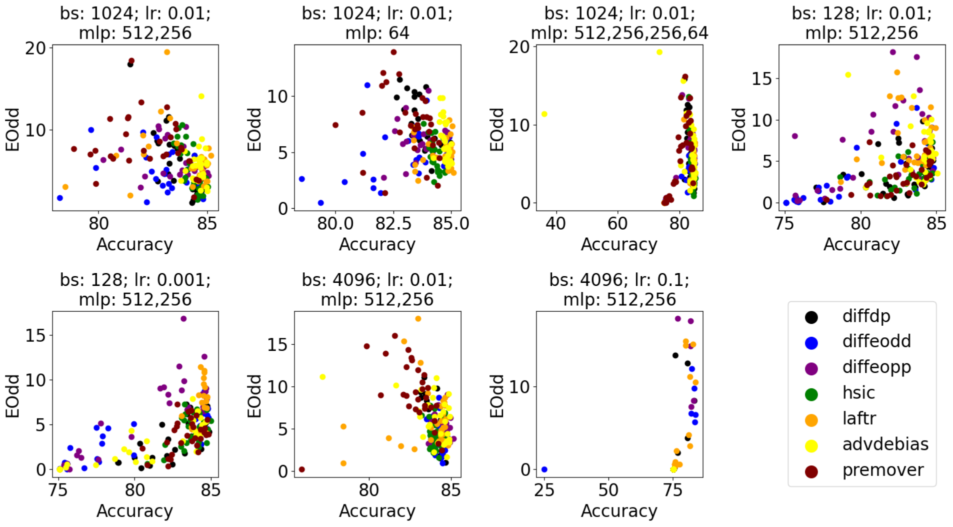

Figure 5: Fairness-utility (equalized odds-accuracy) tradeoff across various settings for the Adult dataset. Each graph represents a different combination of hyperparameters, and each dot in the graph represents a separate training run. Multiple dots for the same mitigation algorithm in the same graph represent runs with changing random seeds and control parameters.

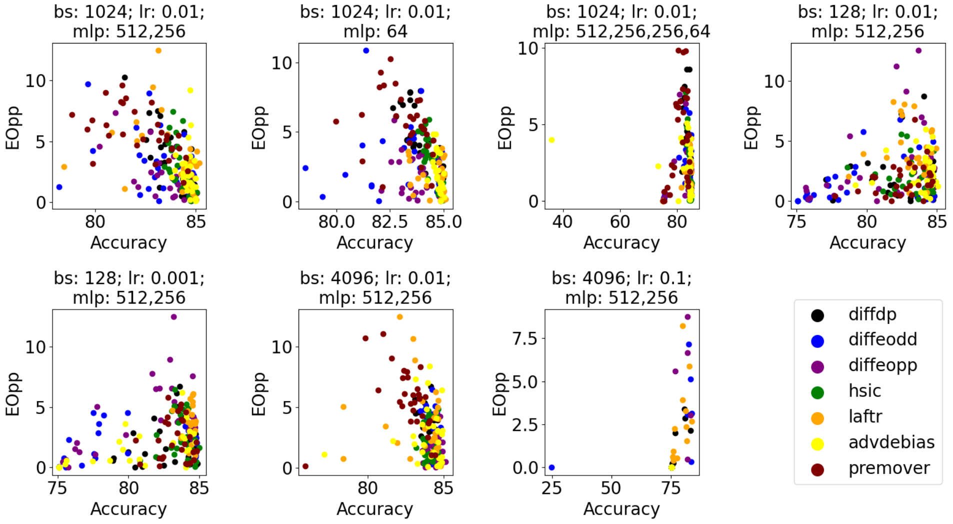

Figure 6: Fairness-utility (equal opportunity-accuracy) tradeoff across various settings for the Adult dataset. Each graph represents a different combination of hyperparameters, and each dot in the graph represents a separate training run. Multiple dots for the same mitigation algorithm in the same graph represent runs with changing random seeds and control parameters.

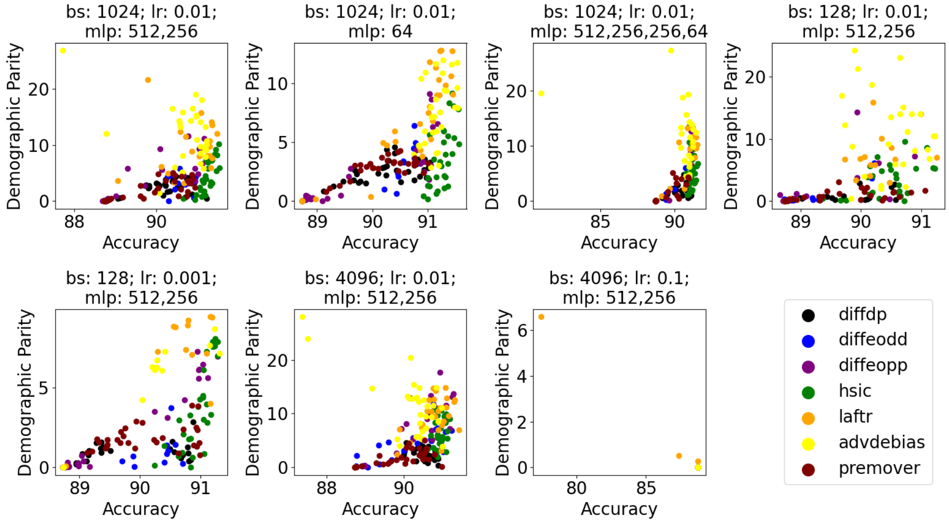

Figure 7: Fairness-utility (demographic parity-accuracy) tradeoff across various settings for the Bank Marketing dataset. Each graph represents a different combination of hyperparameters, and each dot in the graph represents a separate training run. Multiple dots for the same mitigation algorithm in the same graph represent runs with changing random seeds and control parameters.

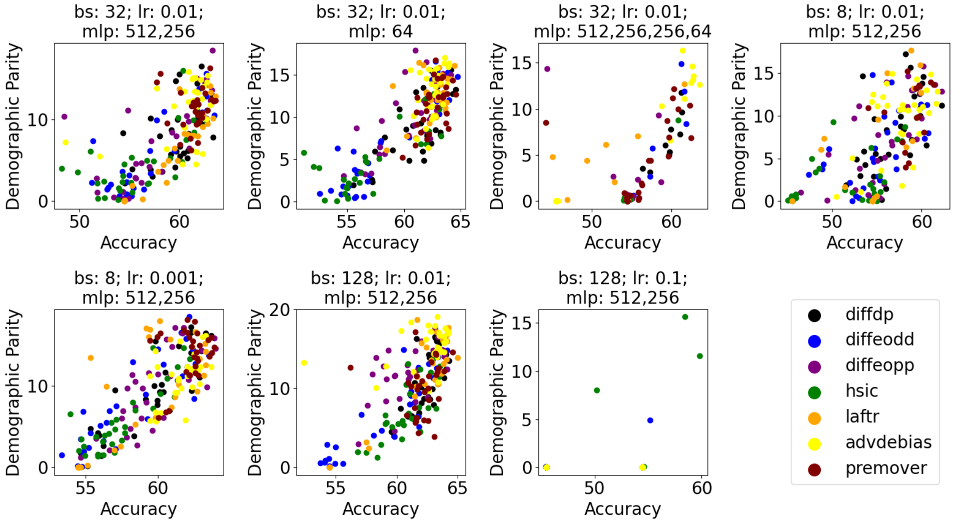

Figure 8: Fairness-utility (demographic parity-accuracy) tradeoff across various settings for the COMPAS dataset. Each graph represents a different combination of hyperparameters, and each dot in the graph represents a separate training run. Multiple dots for the same mitigation algorithm in the same graph represent runs with changing random seeds and control parameters.

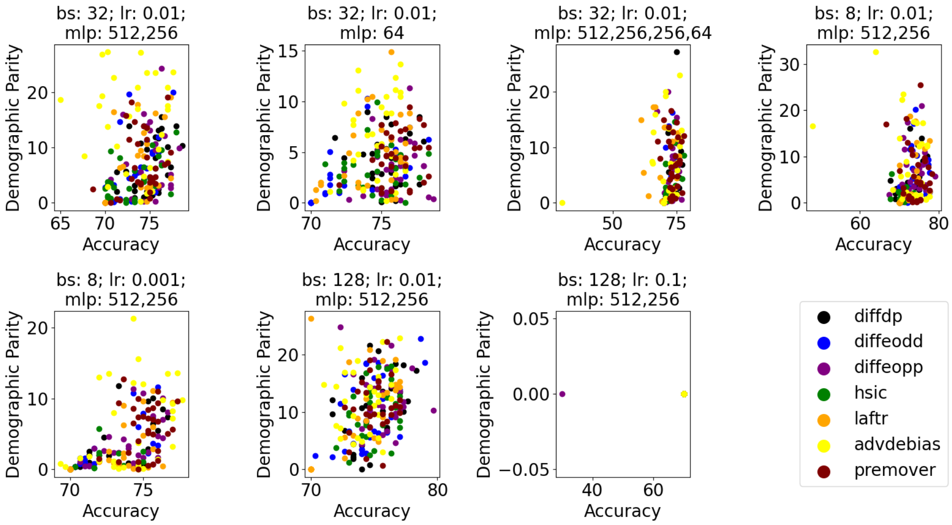

Figure 9: Fairness-utility (demographic parity-accuracy) tradeoff across various settings for the German dataset. Each graph represents a different combination of hyperparameters, and each dot in the graph represents a separate training run. Multiple dots for the same mitigation algorithm in the same graph represent runs with changing random seeds and control parameters.

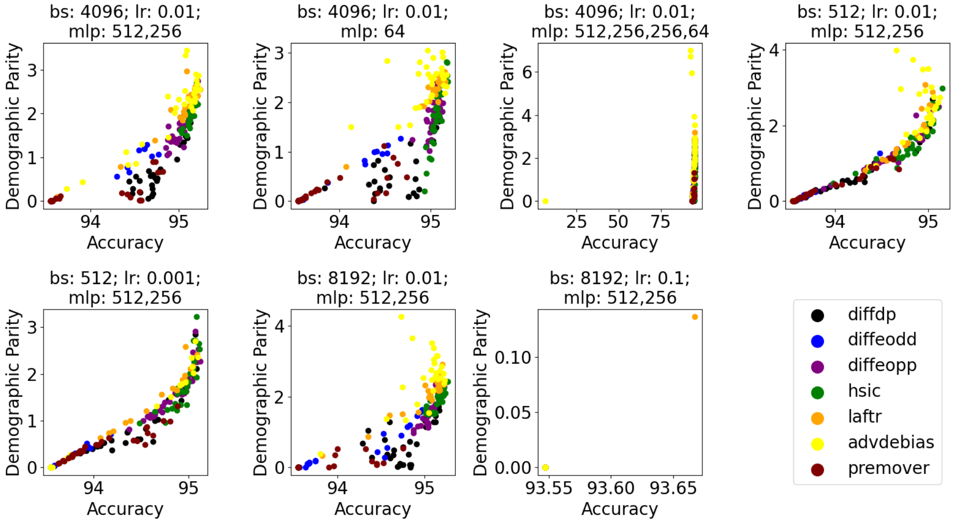

Figure 10: Fairness-utility (demographic parity-accuracy) tradeoff across various settings for the KDDCensus dataset. Each graph represents a different combination of hyperparameters, and each dot in the graph represents a separate training run. Multiple dots for the same mitigation algorithm in the same graph represent runs with changing random seeds and control parameters.

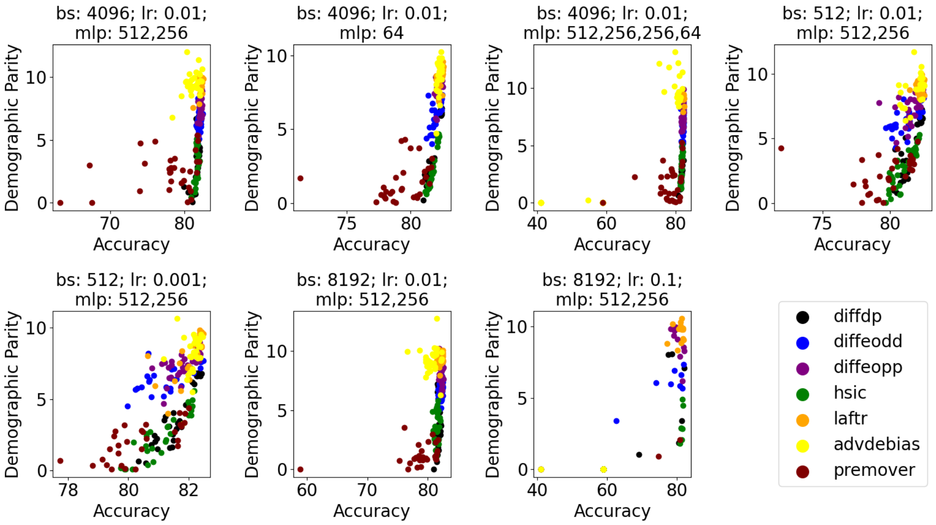

Figure 11: Fairness-utility (demographic parity-accuracy) tradeoff across various settings for the ACS-Income dataset. Each graph represents a different combination of hyperparameters, and each dot in the graph represents a separate training run. Multiple dots for the same mitigation algorithm in the same graph represent runs with changing random seeds and control parameters.

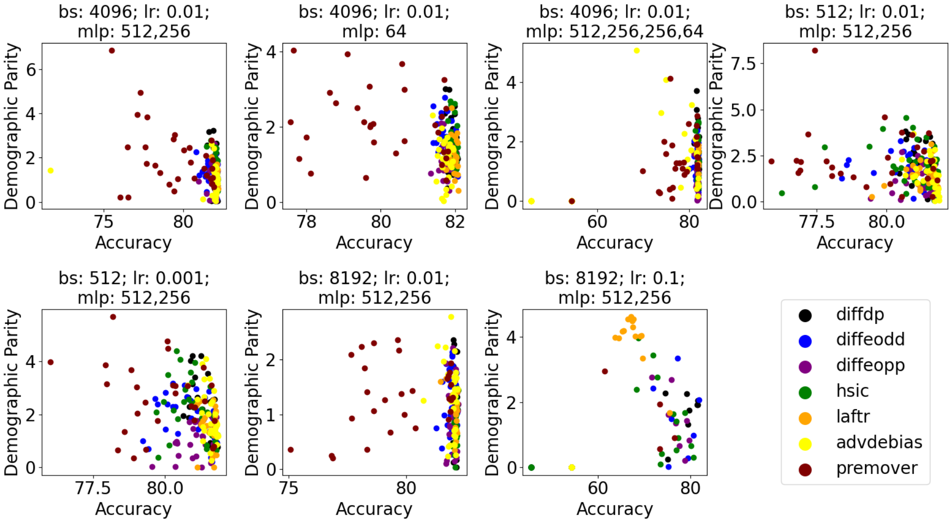

Figure 12: Fairness-utility (demographic parity-accuracy) tradeoff across various settings for the ACS-Employment dataset. Each graph represents a different combination of hyperparameters, and each dot in the graph represents a separate training run. Multiple dots for the same mitigation algorithm in the same graph represent runs with changing random seeds and control parameters.

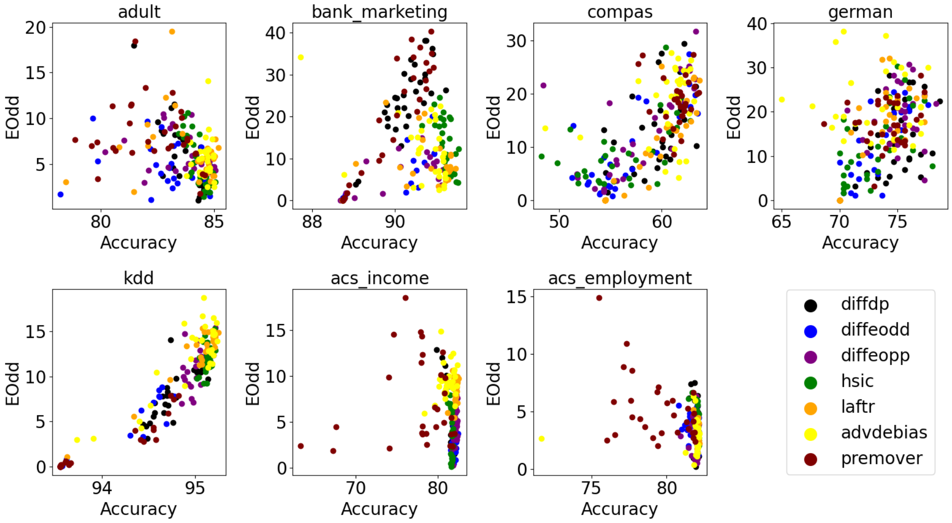

Figure 13: Fairness-utility (equalized odds-accuracy) tradeoff across various datasets, under their default hyperparameters. Each dot in the graph represents a separate training run with changing random seeds and control parameters.

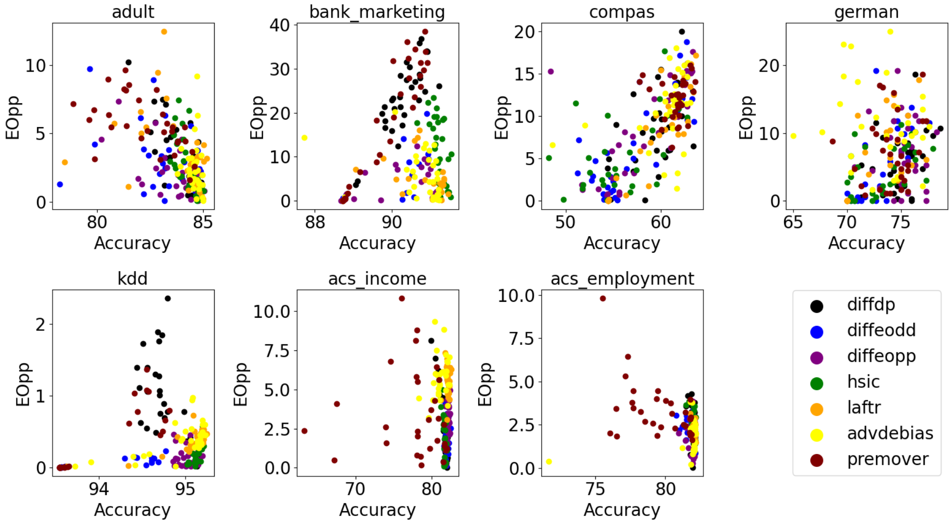

Figure 14: Fairness-utility (equal opportunity-accuracy) tradeoff across various datasets, under their default hyperparameters. Each dot in the graph represents a separate training run with changing random seeds and control parameters.

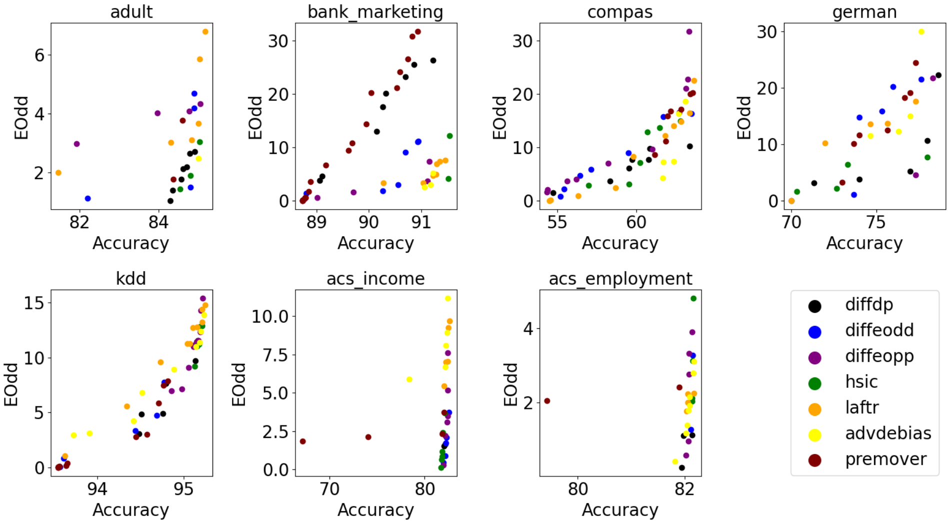

Figure 15: Pareto front of the fairness-utility (equalized odds-accuracy) tradeoff across various datasets. Each dot in the graph represents a separate training run on the pareto front with changing hyperparameters, random seeds and control parameters.

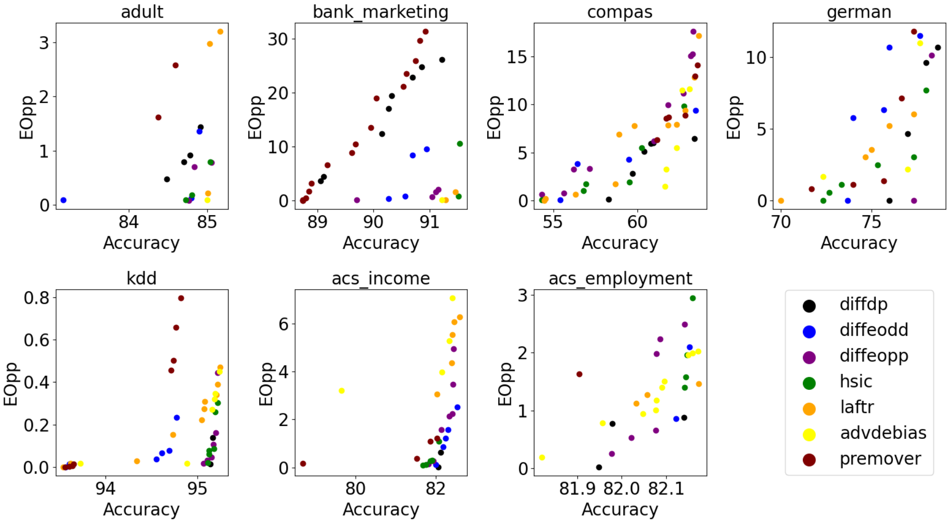

Figure 16: Pareto front of the fairness-utility (equal opportunity-accuracy) tradeoff across various datasets. Each dot in the graph represents a separate training run on the pareto front with changing hyperparameters, random seeds and control parameters.