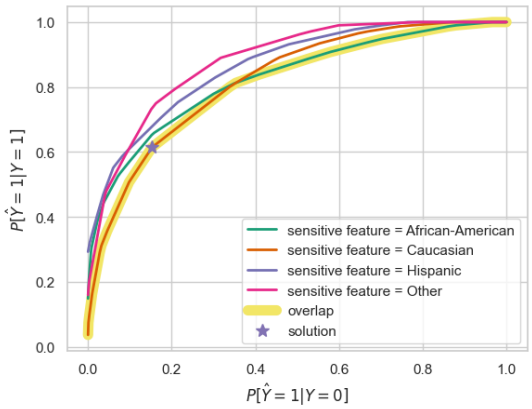
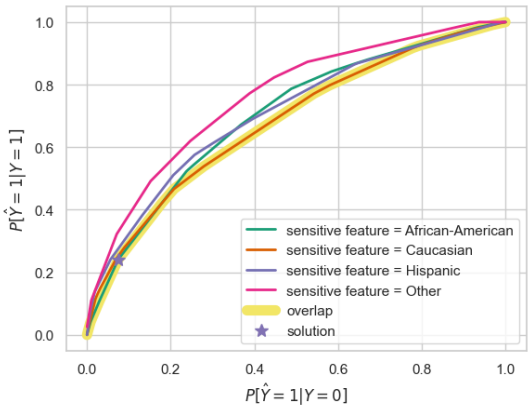
A Miscellaneous

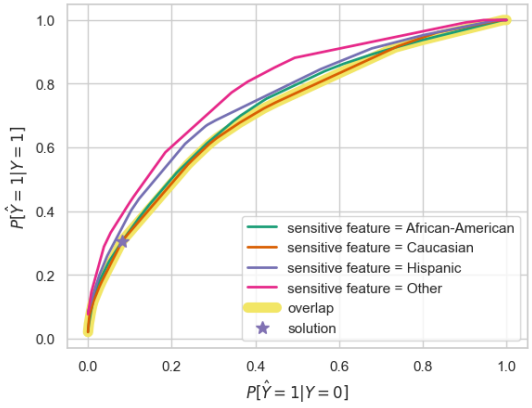


(a) Threshold Optimization - when we trained with y_{bias} and tested with y_{bias}. The overlap rests on the other groups. The optimized threshold lies between 0.60 when the predicted label and the actual label are positive (1) and at 0.19 when the predicted label and actual label have opposite classes (positive (1) and negative (0)).

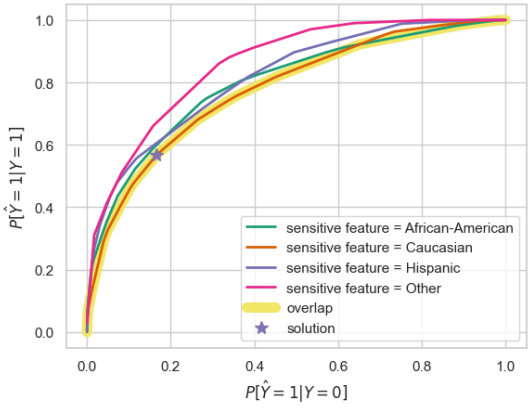


(b) Threshold Optimization - when we trained with y_{bias} and tested with a y_{truth}. The overlap rests on the other groups. The optimized threshold lies between 0.25 when the predicted label and the actual label are positive (1) and at 0.10 when the predicted label and actual label have opposite classes (positive (1) and negative (0)).

Fig. 6. Mitigation: Optimized Threshold - trained with y_{bias}



(a) Threshold Optimization - when we trained with y_{truth} and tested with y_{truth}. The overlap rests on the other groups. The optimized threshold lies between 0.30 when the predicted label and the actual label are positive (1) and at 0.10 when the predicted label and actual label have opposite classes (positive (1) and negative (0)).



(b) Threshold Optimization - when we trained with y_{truth} and tested with a y_{bias}. The overlap rests on the other groups. The optimized threshold lies close to 0.60 when the predicted label and the actual label are positive (1) and at 0.19 when the predicted label and actual label have opposite classes (positive (1) and negative (0)).

Fig. 7. Mitigation: Optimized Threshold - trained with y_{truth}

Table 8. Case II: Performance Assessment After mitigation

Race	**Selection rate**		**TPR**		**FPR**		**PPV**		**NPV**	
Train with y_{truth}	*Real*	*Observed*	*Real*	*Observed*	*Real*	*Observed*	*Real*	*Observed*	*Real*	*Observed*
African-American	0.1605	0.1605	0.2674	0.2695	0.0851	0.0303	0.6887	0.9139	0.6392	0.5266
Caucasian	0.1451	0.1451	0.2634	0.3721	0.0967	0.0305	0.5269	0.8602	0.7500	0.7537
Hispanic	0.0798	0.0798	0.0952	0.2632	0.0744	0.0240	0.3077	0.7692	0.7467	0.8133
Others	0.1353	0.1353	0.3000	0.3125	0.0645	0.0792	0.6667	0.5556	0.7565	0.8087

Table 9. Case II: Fairness Metrics after Mitigation

Race	**Independence**		**Separation**		**Sufficiency**	
Train with y_{truth}	*Real*	*Observed*	*Real*	*Observed*	*Real*	*Observed*
African-American	0.0807	0.0852	0.1829	0.1829	0.4885	0.4885
Caucasian	0.0653	0.0653	0.1905	0.1905	0.2726	0.2726
Hispanic	0.0807	0.0852	0.2147	0.2147	0.4885	0.4885
Others	0.0555	0.0555	0.2144	0.2147	0.3688	0.3688

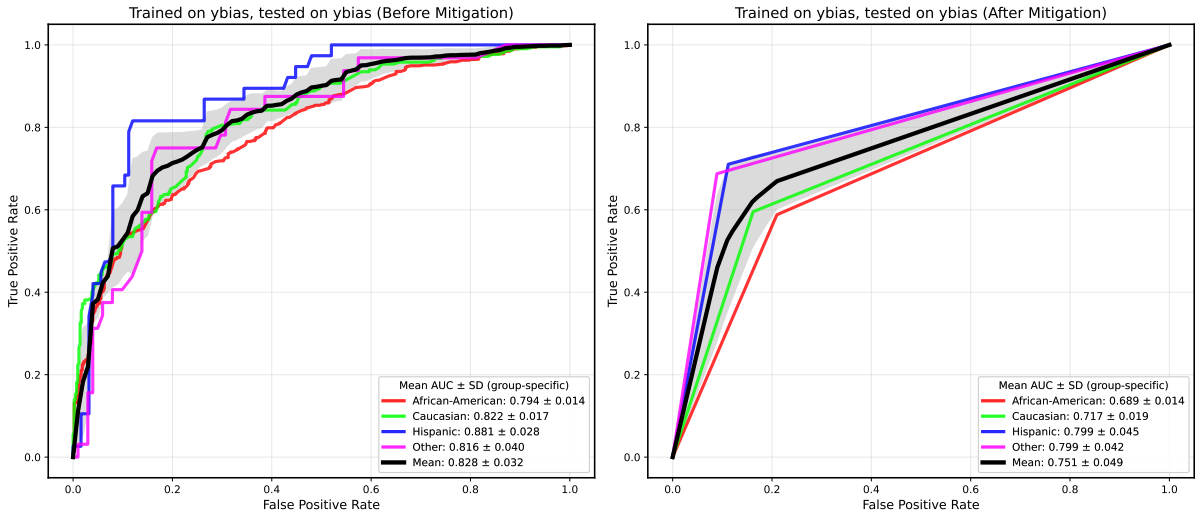


Fig. 8. The Corresponding ROC plot of the Threshold Optimization: trained on y_{bias}, tested on y_{bias}. The plot shows the before and after mitigation effect across groups. Observe the after-mitigation mean ROC $\pm$ SD.

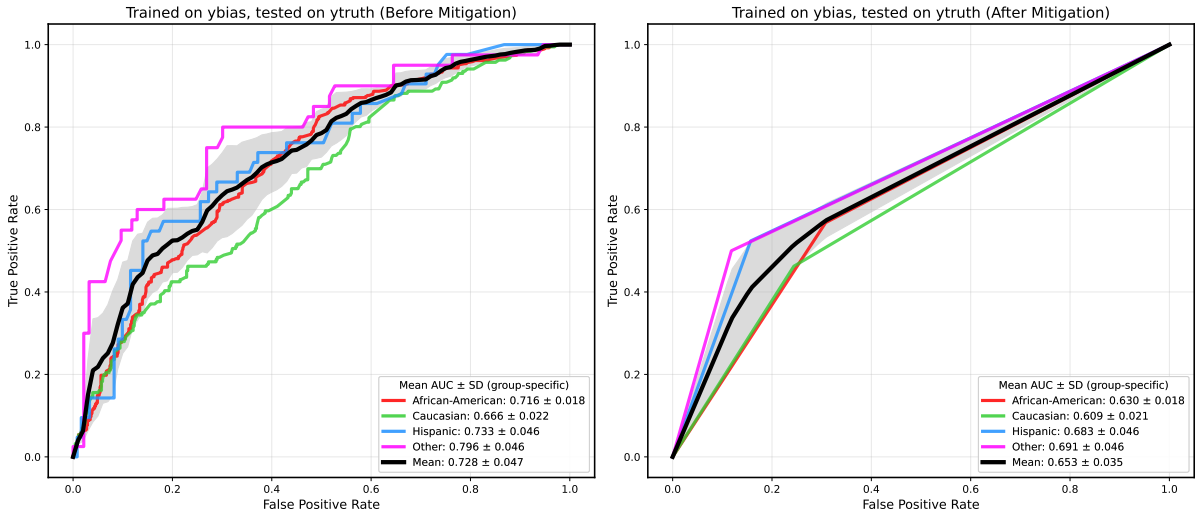


Fig. 9. The Corresponding ROC plot of the Threshold Optimization: trained on y_{bias}, tested on y_{truth}. The plot shows the before and after mitigation effect across groups. The plot shows the before and after mitigation effect across groups. Observe the after-mitigation mean ROC ± SD.

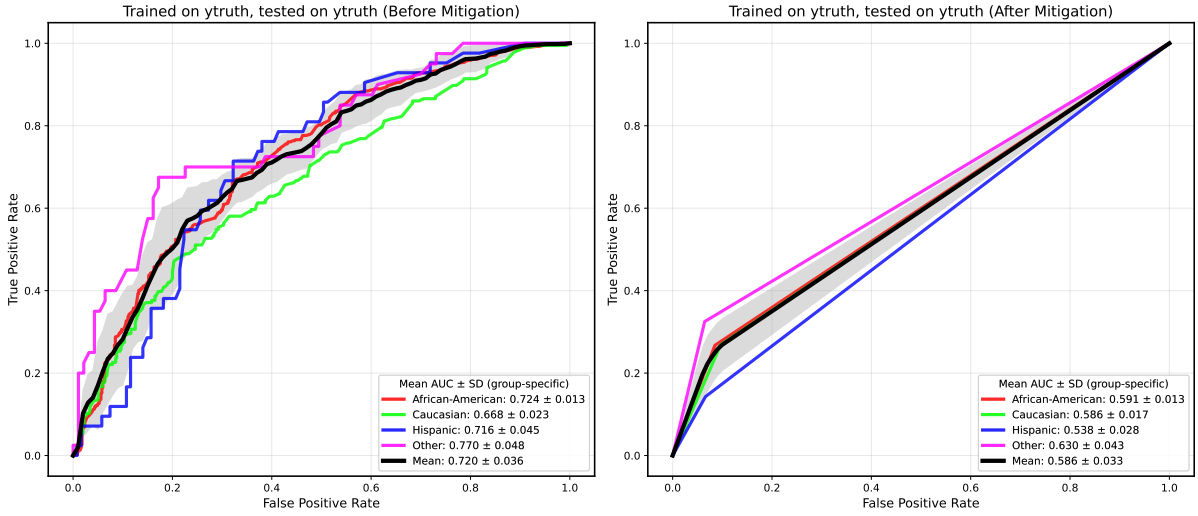


Fig. 10. The Corresponding ROC plot of the Threshold Optimization: trained on y_{truth}, tested on y_{truth}. The plot shows the before and after mitigation effect across groups. The plot shows the before and after mitigation effect across groups. Observe the after-mitigation mean ROC ± SD.

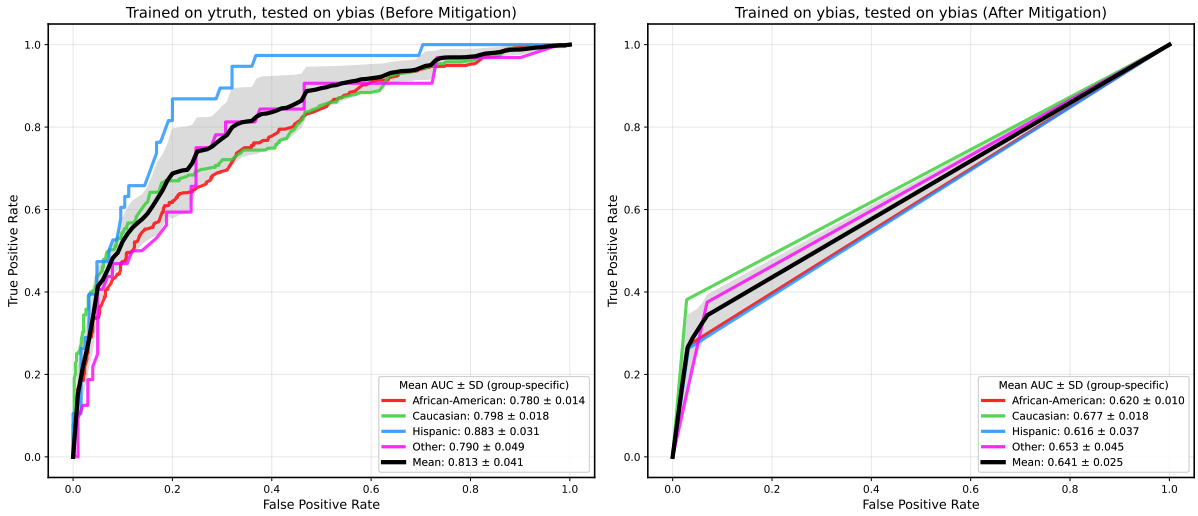


Fig. 11. The Corresponding ROC plot of the Threshold Optimization: trained on y_{truth}, tested on y_{bias}. The plot shows the before and after mitigation effect across groups. The plot shows the before and after mitigation effect across groups. Observe the after-mitigation mean ROC $\pm$ SD.