Challenges in Reducing Bias Using Post-Processing Fairness for Breast Cancer Stage Classification with Deep Learning

Armin Soltan, Peter Washington.

Peter Washington, University of Hawaii at Manoa

MS Plan B / ICS 699, Spring 2024. Published at Algorithms Journal from MDPI

Introduction

Cancer is the second leading cause of mortality worldwide. Breast cancer, lung cancer, and colorectal cancer account for 51% of all new diagnoses among women. Breast cancer has the highest death rate at 32%. However, this death rate is not consistent across different demographic groups. For example, the death rate for Black women is 41% higher than for White women. Three broad classes of algorithms have been investigated to mitigate bias in algorithmic fairness: pre-processing, in-processing, and post-processing. Pre-processing involves changing the data, such as by generative data augmentation, to create equal amounts of data for each demographic group prior to training the model. In-processing methods change the learning algorithm's optimization objective function to enforce a reduction in bias during the training process. These two categories of techniques can function well if modifications to the underlying data or training process are allowed. The final category of methods, post-processing, is applied after the model has been trained, using a separate set of data that was not used during the training phase. Such "black box" approaches are ideal when determining whether modifying the original AI model is impossible or infeasible. In this work, we explore the utility of applying post-processing fairness adjustments to breast cancer stage classification using medical imaging data, testing whether standard post-processing methods adapted to the multi-class setting can mitigate bias in these models.

Motivation

Breast cancer is the most common cancer affecting women globally. Despite the significant impact of deep learning models on breast cancer diagnosis and treatment, achieving fairness or equitable outcomes across diverse populations remain a challenge when some demographic groups are underrepresented in the training data. We quantified the bias of models trained to predict breast cancer stage. The majority of data (over 70%) were from White patients. We found that prior to post-processing adjustments, all deep learning models we trained consistently performed better for White patients than for non-White patients. After model calibration, we observed mixed results, with only some models demonstrating improved performance. This work provides a case study of bias in breast cancer medical imaging models and highlights the challenges in using post-processing to achieve fairness.

Methodology

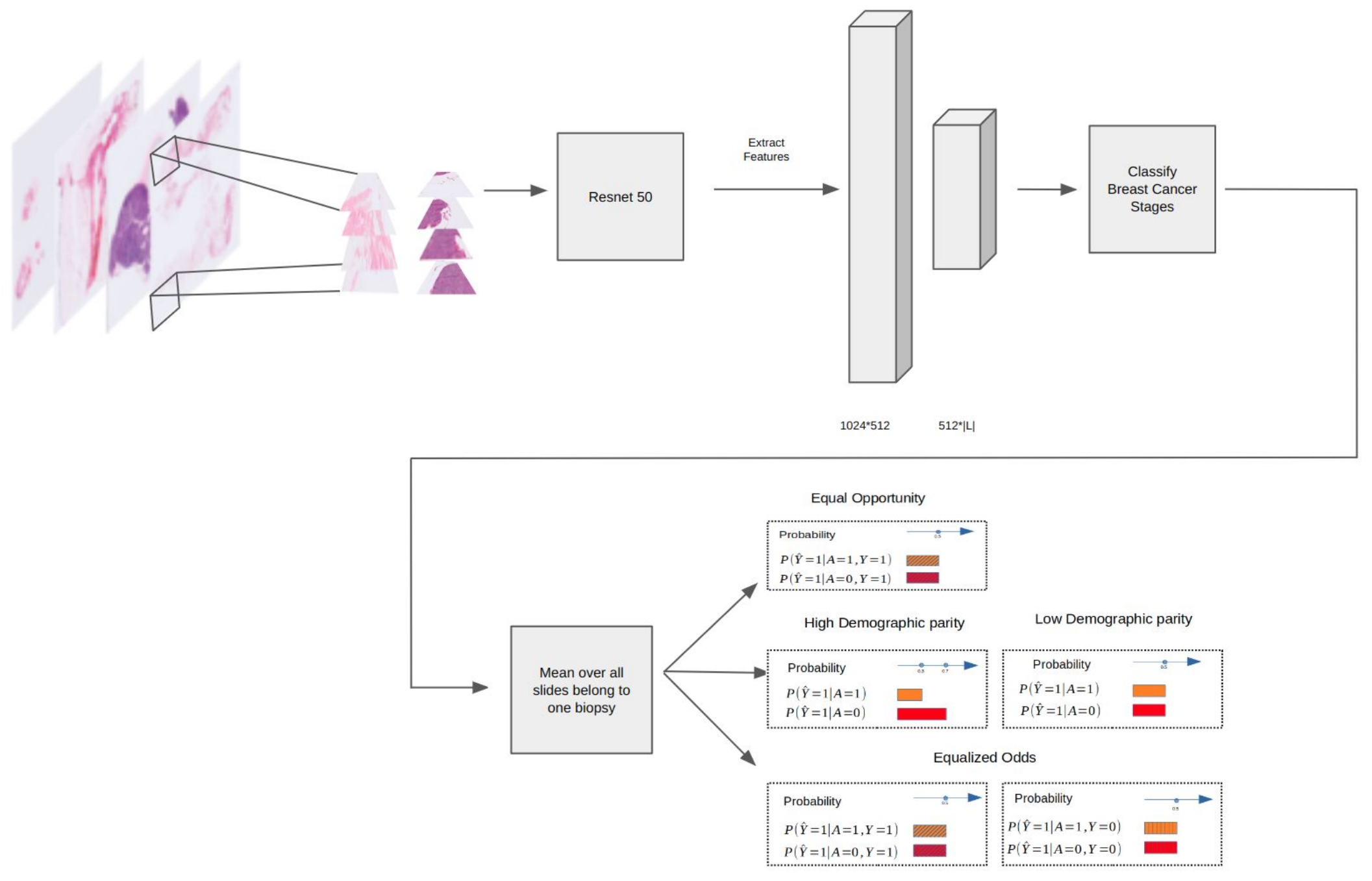
We build our fairness adjustment method upon previous post-processing algorithmic fairness work. Hardt et al. propose a method that helps to adjust the model's outputs to ensure fairness when there are only two possible outcomes. Putzel et al. suggest a way to adapt this method for situations with more than two outcomes, such as the breast cancer stage classification task that we study here. To mitigate the issue of sparse samples for some groups, as is the case with our dataset, we introduce a minor adjustment, an epsilon term, to the TPR and FPR calculations to avoid division errors. By analyzing predicted and true labels alongside sensitive attributes such as race, we engineer 'adjusted' predictions that meet predefined fairness criteria. The resulting predictors aim to balance false positive and true positive rates (for equalized odds) or synchronize true positive rates (for equal opportunity) to ensure fairness across different demographics. We leverage ROC curves to discern optimal fairness thresholds. Aligning ROC curves across groups leads to predictors that fulfill equalized odds, whereas mismatches may necessitate varying thresholds or probabilistic adjustments to achieve fair treatment. We identify optimal predictors by analyzing the intersections of group-specific convex hulls formed from these ROC curves. We manipulate conditional probabilities within the protected attribute conditional probability matrices through linear programming, optimizing against a fairness-oriented loss function. This process also incorporates an element of flexibility, allowing the loss function to penalize inaccuracies differently based on protected group membership. Our fair predictors ensure a balanced representation of demographic groups by equalizing various fairness metrics. We explore two different multi-class fairness criteria, although the method could generalize to other fairness metrics as well. We aim to minimize the same expected loss function for multiple classification that was used by Putzel et al. [19]:

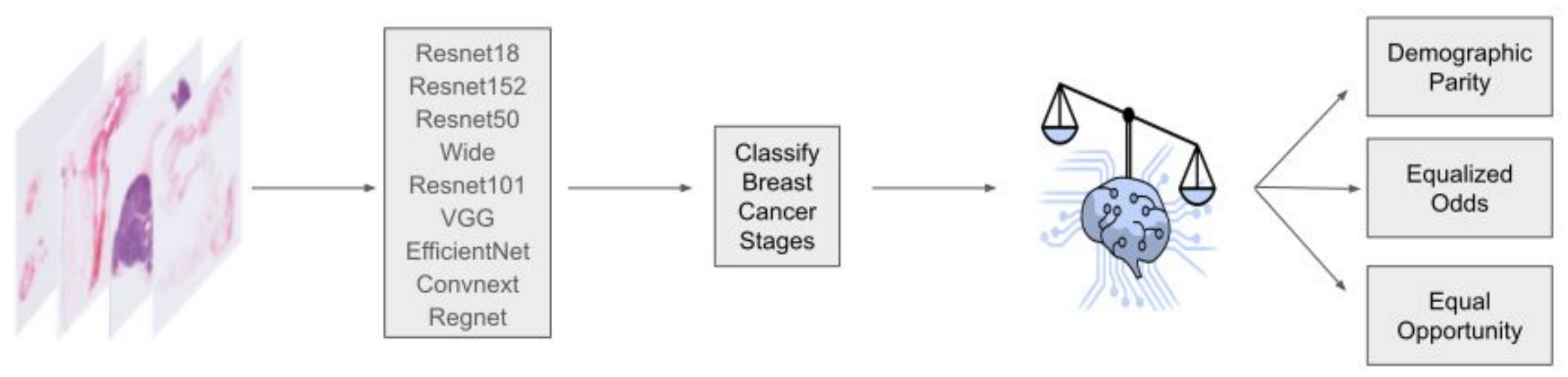
$$E[l(\hat{y}^{adj},y)] = \sum_{\alpha\in A}\sum_{i=1}^{|C|}\sum_{j\neq i} W_{ij}^{\alpha} Pr(A=\alpha, Y=j) l(i,j,\alpha)$$

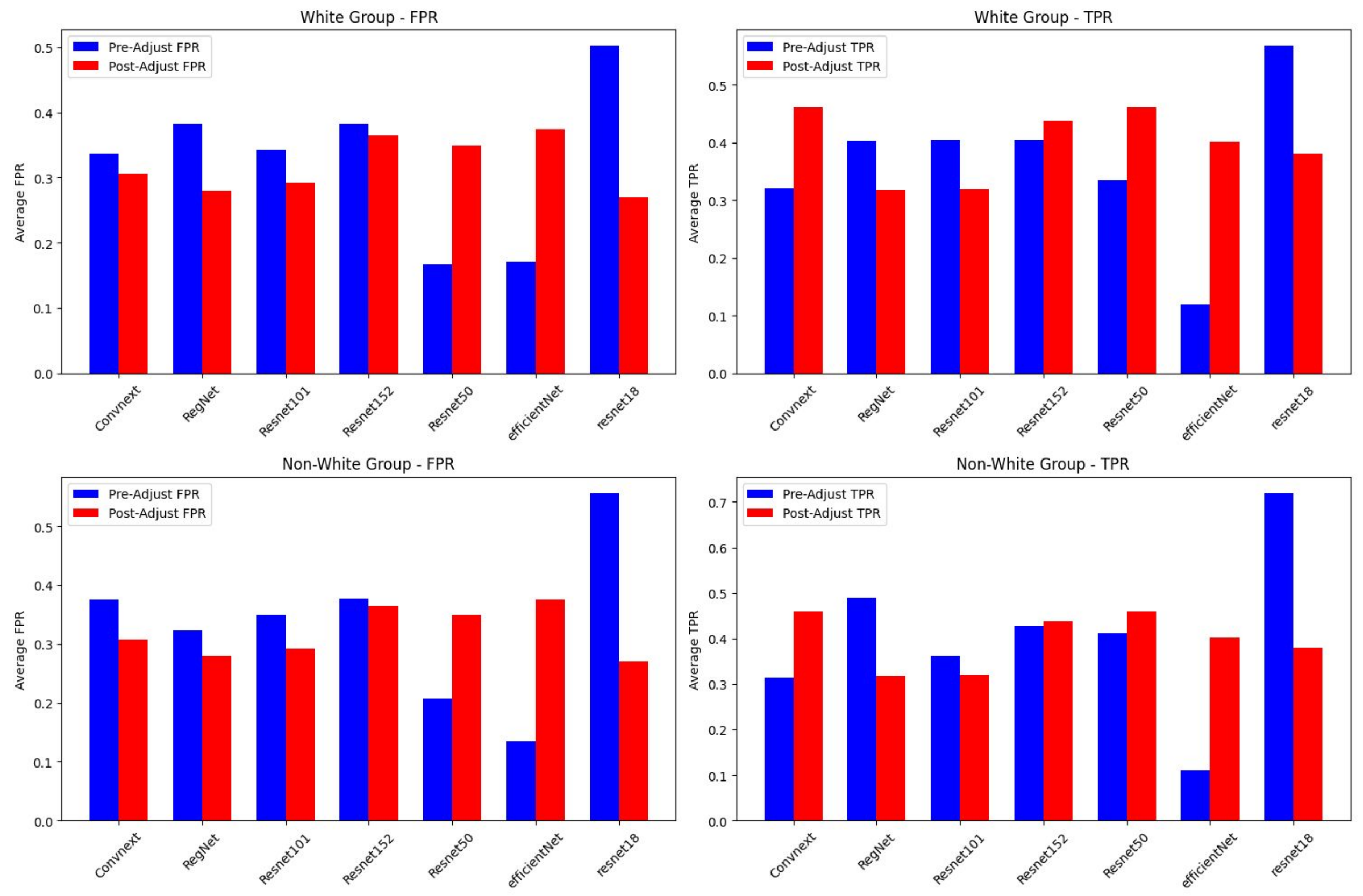
where

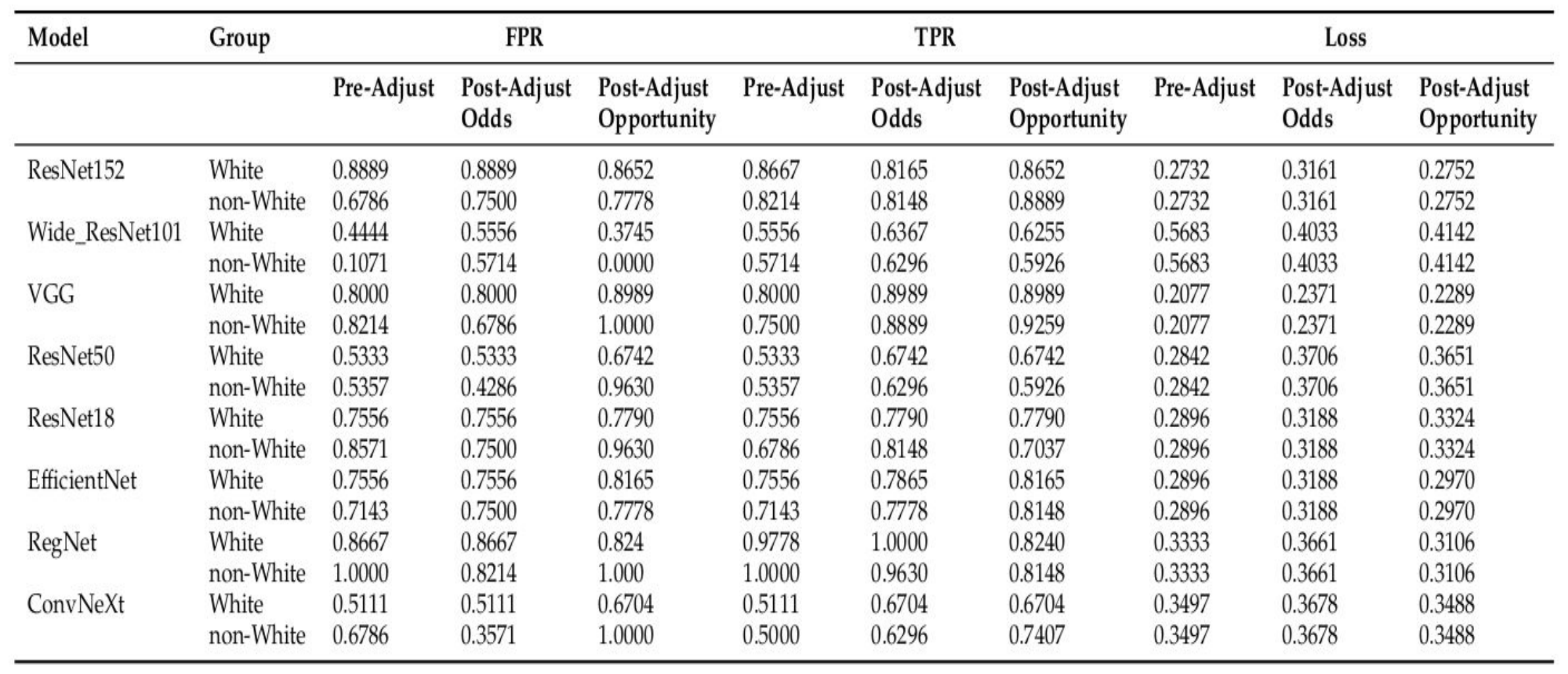
$$(W_{ij}^{\alpha} = Pr(Y_{adj} = i|\hat{Y} = j, A = \alpha))$$

are the protected attribute conditional confusion matrices. To preserve fairness at the individual prediction level, we adopt a stochastic approach. Instead of simply selecting the most probable class, we construct predictions by sampling from the adjusted probabilities. Due to insufficient sample sizes within each demographic group, we encountered instances of zero values for FPs, TPs, FNs, and TNs. To implement our method, we used existing software for calculating fairness metrics, which was originally developed based on binary classification. We add an epsilon term (0.001) to the denominator of each of the four measurements (FPs, TPs, FNs, and TNs) to prevent division errors when calculating the confusion matrix and the fairness metrics (equalized odds and equal opportunity).









Model	Group	FPR			TPR			Loss		
		Pre-Adjust	Post-Adjust Odds	Post-Adjust Opportunity	Pre-Adjust	Post-Adjust Odds	Post-Adjust Opportunity	Pre-Adjust	Post-Adjust Odds	Post-Adjust Opportunity
ResNet152	White	0.8889	0.8889	0.8652	0.8667	0.8165	0.8652	0.2732	0.3161	0.2752
	non-White	0.6786	0.7500	0.7778	0.8214	0.8148	0.8889	0.2732	0.3161	0.2752
Wide_ResNet101	White	0.4444	0.5556	0.3745	0.5556	0.6367	0.6255	0.5683	0.4033	0.4142
	non-White	0.1071	0.5714	0.0000	0.5714	0.6296	0.5926	0.5683	0.4033	0.4142
VGG	White	0.8000	0.8000	0.8989	0.8000	0.8989	0.8989	0.2077	0.2371	0.2289
	non-White	0.8214	0.6786	1.0000	0.7500	0.8889	0.9259	0.2077	0.2371	0.2289
ResNet50	White	0.5333	0.5333	0.6742	0.5333	0.6742	0.6742	0.2842	0.3706	0.3651
	non-White	0.5357	0.4286	0.9630	0.5357	0.6296	0.5926	0.2842	0.3706	0.3651
ResNet18	White	0.7556	0.7556	0.7790	0.7556	0.7790	0.7790	0.2896	0.3188	0.3324
	non-White	0.8571	0.7500	0.9630	0.6786	0.8148	0.7037	0.2896	0.3188	0.3324
EfficientNet	White	0.7556	0.7556	0.8165	0.7556	0.7865	0.8165	0.2896	0.3188	0.2970
	non-White	0.7143	0.7500	0.7778	0.7143	0.7778	0.8148	0.2896	0.3188	0.2970
RegNet	White	0.8667	0.8667	0.824	0.9778	1.0000	0.8240	0.3333	0.3661	0.3106
	non-White	1.0000	0.8214	1.000	1.0000	0.9630	0.8148	0.3333	0.3661	0.3106
ConvNeXt	White	0.5111	0.5111	0.6704	0.5111	0.6704	0.6704	0.3497	0.3678	0.3488
	non-White	0.6786	0.3571	1.0000	0.5000	0.6296	0.7407	0.3497	0.3678	0.3488

Conclusion

We observe biases in the performance of the binary classification model, which consistently performs better on test data corresponding to White individuals. Our work adds further evidence to a wide body of prior work, demonstrating that without care, the integration of AI into diagnostic workflows may amplify existing healthcare disparities. The lack of consistent disparity reductions after fairness adjustments highlights the challenges in applying post-processing techniques to reduce bias in machine learning models trained using medical imaging data. By calibrating the models, we had hoped to improve the equity of AI-enabled diagnostics across different racial groups. However, these methods do not appear to work for deep learning models applied to medical imaging. The primary limitation of this study is the possible lack of generalizability of our findings due to the use of only one dataset for evaluation. Future research on post-processing fairness in medical imaging would benefit from the use of multi-site datasets that cover a broader range of demographic attributes. Another major limitation is that we grouped all non-White patients into a single category for fairness analyses due to the lack of sufficient representation of any race other than White. A more robust analysis would have included performance metrics for each individual race. However, such an analysis requires more samples for the under-represented groups, posing a 'chicken-and-egg problem'.

Future Work

Another interesting area of future work would be studying the explainability of the models in conjunction with fairness. Such a study could aid in the understanding of how different models arrive at their predictions and whether the reasons for arriving at a particular prediction are different across groups.

Refrences

1. Siegel, R.L.; Giaquinto, A.N.; Jemal, A. Cancer statistics, 2024. CA Cancer J. Clin. 2024, 74, 12–49.
2. Golatkar, A.; Anand, D.; Sethi, A. Classification of breast cancer histology using deep learning. In Proceedings of the Image Analysis and Recognition: 15th International Conference, ICIAR 2018, Póvoa de Varzim, Portugal, 27–29 June 2018; Proceedings 15; Springer: Berlin/Heidelberg, Germany, 2018; pp. 837–844.
3. Spanhol, F.A.; Oliveira, L.S.; Petitjean, C.; Heutte, L. Breast cancer histopathological image classification using Convolutional Neural Networks. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; pp. 2560–2567. https://doi.org/10.1109/IJCNN.2016.7727519.
4. Boag, W.; Suresh, H.; Celi, L.A.; Szolovits, P.; Ghassemi, M. Racial Disparities and Mistrust in End-of-Life Care. Proc. Mach. Learn. Res. 2018, 85, 587–602.
5. Adamson, A.S.; Smith, A. Machine Learning and Health Care Disparities in Dermatology. JAMA Dermatol. 2018, 154, 1247–1248. https://doi.org/10.1001/jamadermatol.2018.2348.