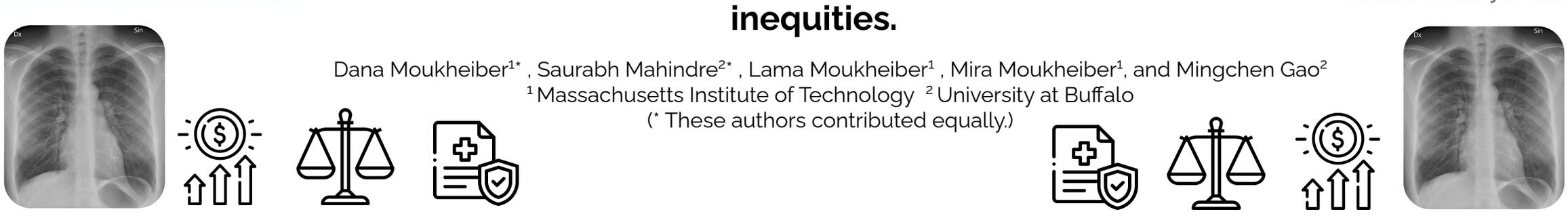
Looking Beyond What You See: An empirical analysis on subgroup intersectional fairness for multi-label chest x-ray classification using social determinants of racial health inequities.

Dana Moukheiber[1*] , Saurabh Mahindre[2*] , Lama Moukheiber[1] , Mira Moukheiber[1], and Mingchen Gao[2]
[1] Massachusetts Institute of Technology [2] University at Buffalo
(* These authors contributed equally.)



Introduction

- In recent years, the integration of deep learning algorithms in real-world health applications, particularly medical imaging, has raised the potential for improved diagnostic accuracy and patient outcomes. However, these algorithms have the capacity to perpetuate biases and maintain stereotypical associations present in the data, particularly to the detriment of marginalized communities.

- Furthermore, it's essential to recognize that race, often misunderstood as a biological characteristic rather than a social construct, is a significant global factor influencing health outcomes. This misclassification further exacerbates health disparities. To comprehensively understand and enhance patient care, it's crucial to investigate how various systems intersect and contribute to the perpetuation of inequities by examining patients' contextual environments.

- Through the lens of social determinants of health (SDOH), we can gain a deeper understanding of the intricate interplay between racism, a social construct, and other determinants of health, shedding light on the mechanisms that sustain healthcare disparities.

Main Contributions

- We assess fair intersectionality in chest X-rays using race and SDOH with eight intersectional groups, recognizing the value of SDOH in capturing contextual information and interconnectedness between race that can influence health outcomes.

- In contrast to previous fairness intersectionality studies in chest X-ray analysis that primarily focus on binary classification, we extend our analysis to a complex multi-label setting, enabling a more comprehensive examination of fairness in chest X-ray diagnostics.

- We adopt a simple, cost-effective, and efficient subgroup robust method and sample a balanced multi-attribute dataset. By adapting equalized odds as a fairness constraint for multi-label settings, we show that our approach is robust to over-fitting and resistant to accuracy and fairness tradeoffs.

Methodology

- All the models are first trained on 2048 X 2048 , anterior-posterior and posterior-anterior MIMIC-CXR images with 14 binary labels. The recently released MIMIC-SDOH dataset allowed us to assess intersectional fairness using both traditional racial attributes from MIMIC-IV and social determinants at a granular level.

- We propose a **fairness class-balanced fine-tuning** method, which involves pre-training a ResNet based model and then re-training the final layer of the model using a balanced sampled dataset while incorporating fairness constraints and addressing class imbalance.

- We add fairness constraints based on the false positive rate and the false negative rate to our overall loss function.

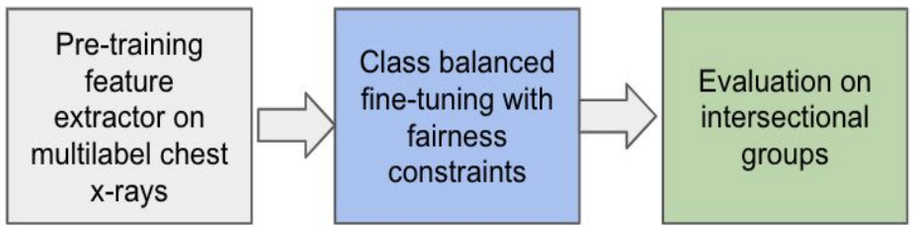
$$L_{fairness} = L_{BCE} + \alpha(fpr + fnr)$$

- We propose to calculate fpr_{kg} and fnr_{kg} for each pair of class k and group g separately and then perform a weighted average to account for class imbalance as follows:

$$fpr = \frac{1}{|G|}\sum_{g}\frac{1}{W}\sum_{k} fpr_{kg}w_k \quad , \quad fnr = \frac{1}{|G|}\sum_{g}\frac{1}{W}\sum_{k} fnr_{kg}w_k$$

$$w_k = \frac{(n - n_k)}{n} \quad , \quad W = \sum_{k} w_k$$

, where nk is the frequency of positive labels for that class.



Overview of our proposed methodology

- Our overall framework consists of two stages:

1. **Pre-training** We initially pretrain a neural network to extract features from chest X-ray images on the training data set of 190k study ids. We use a residual network architecture for the feature extractor trained for multi-label classification using binary cross-entropy loss .

2. **Fine-tuning** We compare our **fairness class-balanced fine-tuning** method to the following methodologies:
   - **ERM (Equalized Odds Re-weighting)**: This method involves training a pre-trained without considering sensitive attributes.
   - **Fine-tuning**: In this approach, the final layer of the model is re-trained using an imbalanced sampled dataset.
   - **Deep Feature Reweighting (DFR)**: DFR involves re-training the final layer of the model using a balanced sampled dataset.

Experiments & Results

Income	Insurance	Race	No. Studies
Low	Low	White	20,638
Low	Low	Non-White	10,650
Low	High	White	20,308
Low	High	Non-White	26,261
High	Low	White	50,499
High	Low	Non-White	9,666
High	High	White	13,214
High	High	Non-White	5,261
Total			193,730

Table 1: Number of study-ids present across eight intersectional groups. Based on the tract-level and county-level social determinants, we select income and insurance coverage, respectively. Both social determinant attributes are stratified for high-level and low-level. We also include race stratified for whites and non-whites.

- We find that ERM achieves higher AUC and weighted-accuracy (WACC) values compared to other methods but worse fairness metrics.

- Our fairness class balanced fine-tuning method has the lowest equalized odds difference (EO_Diff) incidence, indicating reduced disparities and biases in model predictions across different subgroups.

- Our method also exhibits the most favorable fairness metrics and overall performance, as indicated by the highest accuracy-fairness (AF) value (i.e., WACC - EO Diff).

These findings support the efficacy of our approach, which involves adopting fairness class-balanced fine-tuning in the context of multi-label classification.

Method	AUC_{avg}	EO_Diff_{avg}	$WACC_{avg}$	AF_{avg}
ERM	**0.7862**	0.5050	**0.6730**	0.1680
Finetuning	0.7436	0.5047	0.6537	0.1490
DFR	0.7399	0.4788	0.6520	0.1731
Fairness class-balanced finetuning	0.7429	**0.4622**	0.6550	**0.1940**

Table 2: Model performance on MIMIC-CXR for multi-label classification incorporating equalized odds difference fairness constraint. Averaged values are reported over 100 random trials.

Conclusion

- In this study, we introduce a framework that focuses on achieving accurate diagnostic results while ensuring fairness across diverse intersectional groups in high-dimensional chest X-ray multi-label classification. We adopt an intersectional multi-attribute fairness approach, extending beyond traditional protected attributes to consider complex interactions within sensitive attributes.

- Our method involves fine-tuning pretrained models with equalized odds (EO) as a fairness constraint and incorporating weights for multi-label settings, effectively mitigating fairness bias without altering model predictions during testing.

- Evaluation on the MIMIC-CXR dataset reveals that our fairness class-balanced fine-tuning approach outperforms other methods in both fairness metrics (EO Diff) and accuracy-fairness (AF) evaluation. By recognizing the influence of social determinants of health, our approach aims to advance health equity and represents a promising step in improving medical algorithms.