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] Mingchen Gao[2]

[1]Massachusetts Institute of Technology [2]University at Buffalo

Abstract

There has been significant progress in implementing deep learning models in disease diagnosis using chest X-rays. Despite these advancements, inherent biases in these models can lead to disparities in prediction accuracy across protected groups. In this study, we propose a framework to achieve accurate diagnostic outcomes and ensure fairness across intersectional groups in high-dimensional chest X-ray multi-label classification. Transcending traditional protected attributes, we consider complex interactions within social determinants, enabling a more granular benchmark and evaluation of fairness. We present a simple and robust method that involves retraining the last classification layer of pre-trained models using a balanced dataset across groups. Additionally, we account for fairness constraints and integrate class-balanced fine-tuning for multi-label settings. The evaluation of our method on the MIMIC-CXR dataset demonstrates that our framework achieves an optimal tradeoff between accuracy and fairness compared to baseline methods.

1. Introduction

Deep learning algorithms trained on large-scale medical imaging data have been increasingly employed in real-world health applications in the past few years due to their potential to improve diagnostic accuracy and patient outcomes [36, 15, 2, 48, 19]. However, these models also have the potential to learn and exacerbate pre-existing biases in data, perpetuating stereotypical connections and leading to negative consequences for marginalized communities [43]. The advancements and rapid deployment of computer vision models in healthcare settings highlight the necessity of acknowledging the possibility of inherent biases and risks

*Equal contribution.

across protected patient attributes related to disease, including race and gender.

The misclassification of race as a biological aspect, rather than a social construct, further exacerbates health disparities. To gain a more comprehensive understanding and enhance patient care, it is important to explore the patient's contextual environment, as a patient's clinical profile offers only a limited glimpse of the factors impacting their health [38, 16, 47]. By examining the broader social, economic, and environmental factors through the lens of social determinants of health (SDOH), we can better comprehend the intricate interplay between racial biases and other determinants of health, shedding light on the mechanisms that perpetuate inequities in healthcare [5].

Several studies have been conducted to evaluate subgroup robust methods for addressing fairness bias across various domains of images with real-world applications [56, 13, 28, 50]. These studies mostly benchmark bias and fairness algorithms for binary classification [17, 53, 18, 55] focusing on basic protective attribute categories along single dimensions. However, there has been limited research investigating intersectional bias with multiple demographic dimensions [54, 44, 55, 29], and the integration of SDOH into these studies remains unexplored, primarily due to privacy concerns[10, 23].

Herein, we adopt a simple and cost-effective method to fine-tune multi-label classification models for intersectional fairness across multiple attributes. Specifically, we measure intersectional group fairness in chest X-rays by considering race and two (SDOHs) — health insurance and income, forming eight distinct intersectional groups. Our approach adapts current subgroup robust methods to a balanced sampled multi-attribute dataset and addresses fairness constraints and class imbalance for multi-label settings. This adaptation extends beyond previous fairness intersectionality studies, which primarily focus on binary classification and intersections of two attributes. When applied to the MIMIC-CXR multi-label dataset, we show that our method performs well not only in classification perfor-

mance but also in fairness metrics, outperforming established baselines.

2. Related Works

There have been several works proposed to combat fairness bias across benchmarks and real-world datasets:

Debiasing Algorithms: Debiasing algorithms play a crucial role in enhancing performance across minority subpopulations, thereby improving accuracy uniformly across different groups. These strategies can be broadly categorized into two types: those that rely on attributes being explicitly labeled and available during training, and attribute-agnostic methods, which do not necessitate direct access to such attributes during the training process. Empirical Risk Minimization (ERM)[45] , a foundational concept in machine learning, aims to minimize the average error across all samples by offering a generalized framework that can be adapted to include more sophisticated debiasing techniques. Some of these advanced methods, such as up-weighting poorly performing samples[32, 37, 30], directly build upon the ERM framework to ensure equitable performance across diverse groups. For instance, the Deep Feature Re-weighting (DFR) technique[17] begins with training a model using ERM and subsequently refines it by re-training with a balanced sample set across groups, illustrating a practical application of debiasing algorithms. These methodologies are critically evaluated using a variety of fairness metrics, ensuring measurable improvements in algorithmic fairness.

Fairness Metrics: Bias mitigation algorithms are evaluated using various fairness metrics. Some well-known metrics in this context are demographic parity [12], equalized odds, and equality of opportunity. Demographic parity involves comparing the average prediction score across different subgroups. Equality of Opportunity takes the label distribution into account and assesses the True Positive Rate (TPR) gap among different groups. Equalized Odds [14] measures both the TPR and False Positive Rate (FPR) gaps across various groups.

Pre-processing, In-processing, and Post-processing Fairness Techniques: Several prior works on mitigating fairness bias [11, 35] have been proposed, including pre-processing techniques [49, 24, 6, 4], train-time techniques [26, 25], and post-processing techniques [14, 24, 40, 51, 3, 42], which aim to mitigate fairness bias before, during, or after model training, respectively. However, as presented by [8, 41], over-parameterized models overfit to fairness objectives, which is an acute problem, especially in high-stake clinical settings. In addition, as outlined by [9], there are inherent imbalances that extend beyond class labels to include sensitive attributes, posing a challenge to the generalizability of fairness properties in over-parameterized models like neural networks. This issue becomes particularly pronounced in large and challenging datasets, especially when the model tends to favor minority attributes.

Intersectional Fairness: Intersectional biases occur when protected attributes interact with each other. Previous research on bias evaluation in chest X-ray imaging has primarily focused on examining protected attributes such as race, age, and gender as mutually exclusive categories with single dimensions [56, 31]. Yet, only a few studies have explored protected attributes as non-mutually exclusive categories with multiple dimensions. Such an approach is critical, given that the disparities affecting intersectional subgroups can be profound, highlighting the need to address the compounded biases within marginalized communities. [44] evaluated a DenseNet model on chest X-rays and revealed lower true positive rates (TPRs) across four categories of protected attributes, namely gender, age, race, and insurance type. However, their evaluation only focused on insurance as a proxy for social determinants of health (SDOH). Here, we focus on equalized odds as a metric to facilitate comparisons with previous work on studying fairness in medical imaging [53, 54, 56]. Our study goes beyond race, gender, and age, avoiding pre-defined proxies of SDOH. Instead, we use and incorporate social determinant attributes sourced from country and tract-level data to better understand the complex interactions between various SDOH factors for more granular benchmarking and evaluation of fairness in clinical settings.

3. Methods

This section describes our methods and details the implementation process. We assume we have a dataset of n chest X-ray samples and C target binary classes. Each sample can belong to one or more classes due to the multi-label nature of chest X-rays. The binary variable y_{ik} denotes whether sample i has class k as positive, where $k \in C$. For intersectional groups, we define a set denoted as G. The binary variable a_{ig} indicates whether a sample i belongs to group g, where $g \in G$. We conduct experiments on MIMIC-CXR, which is divided into 190,000 training samples and 2,500 testing samples. We link MIMIC-CXR to MIMIC-IV and MIMIC-SDOH to create intersectional groups (see dataset details in Appendix 6.1). The number of samples used across eight intersectional groups is presented in Table 1.

3.1. Pre-training a residual network for feature extraction

We first train a neural network to extract features from chest X-ray images using the training data set. We adopt a residual network architecture as the feature extractor similar to prior studies on chest X-ray classification [7, 19, 36, 20]. ERM is a standard training method that minimizes the average loss across samples. It has demonstrated its effective-

Table 1. Number of samples present across eight intersectional groups.

Income	Insurance	Race	No. Samples
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

ness in extracting meaningful features for both single and multi-label chest X-ray classification, and it has been incorporated into diverse fairness benchmarks [32, 37, 17]. Consequently, we train the residual network via ERM. Since our training data set exhibits class imbalance, we use class weights, p_k, for each class for effective pre-training. The weighted version of binary cross-entropy loss used in pre-training is described in 1,

$$L_{BCE} = - \sum_{i=1...n} \sum_{k=1...|C|} p_k y_{ik} \log \hat{y_{ik}} + (1 - y_{ik}) \log (1 - \hat{y_{ik}}), \tag{1}$$

where $p_k = \frac{\text{\# of samples with no findings class}}{\text{\# of samples in class } k}$ is the positive weight for each class k and $\hat{y_{ik}}$ represents the predicted probability of class k for the i-th sample, obtained after applying the sigmoid activation function to the logits.

3.2. Class balanced fine-tuning on a sampled dataset with fairness constraints

For fine-tuning, we freeze the feature extractor and re-train a new final classification layer on a sampled data set with balanced group distribution [27] for robustness across intersectional groups[1]. Next, we add fairness constraints based on the false positive rate, fpr, (6) and false negative rate, fnr, (7) [34] [39] to our overall loss function. Intersectional groups and multi-label samples pose a challenge since a sample can belong to multiple classes and one of many intersectional groups. Compared to prior work [33] which only considers two values of a sensitive attribute in single label settings, we propose to calculate fpr_{kg} (3) and fnr_{kg} (4) for each pair of class k and group g separately to accommodate intersectional groups in multi-label settings.

$$L_{fairness} = L_{BCE} + \alpha(fpr + fnr) \tag{2}$$

[1]The pre-trained model weights and classifier weights will be made available on HuggingFace.

$$fpr_{kg} = \left| \frac{\sum_i \hat{y_{ik}}(1 - y_{ik})a_{ig}}{\sum_i a_{ig}} - \frac{\sum_i \hat{y_{ik}}(1 - y_{ik})(1 - a_{ig})}{\sum_i (1 - a_{ig})} \right| \tag{3}$$

$$fnr_{kg} = \left| \frac{\sum_i (1 - \hat{y_{ik}})y_{ik}a_{ig}}{\sum_i a_{ig}} - \frac{\sum_i (1 - \hat{y_{ik}})y_{ik}(1 - a_{ig})}{\sum_i (1 - a_{ig})} \right| \tag{4}$$

To account for class imbalance, we propose to use a weighted average of fpr_{kg} and fnr_{kg} for all $k \in C$ while aggregating for group g. The weights w_k (5) are devised based on frequency n_k of the positive label y_k in the training set of n samples. The final loss in the fine-tuning step is (2).

$$w_k = \frac{(n - n_k)}{n}, W = \sum_k w_k \tag{5}$$

$$fpr = \frac{1}{|G|} \sum_g \frac{1}{W} \sum_k fpr_{kg}w_k \tag{6}$$

$$fnr = \frac{1}{|G|} \sum_g \frac{1}{W} \sum_k fnr_{kg}w_k \tag{7}$$

Compared Methods. We consider three baseline methods, followed by our proposed method:

ERM: Training a residual network where the final layer is a classification layer without considering intersectional groups.
Fine-tuning: Re-training a new final classification layer of the residual network using an imbalanced sampled dataset [27].
Deep Feature Reweighting (DFR): Re-training a new final classification layer of the residual network using a balanced sampled dataset across intersectional groups.
Fair Class-balanced Fine-tuning (Ours): Re-training a new final classification layer of the residual network using a sampled dataset balanced across inter-sectional groups, considering both fairness constraints and class imbalance.

Evaluation Metrics. To evaluate the performance of our model on the unseen test set, we use weighted accuracy (WACC) and Area under the ROC Curve (AUC) [34]. We use two metrics for classification: WACC, which accounts for the class imbalance and AUC. For multi-label settings, these metrics are averaged across all labels. For fairness evaluation, we adopt two metrics: equalized odds difference (EO_Diff) [14] and accuracy-fairness (AF) [34]. EO_Diff

Table 2. Model performance on MIMIC-CXR for multi-label classification incorporating the equalized odds difference fairness constraint. We use demographic attributes from MIMIC and social determinant attributes from MIMIC-SDOH dataset to form the intersectional groups. Averaged values are reported over 100 random trials.

Method	AUC_{avg} ($\uparrow$)	EO_Diff_{avg} ($\downarrow$)	$WACC_{avg}$ ($\uparrow$)	AF_{avg} ($\uparrow$)
ERM	**0.7861**	0.4243	0.6438	0.2195
Fine-tuning	0.7800	0.3811	0.6292	0.2481
DFR	0.7806	0.3677	**0.6526**	0.2579
Fair Class-balanced Fine-tuning (Ours)	0.7763	**0.3224**	0.6045	**0.2820**

calculates the maximum between two differences: the difference between the minimum and maximum true positive rates and the difference between the minimum and maximum false positive rates across all intersectional groups [14] [46]. For multi-label settings, we calculate EO_Diff for each label separately and then average it as shown in 9.

$$EO\_Diff_k = \max\{\Delta tpr_k, \Delta fpr_k\} \qquad (8)$$

$$EO\_Diff_{avg} = \frac{\sum_k EO\_Diff_k}{|C|} \qquad (9)$$

AF is a metric derived from an equal-weight linear combination of weighted accuracy and fairness: AF = WACC - EO_Diff [34].

3.3. Implementation Details

Pre-training details: For pre-training, we use Adam optimizer with a learning rate of 10^{-4} and weight decay of 10^{-4}. The mini-batch size is set to 32 due to hardware limitations and we perform training on the whole training set for three epochs.

Training details: In order to create a balanced data set for fine-tuning, we sample 2500 study-ids per group at random from the training data set. For fine-tuning, we use Adam optimizer with a learning rate of $5x10^{-5}$, weight decay of 10^{-3}, and batch size of 32. We set the value of α to 1 for simplicity and perform fine-tuning for one full epoch.

Evaluation: While computing the weighted accuracy, we use a 0.5 threshold to convert the output probabilities to predicted labels.

4. Results and Discussion

We conduct multi-label classification on MIMIC-CXR to predict the 14 base classes, evaluating our proposed method compared to three baseline methods. The results of the model performance, along with fairness metrics, are presented in Table 2. We assess several metrics, including AUC_{avg}, $WACC_{avg}$, and AF_{avg}, where higher values indicate better performance. Conversely, for EO_Diff_{avg}, a

smaller value is considered more desirable. Our findings reveal that ERM achieves the highest AUC and second-highest WACC values but the lowest fairness metrics. Additionally, we observe that fair class balanced fine-tuning results in the lowest EO_Diff_{avg}, suggesting reduced disparities and biases in model predictions among intersectional groups. Overall, fair class balanced fine-tuning exhibits the most favorable fairness metrics and overall performance. This is indicated by the highest AF value, although the $WACC_{avg}$ and AUC_{avg} are slightly reduced.

5. Conclusion

In this study, we introduce a framework aimed at promoting equitable representation across diverse intersectional groups in high-dimensional, multi-label chest X-ray classification. We adopt an intersectional multi-attribute fairness perspective, we consider complex interactions within sensitive attributes, going beyond traditional protected attributes to include social determinants of health such as insurance and income. Our methods involves retraining the final classification layer of pre-trained models using a balanced sampled multi-attribute dataset. We also consider both fairness constraints in our overall loss and accommodate intersectional groups in multi-label settings, providing a simple and robust method for assessing fairness in real-world clinical applications. Our evaluation on the MIMIC-CXR dataset demonstrates that our method improves equalized odds difference and accuracy-fairness metrics marking a promising step forward in medical algorithms.

Acknowledgements

References

[1] Agency for Healthcare Research and Quality. Social Determinants of Health Database. https://www.ahrq.gov/sdoh/data-analytics/sdoh-data.html. Accessed: March 28, 2024.

[2] Nkechinyere N Agu, Joy T Wu, Hanqing Chao, Ismini Lourentzou, Arjun Sharma, Mehdi Moradi, Pingkun Yan, and James Hendler. Anaxnet: anatomy aware multi-label finding classification in chest x-ray. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pages 804–813. Springer, 2021.

[3] Ibrahim M Alabdulmohsin and Mario Lucic. A near-optimal algorithm for debiasing trained machine learning models. *Advances in Neural Information Processing Systems*, 34:8072–8084, 2021.

[4] Ibrahim M Alabdulmohsin, Jessica Schrouff, and Sanmi Koyejo. A reduction to binary approach for debiasing multiclass datasets. *Advances in Neural Information Processing Systems*, 35:2480–2493, 2022.

[5] Leo Anthony Celi, Jacqueline Cellini, Marie-Laure Charpignon, Edward Christopher Dee, Franck Dernoncourt, Rene Eber, William Greig Mitchell, Lama Moukheiber, Julian Schirmer, Julia Situ, et al. Sources of bias in artificial intelligence that perpetuate healthcare disparities—a global review. *PLOS Digital Health*, 1(3):e0000022, 2022.

[6] L Elisa Celis, Vijay Keswani, and Nisheeth Vishnoi. Data preprocessing to mitigate bias: A maximum entropy based approach. In *International conference on machine learning*, pages 1349–1359. PMLR, 2020.

[7] Geeticka Chauhan, Ruizhi Liao, William Wells, Jacob Andreas, Xin Wang, Seth Berkowitz, Steven Horng, Peter Szolovits, and Polina Golland. Joint modeling of chest radiographs and radiology reports for pulmonary edema assessment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 529–539. Springer, 2020.

[8] Valeriia Cherepanova, Vedant Nanda, Micah Goldblum, John P Dickerson, and Tom Goldstein. Technical challenges for training fair neural networks. *arXiv preprint arXiv:2102.06764*, 2021.

[9] Zhun Deng, Jiayao Zhang, Linjun Zhang, Ting Ye, Yates Coley, Weijie J Su, and James Zou. Fifa: Making fairness more generalizable in classifiers trained on imbalanced data. *arXiv preprint arXiv:2206.02792*, 2022.

[10] Rebecca Driessen, Neil Bhatia, Judy Wawira Gichoya, Nabile M Safdar, and Patricia Balthazar. Sociodemographic variables reporting in human radiology artificial intelligence research. *Journal of the American College of Radiology*, 2023.

[11] Mengnan Du, Fan Yang, Na Zou, and Xia Hu. Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems*, 36(4):25–34, 2020.

[12] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 259–268, New York, NY, USA, 2015. Association for Computing Machinery.

[13] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

[14] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[15] Nasir Hayat, Hazem Lashen, and Farah E Shamout. Multi-label generalized zero shot learning for the classification of disease in chest radiographs. In *Machine Learning for Healthcare Conference*, pages 461–477. PMLR, 2021.

[16] Casey Holmes Fee, Rachel Scarlett Hicklen, Sidney Jean, Nebal Abu Hussein, Lama Moukheiber, Michelle Foronda de Lota, Mira Moukheiber, Dana Moukheiber, Leo Anthony Celi, and Irene Dankwa-Mullan. Strategies and solutions to address digital determinants of health (ddoh) across underinvested communities. *PLOS digital health*, 2(10):e0000314, 2023.

[17] Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew G Wilson. On feature learning in the presence of spurious correlations. *Advances in Neural Information Processing Systems*, 35:38516–38532, 2022.

[18] Sarah Jabbour, David Fouhey, Ella Kazerooni, Michael W Sjoding, and Jenna Wiens. Deep learning applied to chest x-rays: Exploiting and preventing shortcuts. In *Machine Learning for Healthcare Conference*, pages 750–782. PMLR, 2020.

[19] Zhanghexuan Ji, Mohammad Abuzar Shaikh, Dana Moukheiber, Sargur N Srihari, Yifan Peng, and Mingchen Gao. Improving joint learning of chest x-ray and radiology report by word region alignment. In *International Workshop on Machine Learning in Medical Imaging*, pages 110–119. Springer, 2021.

[20] Zhanghexuan Ji, Mohammad Abuzar Shaikh, Dana Moukheiber, Sargur N Srihari, Yifan Peng, and Mingchen Gao. Improving joint learning of chest x-ray and radiology report by word region alignment. In *Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings*, page 110–119, Berlin, Heidelberg, 2021. Springer-Verlag.

[21] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

[22] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

[23] Nathan Kallus, Xiaojie Mao, and Angela Zhou. Assessing algorithmic fairness with unobserved protected class using

data combination. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 110, New York, NY, USA, 2020. Association for Computing Machinery.

[24] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012.

[25] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*, pages 35–50. Springer, 2012.

[26] Michael P Kim, Amirata Ghorbani, and James Zou. Multi-accuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254, 2019.

[27] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022.

[28] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[29] Narine Kokhlikyan, Bilal Alsallakh, Fulton Wang, Vivek Miglani, Oliver Aobo Yang, and David Adkins. Bias mitigation framework for intersectional subgroups in neural networks. *arXiv preprint arXiv:2212.13014*, 2022.

[30] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33:728–740, 2020.

[31] Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, 2020.

[32] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021.

[33] Padala Manisha and Sujit Gujar. Fnnc: Achieving fairness through neural networks. *arXiv preprint arXiv:1811.00247*, 2018.

[34] Yuzhen Mao, Zhun Deng, Huaxiu Yao, Ting Ye, Kenji Kawaguchi, and James Zou. Last-layer fairness fine-tuning is simple and effective for neural networks. *arXiv preprint arXiv:2304.03935*, 2023.

[35] Ricards Marcinkevics, Ece Ozkan, and Julia E. Vogt. De-biasing deep chest x-ray classifiers using intra- and post-processing methods. In Zachary Lipton, Rajesh Ranganath, Mark Sendak, Michael Sjoding, and Serena Yeung, editors, *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pages 504–536. PMLR, 05–06 Aug 2022.

[36] Dana Moukheiber, Saurabh Mahindre, Lama Moukheiber, Mira Moukheiber, Song Wang, Chunwei Ma, George Shih, Yifan Peng, and Mingchen Gao. Few-shot learning geometric ensemble for multi-label classification of chest x-rays. In *Data Augmentation, Labelling, and Imperfections: Second MICCAI Workshop, DALI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings*, pages 112–122. Springer, 2022.

[37] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020.

[38] Lama H Nazer, Razan Zatarah, Shai Waldrip, Janny Xue Chen Ke, Mira Moukheiber, Ashish K Khanna, Rachel S Hicklen, Lama Moukheiber, Dana Moukheiber, Haobo Ma, et al. Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS Digital Health*, 2(6):e0000278, 2023.

[39] Manisha Padala and Sujit Gujar. Fnnc: Achieving fairness through neural networks. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, 2021.

[40] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *Advances in neural information processing systems*, 30, 2017.

[41] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR, 2020.

[42] Jessica Schrouff, Natalie Harris, Sanmi Koyejo, Ibrahim M Alabdulmohsin, Eva Schnider, Krista Opsahl-Ong, Alexander Brown, Subhrajit Roy, Diana Mincu, Christina Chen, et al. Diagnosing failures of fairness transfer across distribution shift in real-world medical settings. *Advances in Neural Information Processing Systems*, 35:19304–19318, 2022.

[43] Kenneth P Seastedt, Dana Moukheiber, Saurabh A Mahindre, Chaitanya Thammineni, Darin T Rosen, Ammara A Watkins, Daniel A Hashimoto, Chuong D Hoang, Jacques Kpodonu, and Leo A Celi. A scoping review of artificial intelligence applications in thoracic surgery. *European Journal of Cardio-Thoracic Surgery*, 61(2):239–248, 10 2021.

[44] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew BA McDermott, Irene Y Chen, and Marzyeh Ghassemi. Under-diagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine*, 27(12):2176–2182, 2021.

[45] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

[46] Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*, pages 1–7, 2018.

[47] Stephen Waite, Jinel Scott, and Daria Colombo. Narrowing the gap: imaging disparities in radiology. *Radiology*, 299(1):27–35, 2021.

[48] Ryan Wang, Li-Ching Chen, Lama Moukheiber, Kenneth P Seastedt, Mira Moukheiber, Dana Moukheiber, Zachary Zaiman, Sulaiman Moukheiber, Tess Litchman, Hari Trivedi, et al. Enabling chronic obstructive pulmonary disease diagnosis through chest x-rays: A multi-site and multi-modality study. *International Journal of Medical Informatics*, 178:105211, 2023.

[49] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5310–5319, 2019.

[50] Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre Alvise-Rebuffi, Ira Ktena, Krishnamurthy Dvijotham, and Taylan Cemgil. A fine-grained analysis on distribution shift, 2021.

[51] Forest Yang, Mouhamadou Cisse, and Sanmi Koyejo. Fairness with overlapping groups; a probabilistic perspective. *Advances in neural information processing systems*, 33:4067–4078, 2020.

[52] Ming Ying Yang, Gloria Hyunjung Kwak, Tom Pollard, Leo Anthony Celi, and Marzyeh Ghassemi. Evaluating the impact of social determinants on health prediction. *arXiv preprint arXiv:2305.12622*, 2023.

[53] Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. Change is hard: A closer look at subpopulation shift. *arXiv preprint arXiv:2302.12254*, 2023.

[54] Haoran Zhang, Natalie Dullerud, Karsten Roth, Lauren Oakden-Rayner, Stephen Pfohl, and Marzyeh Ghassemi. Improving the fairness of chest x-ray classifiers. In *Conference on Health, Inference, and Learning*, pages 204–233. PMLR, 2022.

[55] Haoran Zhang, Natalie Dullerud, Laleh Seyyed-Kalantari, Quaid Morris, Shalmali Joshi, and Marzyeh Ghassemi. An empirical framework for domain generalization in clinical settings. In *Proceedings of the conference on health, inference, and learning*, pages 279–290, 2021.

[56] Yongshuo Zong, Yongxin Yang, and Timothy Hospedales. Medfair: Benchmarking fairness for medical imaging. *arXiv preprint arXiv:2210.01725*, 2022.

6. Appendix

6.1. Dataset Description

We utilize MIMIC-IV and MIMIC-SDOH to create attribute intersectional groups, which are then matched to the MIMIC-CXR images.

- **MIMIC-CXR**: A comprehensive collection of de-identified Chest X-ray (CXR) data acquired from Beth Israel Deaconess Medical Center in Boston, United States. MIMIC-CXR includes 14 binary labels that indicate the presence or absence of pathology.[22]. Specifically, we follow [36] to resize the images into 2048 X 2048 and we only use anterior-posterior (AP) and posterior-anterior (PA) radiographs view positions.

- **MIMIC-IV**: An electronic health records database that contains data on patients admitted to the intensive care unit at Beth Israel Deaconess Medical Center. This extensive database comprises comprehensive information about patients' clinical records, demographic details, laboratory findings, and other pertinent medical data [21]. In our study, we extract racial information from MIMIC-IV and categorize the race into two groups: white and non-white.

- **MIMIC-SDOH**: A database resulting from the integration of the MIMIC-IV clinical database with Social Determinants of Health (SDOH) databases: County Health Rankings (CHR), Social Vulnerability Index (SVI), and Social Determinants of Health Database (SDOHD) [52]. Here, we focus on the variables available in the SDOHD [1] which offers more detailed and granular SDOH data. The SDOHD provides information at the county, census tract, and ZIP code levels, encompassing variables related to economic, healthcare, education, and social contexts as well as physical infrastructure. From this data, we extract health insurance information at the county level, focusing on the estimated percentage of the uninsured population for all income levels (under 65 years). Additionally, we gather income information at the tract level, specifically the median household income (dollars, inflation-adjusted to the data file year). Subsequently, we categorize both the health insurance and income data into two distinct groups each: high and low income, as well as high and low insurance coverage.