Lightweight Model Adaptation for Mitigating Bias in Deep Learning Models for Chest X-Ray Analysis

Clemence Mottez
Institute of Computational and Mathematical Engineering
Center of Artificial Intelligence in Medicine and Imaging
Stanford University

Abstract

Deep learning models have demonstrated significant potential in improving chest X-ray diagnosis. However, these models may exacerbate healthcare disparities. Addressing the inherent biases of deep learning models is essential to ensure their safe and reliable deployment in clinical practice. We extend a recent method that combines embeddings extracted by a convolutional neural network (CNN) with an eXtreme Gradient Boosting (XGBoost) classifier to effectively mitigate bias across the sex, age, and race subgroups. We first demonstrate the generalizability of our method across multiple medical conditions. Then, we show that retraining the CNN head with XGBoost achieves greater bias reduction compared to alternative classifiers. We further explore the adaptability of our lightweight approach by applying it to image-feature extraction models, including foundation models based on Contrastive Language-Image Pre-Training and Distillation with No Labels architectures, noting weaker results compared to the CNN. Moreover, we show that our lightweight bias mitigation technique outperforms existing bias mitigation techniques, such as data augmentation, active learning, and adversarial attacks, that require the full model retraining. Finally, we show that combining our XGBoost head retraining with active learning leads to the optimal balance, significantly reducing bias without compromising overall model performance.

1. Introduction

Deep Learning (DL) models have the potential to transform healthcare by increasing diagnostic accuracy, personalizing treatment, and improving patient outcomes[1]. However, these technologies risk exacerbating healthcare disparities if their performance varies across different subgroups of patients, for example according to sex, age, and race [14]. These biases may arise from training data that underrepresents certain populations, algorithm designs that overlook the unique characteristics of different groups, or disparities in healthcare access[5]. Biases are among the many barriers that prevent the deployment of these models in clinical practice, where equitable outcomes are crucial[12]. Current bias mitigation methods involve trade-offs between fairness and accuracy. Techniques such as rebalancing training datasets or modifying algorithms often require extensive model retraining[13] and are thus impractical in healthcare due to data scarcity and resource constraints. To address these limitations, we propose an efficient and lightweight method that significantly reduces bias without retraining the full model, offering a practical solution for resource-constrained clinical settings.

2. Related Work

This project is based on the extension of a recent study [3]. It proposes a lightweight model adaptation strategy to mitigate biases related to sex, age, and race in Chest X-ray (CXR) diagnosis by replacing the final classification layer of a CNN with an XGBoost model which is then retrained on a curated subset of data. However, this study is limited to one disease, does not compare the use of XGBoost with other models, is limited to only one model architecture, does not combine it with existing bias mitigation strategies, and does not compare it with existing work. Our research aims to tackle these limitations.

Similar work has been done on last layer retraining to tackle spurious correlations[8]. This paper demonstrate that simple last layer retraining can match or outperform state-of-the-art approaches on spurious correlation benchmarks, but with profoundly lower complexity and computational expenses. However, they only retrain the head with a linear classifier and don't explore other models.

Some research involves combining a CNN with an XGBoost classifier head to improve the model's performance[6, 10], but they explore this combination as a way to mitigate bias.

	Train-valid-test size	Sex ratio (F-M)	Age ratio (Y-O)	Race ratio (W-B-A)
CheXpert	67,263-4,484-40,358	44.8-55.2	44.4-55.6	80.6-15.5-3.9
MIMIC	- - 37,446	42.0-58.0	62.9-37.1	78.2-7.1-14.7

Figure 1: Dataset information

3. Data

We evaluate our method on the two largest CXR publicly available datasets to ensure the robustness and generalization of our model across different clinical environments.

- CheXpert Plus[2], a dataset consisting of 224,316 CXRs obtained at Stanford Health Care.
- Medical Information Mart for Intensive Care (MIMIC)[7] comprising 377,110 CXRs performed at the Beth Israel Deaconess Medical Center.

Both of these datasets contain demographic information such as sex, age, and race for each patient. For sex, we focused on the difference in performance between males and females; for age, we used a threshold of 70 years old; for race, we focused our analysis on White, Black, and Asian. The splits used, as well as the class imbalance information are in Fig.1. We don't provide Train and Valid splits for MIMIC since we will only train our models on CheXpert. We will then test them on both CheXpert and MIMIC to ensure the consistency of our results in distribution (ID) and Out-Of-Distribution (OOD). We had to do some data preprocessing. First, we only kept posterio anterior and anterio posterior images by removing all the lateral images, resulting in 112,105 CXR for CheXpert and 139,508 for MIMIC. Then, we resized the images to 224x224 since the DenseNet121 model takes as input 224x224 images.

4. Method

4.1. Metrics:

First of all, there is no universal definition of bias. It is actually hard to define clinical bias. There is always a trade-off between fairness and other important metric such as overall performance. In this paper, we state that if we decrease the bias, we should at least keep the same overall performance that we had before reducing the bias. This might disadvantage some subgroup, but it is the fairer approach we could think about.

For the metrics, we want to:

- Have a good overall performance: AUPRC provide a good trade-off between precision and recall and is effective when dealing with imbalanced datasets.
- Have similar performance across subgroups: Δ AUPRC. If we have more than one subgroup (for race for example we are studying White, Black, and Asian, we take the maximum of the Δ AUPRC.

4.2. Models:

We extract embeddings from the CXR images using pretrained models to leverage the extensive feature learning these models have acquire.

- DenseNet-121 from TorchXrayVision, the same model used in the original paper presenting the new hybrid method. This choice ensures a fair comparison with prior work. DenseNet-121 is a well-established architecture for medical image analysis, known for its dense connectivity, which promotes feature reuse and gradient flow, making it highly effective for extracting nuanced medical features.
- MedImageInsight from Microsoft, employs a dual-encoder architecture inspired by Contrastive Language-Image Pre-Training (CLIP), using separate encoders for images and text. Both encoders are trained using contrastive learning to map inputs into a shared embedding space, facilitating tasks like image-text matching and zero-shot classification. It is trained on a vast and diverse dataset of text-image pairs.
- RAD-DINO from Microsoft, based on the Distillation with No Labels (DINO) self-supervised learning framework, using a vision transformer (ViT) architecture. It employs a teacher-student setup where both networks process different augmented views of the same image, enabling the model to learn meaningful representations without labeled data. It is specifically designed for extracting embeddings from medical imaging.

For the second part of our experiments, we retrain our own CNN DenseNet-121 model from scratch. This allow us to have control on the splits we use and on the number of outputs (only outputting the medical conditions we want to study). This is more fair. We initialize the weights with the ones from CheXNet [9], a DenseNet121 model trained to

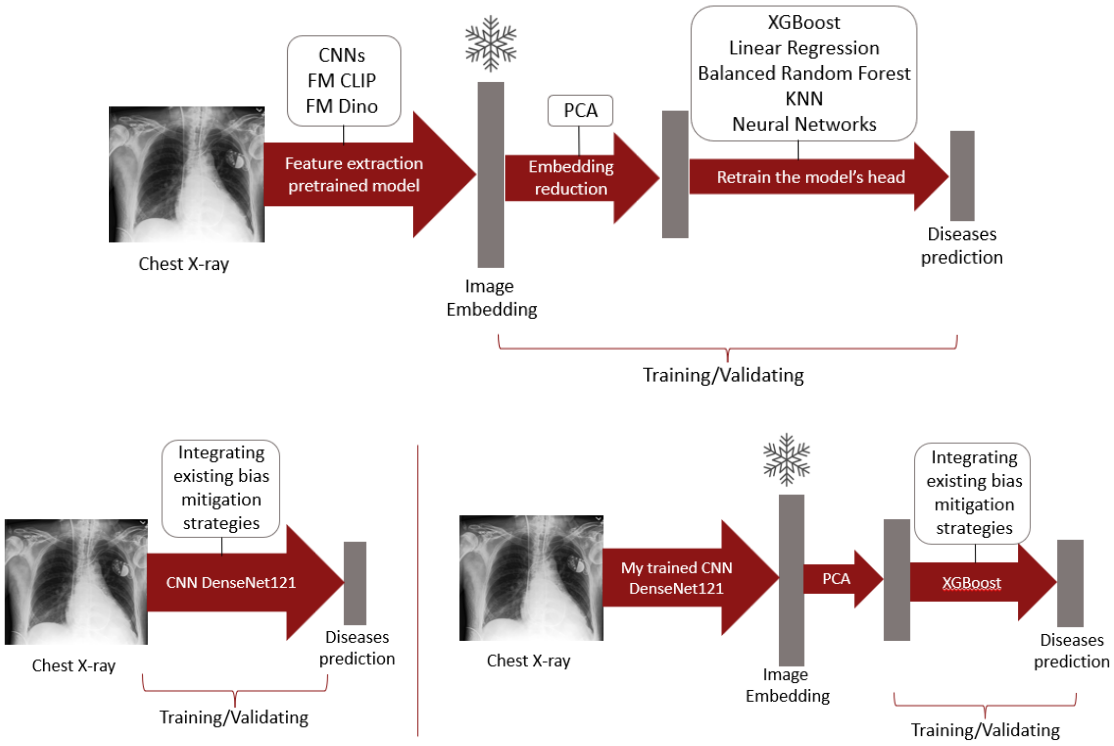

Figure 2: Overall pipeline of the proposed experiments. Top: Extending the CNN-XGboost hybrid bias mitigation method to other model architectures, other models to retrain the head, and more medical conditions. Bottom: Train our own DenseNet121 model to integrate existing bias mitigation techniques and compare the results when inegrating the same techniques when only retraining the head with a XGBoost or a LR model.

predict pneumonia from CXR (note that we our study does not involve pneumonia). The parameters used in training are similar to the ones used by TorchXrayVision [4]: Learning rate: 1e-4, Num epochs: 10, Batch size: 16, Criterion: BCEWithLogitsLoss, Optimizer: Adam, Scheduler: ReduceLROnPlateau, Training transformation: RandomRotation(10), ColorJitter(brightness=0.1, contrast=0.1), RandomHorizontalFlip(). We train it with early stopping.

4.3. Pipeline

The pipeline is presented in Figure 2.

1. To extend the current method to more than one medical condition, we changed the final XGBoost classifier to a multi-head classifier of the size of the number of medical conditions we want to study. This number is determined by the original model performance. We select medical conditions that can be predicted by the model with a performance above a certain threshold. Indeed, it does not make sense to study the bias if the original disease classification performance is low for every subgroup.

2. To retrain the head of the CNN with different models, we replace the XGBoost model of the MultiOutputClassifier by Linear Regression (LR), Decision Tree (DT), Random Forest (RF), Neural Networks (NN), K Nearest Neighbors (KNN), and Balanced Random Forest (BRF). For each model, we run some hyperparameter tuning on the validation set to select the best parameters.

3. To extend the current CNN-XGBoost method to other models than can extract features from images such as the CLIP based and Dino based FM, we extract embeddings from these pretrained models and retrain the head of the models using an XGBoost multi-head classifier. We then compare the performance and bias in the original full model with our retrained new version.

4. To compare existing bias mitigation strategies with the XGBoost head retraining, we first retrain a CNN from scratch, to have more control on the model. Then we retrain it with existing bias mitigation techniques and compare the results with our XGBoost head retraining results. Existing bias mitigation techniques include weighted sampling (we re-weight the subgroups according to their proportion in training data); adver-

sarial training (we add a secondary adversarial branch that tries to predict the sensitive attribute sex, race, age and train the main network to be "demographically agnostic" by minimizing this branch's accuracy), data augmentation (on subgroups that don't perform well on validation data), and active learning (we use uncertainty sampling or diversity-based selection to preferentially add underrepresented samples during model training).

5. Finally, we combine these exiting bias mitigation strategies with the XGBoost head retraining and compare this computationally efficient results with using these exiting bias mitigation on the full model retraining.

5. Experiments, results, and discussion

1. First, we extended the current method to more than one output. Instead of only focusing on Pleural Effusion as in the initial paper [3], we integrated Cardiomegaly, Lung Opacity, and Edema in the analysis. The studied medical conditions were chosen based on their performance: their AUPRC should be greater than 50% to be integrated in the analysis. Results 3 show that the method adapts well to multiple outputs: overall performance increases will bias according to the sex, age, and race subgroups decreases, ID on CheXpert and OOD on MIMIC.

2. Then, we retrained the head of the CNN with different models such as LR, DT, RF, NN, KNN, and BRF. Results 4 show that XGBoost has the best trade-off between performance and bias, closely followed by LR. It is not surprising that LR is working well since the last layer of a DenseNet121 is usually a linear layer. Therefore, the extracted embeddings are trained to be combined with a linear layer. Moreover, last layer retraining has been proven to improve performance and reduce bias, even ID[8]. DT and RF, being as good as random answers, were not included in the results. BRF is also working well, certainly due to its nature at handling imbalance datasets, which can be beneficial in handling bias. One interesting observation is that the different models don't have the same bias reduction pattern across the subgroups. For example, LR is working well in reducing bias for sex and age but not for race, while XGBoost and BRF are more consistent across the subgroups. This can be due to the data imbalanced: races are much more imbalanced than sex and age. LR, by its simple nature, might not be able to handle that as well as other models that are known to work well at handling imbalance datasets.

3. We then tried to generalize the hybrid CNN-XGBoost model to other model architectures. Since FM are more and more studied, we chose two FM with different architectures: CLIP, and RAD-DINO. Since XGBoost and LR models to retrain the head of the network seem to work the best, we only studied bias mitigation using these two models. The RAD-DINO model has no baseline, since it was trained to extract embeddings. Thus, we can only compare the difference between training the head of the model using a LR or a XGBoost model. Results, as shown in Figure **??** are a bit surprising. Indeed, there is no clear pattern. First, in most cases, retraining the head of the models improve the overall performance, with no major differences between retraining with XGBoost or retraining with LR. On embeddings extracted by DenseNet, retraining the head with XGBoost reduces the bias, both ID and OOD, for sex, age, and race (except for race OOD). XGBoost is a bit better than LR but results are not significant. On embeddings extracted by MedImageInsight, results are very different on the two datasets. Bias is decreasing for age on both datasets. On CheXpert bias is increasing for race and sex, and on MIMIC, bias is decreasing for sex and not changing for race. Again, there is no major difference between LR and XGBoost. On embeddings extracted by RAD-DINO, results are also different on the two datasets. Bias is smaller when training with LR for race, similar for LR and XGBoost for age and sex on CheXpert, and smaller when training with XGBoost on MIMIC. We cannot draw any conclusion from these results, no statistically significant pattern emerges. To better understand these differences, we explored the differences in the embeddings extracted by the three different models using Principal Component Analysis (PCA) density and (t-distributed Stochastic Neighbor Embedding[11] (t-SNE) plots. We can see that MedImageInsight and RAD-DINO embeddings have more distinctions according to the different subgroups 5. We also compared the performance of a RF classifier to predict the subgroups. Results 5 show that it is easier to predict subgroups using the RAD-DINO embeddings, closely followed by MedImageInsight embeddings. CNN subgroup predictions always result in the lowest performance. We compared these results with the bias of the three models those heads have been retrained using XGBoost. We hypothesized that embeddings that can easily predict some subgroups would ne more biased related to these subgroups. We cannot conclude that from the results, since again, there is not clear pattern.

4. Then, we retrained a CNN DenseNet121 from scratch. We then reassessed the overall performance and bias ID and OOD, with and without the XGBoost head retraining. Results match previous results with the pre-

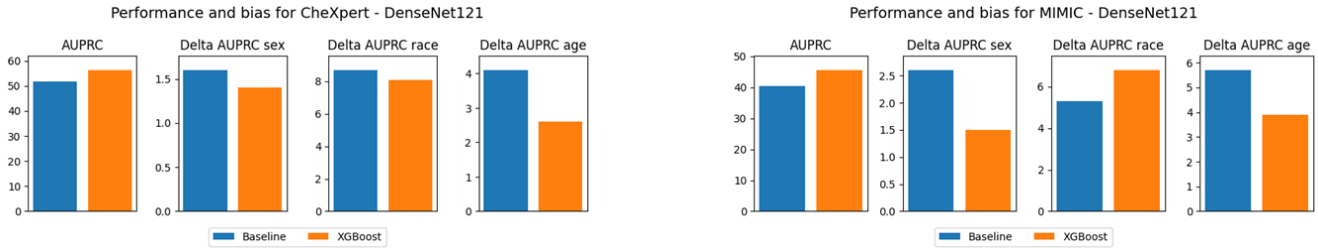

Figure 3: Performance and bias between the baseline model and the model with the retrained head with an XGBoost, ID (left) and OOD (right).

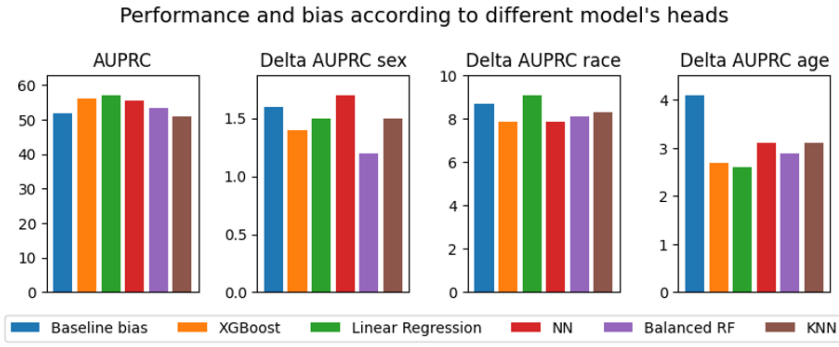

Figure 4: Differences in performance and bias when retraining the head of the DenseNet121 with different models on CheXpert.

trained TorchXRayVision model. We notice an increase in the overall performance, and a decrease in bias when retraining the head with an XGBoost model in comparison to the baseline. Then we compare the XGBoost head retraining computational efficient way of mitigating bias, with the existing bias mitigation that require the full model retraining. We can see in Fig 6 that both ID and ODD, XGBoost head retraining is as good (and even better) than existing bias mitigation techniques that require the full model retraining. Moreover, it has a lower computational cost since it only requires to retrain a single layer.

5. Finally, we combined existing bias mitigation methods with the XGBoost head retraining and compare that with retraining the full model. We can see in Fig 7 that combining existing bias mitigation techniques with the XGBoost head is better than retraining the full model. And again, it has a much lower computational cost.

   The final results show that combining active learning with XGBoost head retraining leads to the highest decrease in bias among sex, age, and race, ID and OOD.

6. Conclusion and Future Work

In this study, we presented a lightweight bias mitigation approach for DL models in CXR analysis. By retraining the final classification layer with an XGBoost model, our method effectively reduced biases related to sex, age, and race without requiring full model retraining. Our approach demonstrated robustness across multiple medical conditions and datasets, maintaining performance both ID (CheXpert) and OOD (MIMIC). Additionally, our method outperformed traditional bias mitigation techniques, including data augmentation, active learning, and adversarial training. Combining our XGBoost head retraining with active learning resulted in an optimal balance between fairness and diagnostic accuracy.

Future work includes focusing on interpretability. We want to find out why do we observe these results. We will focus more on where does the bias come from. Experimental ideas include studying linear probing from previous CNN layers to find out at which step the model captures information linked to the specific medical conditions and the demographic subgroups; studying why the different classification heads have different bias reduction across subgroups; studying the impact of different subgroup balancing in the data on the bias.

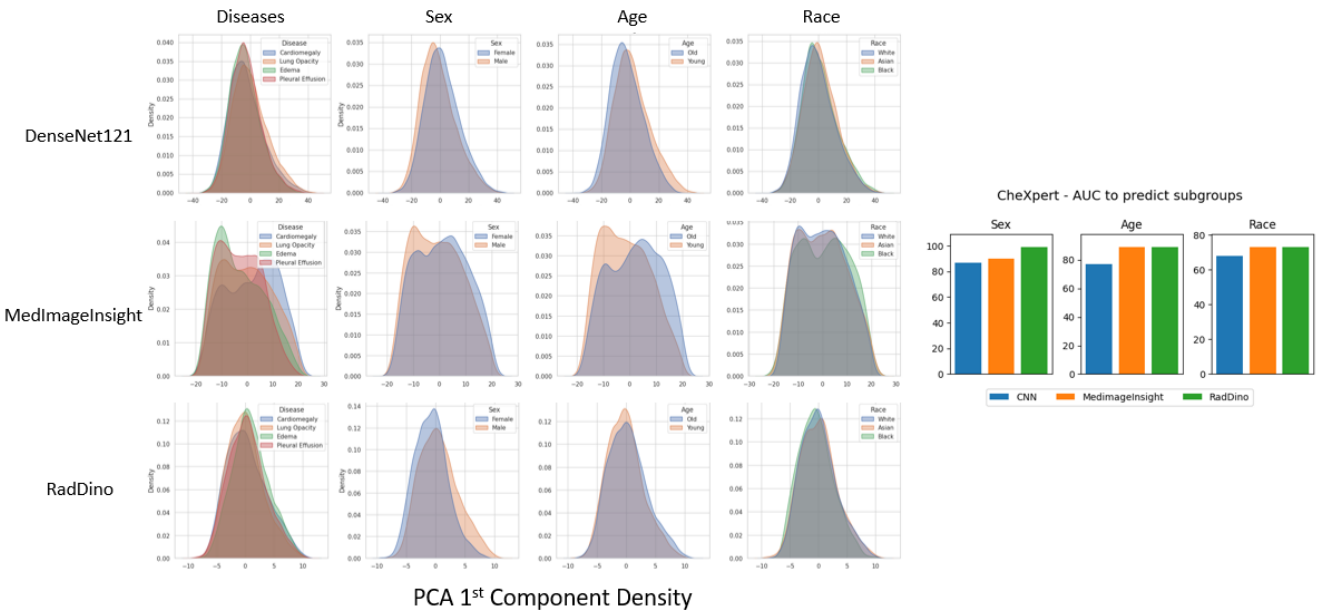

Figure 5: Left: 1st Component Density extracted from PCA according to the different subgroups for the different models to extract embeddings. Right: Performance of a RF classifier in predicting the subgroups from the embeddings extracted by the three different models.

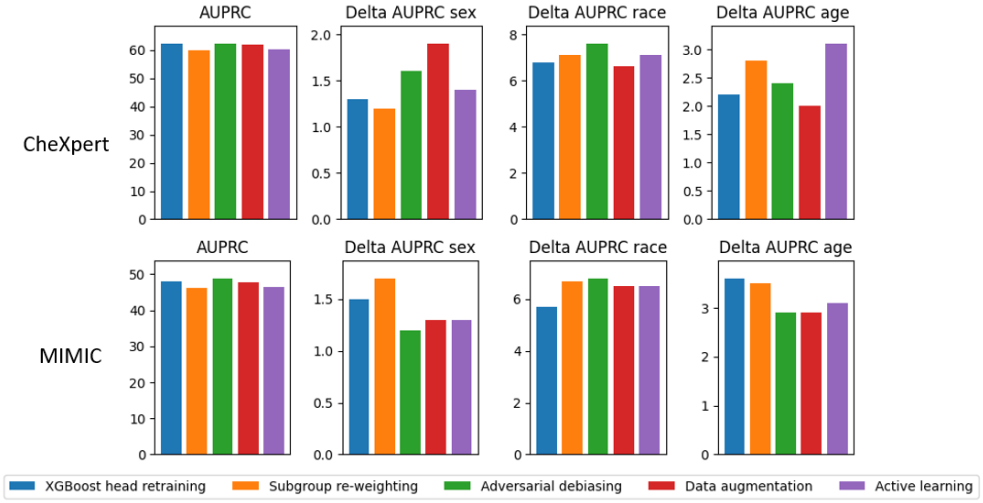

Figure 6: Comparison of our new lightweight bias mitigation method (in bleu) with existing methods that require full model retraining.

Acknowledgment

Thanks to the Center for Artificial Intelligence in Medicine and Imaging for their support and insights throughout this project. Special thanks to my PI, Professor Curtis Langlotz, as well as the PhD students and postdoc Louisa Fay, Maya Varma, and Sophie Ostmeier for their guidance and feedback.

References

[1] S. Alowais, S. Alghamdi, and N. Alsuhebany. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Med Educ*, 2023. 1

[2] T. S. S.-C. H. Z. C. M. V. S. Q. T. C. T. C. C. P. L. Chambon, Jean-Benoit Delbrouck. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demo-

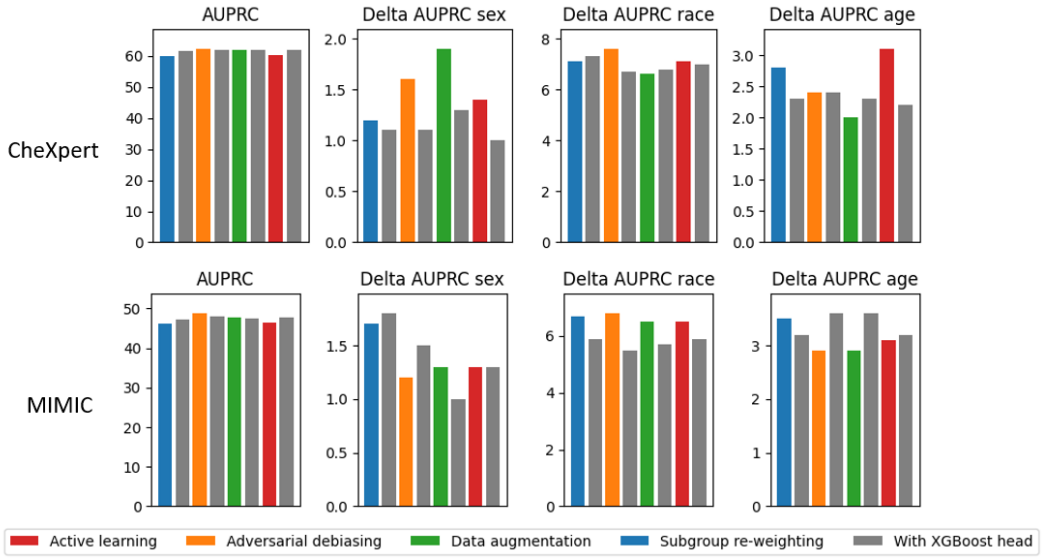

Figure 7: Comparison of existing bias mitigation methods that require full model retraining with combine the method with our XGBoost head retraining (in grey).

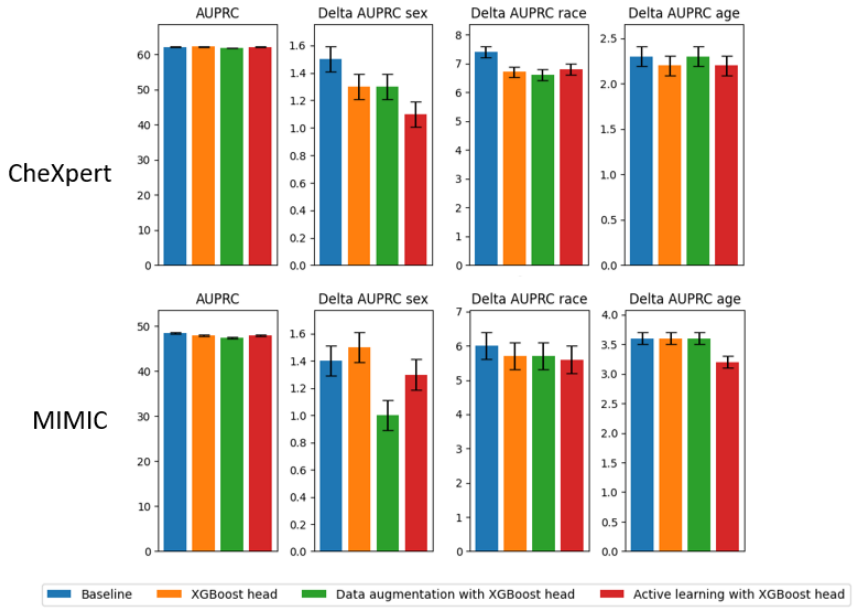

Figure 8: Comparison of initial performance and bias (baseline in blue) with XGBoost head retraining (orange), XGBoost head retraining combined with data augmentation (green) and active learning (red), ID on CheXpert and OOD on MIMIC.

graphics and additional image formats. AAAI Press, 2024. 2

[3] J.-B. D.-C. L. Clemence Mottez, Louisa Fay. Lightweight model adaptation for mitigating bias in deep learning models for chest x-ray analysis. 2025. 1, 4

[4] J. Cohen, J. Viviano, P. Bertin, P. Morrison, P. Torabian, M. Guarrera, M. Lungren, A. Chaudhari, R. Brooks, M. Hashir, and H. Bertrand. Torchxrayvision: A library of chest x-ray datasets and models, 10 2021. 3

[5] Y. J.-S. G. Gianfrancesco, Tamang S. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med*, 2018. 1

[6] Y. Hedhoud, T. Mekhaznia, and M. Amroune. An improvement of the cnn-xgboost model for pneumonia disease clas-

sification. *Polish Journal of Radiology*, 88:483–493, 2023. 1

[7] P. T. B.-S. e. a. Johnson, A.E.W. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data, 2019. 2

[8] P. Kirichenko, P. Izmailov, and A. G. Wilson. Last layer re-training is sufficient for robustness to spurious correlations, 2023. 1, 4

[9] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning, 2017. 2

[10] E. Sugiharti, R. Arifudin, D. Wiyanti, and A. Susilo. Integration of convolutional neural network and extreme gradient boosting for breast cancer detection. *Bulletin of Electrical Engineering and Informatics*, 11:803–813, 04 2022. 1

[11] H. Van der Maaten. Visualizing data using t-sne. 2008. 4

[12] S. M.-e. a. Wiens, Saria S. Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine*, 2019. 1

[13] Y. Yang, H. Zhang, D. Katabi, and M. Ghassemi. On mitigating shortcut learning for fair chest x-ray classification under distribution shift. In *NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models*, 2024. 1

[14] G. J. e. a. Yang Y., Zhang H. The limits of fair medical imaging ai in real-world generalization. *Nature Medicine*, 2024. 1