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

Clemence Mottez[1,2] CMOTTEZ@STANFORD.EDU
[1] *Center for Artificial Intelligence in Medicine and Imaging, Stanford University*
[2] *Institute of Computational and Mathematical Engineering, Stanford University*
Louisa Fay[1] LFAY@STANFORD.EDU
Jean-Benoit Delbrouck[1] JBDEL@STANFORD.EDU
Curtis Langlotz[1] LANGLOTZ@STANFORD.EDU

Editors: Under Review for MIDL 2025

Abstract

Deep learning (DL) models have demonstrated significant potential in improving chest X-ray (CXR) diagnosis. However, these models may exacerbate healthcare disparities. Addressing the inherent biases of DL models is essential to ensure their safe and reliable deployment in clinical practice. We suggest a novel bias mitigation approach that combines embeddings extracted by a Convolutional Neural Network (CNN) with an eXtreme Gradient Boosting (XGBoost) classifier. Our results show that this hybrid model significantly reduces bias across the sensitive attributes sex, age, and race, while maintaining comparable overall diagnostic performance and without the need for expensive model retraining. Our approach demonstrates that integrating simple, interpretable, and computationally efficient modifications into existing models can effectively enhance fairness in medical imaging.

Keywords: Bias Mitigation, Chest X-ray, Convolutional Neural Network, eXtreme Gradient Boosting

1. Introduction

DL models have the potential to transform healthcare by increasing diagnostic accuracy, personalizing treatment, and improving patient outcomes(Alowais, 2023). However, these technologies risk exacerbating healthcare disparities if their performance varies among different subgroups of patients, for example according to sex, age, and race (Yang Y., 2024). 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(Gianfrancesco, 2018). Biases are among the many barriers that prevent the deployment of these models in clinical practice, where equitable outcomes are crucial(Wiens, 2019). Current bias mitigation methods involve tradeoffs between fairness and accuracy. Techniques such as rebalancing training datasets or modifying algorithms often require extensive model retraining(Yang et al., 2024) and are thus impractical in healthcare due to data scarcity and resource constraints. In response, our study proposes a novel lightweight model adaptation strategy to mitigate biases related to sex, age, and race in CXR diagnosis. Our solution is to replace the final classification layer of CNNs with an XGBoost model, which is then retrained on a curated subset of data. This hybrid approach leverages CNNs' feature extraction capabilities and XGBoost's effectiveness in

© 2025 CC-BY 4.0, C. Mottez, L. Fay, J.-B. Delbrouck & C. Langlotz.

handling class imbalances to reduce bias. Previous studies have shown better classification performance when the last layer of CNNs is replaced by an XGBoost classifier(Shanmugam, 2023; Sugiharti et al., 2022; Hedhoud et al., 2023), but none have explored this combination to mitigate bias. Our hybrid CNN-XGBoost framework demonstrates that equitable diagnostics are achievable without sacrificing accuracy, reducing bias across sex, age, and race subgroups by 62.6% while improving the overall performance by 8.9%.

2. Method

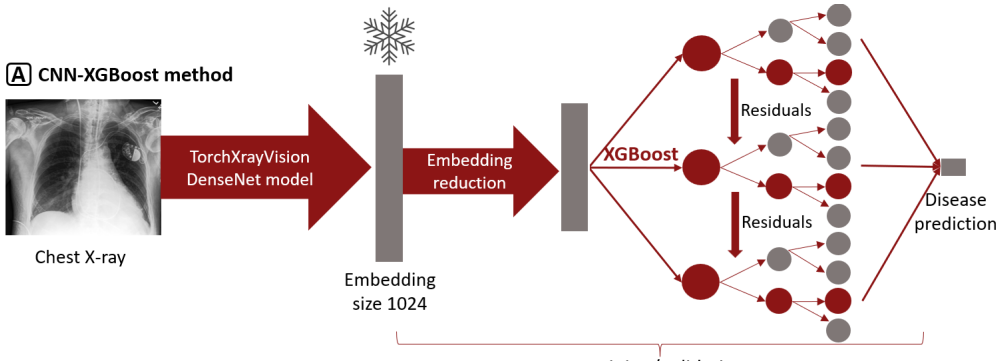
We use the DenseNet-121 model from TorchXrayVision(Cohen et al., 2021) pretrained on the CheXpert(Irvin et al., 2019) dataset to encode each X-ray image. For each image, we first extract the last hidden layer of the CNNs, resulting in a 1024-dimensional embedding. Next, we reduce their dimension with reduction techniques including Principal Component Analysis (PCA) and encoder-decoder architectures. We select the method that maximizes performance and minimizes bias on the validation data. The resulting vector serves as input for an XGBoost classifier, which is then trained. XGBoost was chosen due to its ensemble nature. Multiple trees are trained to correct the errors of the previous trees, inherently focusing on harder-to-classify examples such as underrepresented groups. We expect this integration to both reduce the performance gap among subgroups and increase the overall performance. The pipeline is presented in Figure 1A.
This method is model-agnostic, which means that it can be adapted to other model architectures designed for image feature extraction. Moreover, existing bias mitigation techniques, including adversarial training, sample weighting, and data augmentation(Yang et al., 2024) can easily be integrated into our framework by only retraining the XGBoost head.

3. Experiments

Datasets: We evaluate our method on two datasets to ensure the robustness and generalization of our model across different clinical environments. First, in-distribution data from CheXpert(Chambon, 2024), a dataset consisting of 224,316 CXRs obtained at Stanford Health Care. Second, Out-Of-Distribution (OOD) data from Medical Information Mart for Intensive Care (MIMIC)(Johnson, 2019) comprising 377,110 CXRs performed at the Beth Israel Deaconess Medical Center. Detailed dataset information are presented in Figure 1B.
Embedding Analysis: We first analyzed the embeddings extracted with DenseNet using PCA, t-distributed Stochastic Neighbor Embedding (t-SNE)(Van der Maaten, 2008) plots, and statistical tests such as the two-sample Kolmogorov–Smirnov test. Results, as shown in the Appendix Figure 2, indicate significant differences across subgroups, suggesting that the model could use shortcuts for disease classification, potentially leading to biased results.
Embeddings reduction: Results in the Appendix Table 1 show that reducing the size of the embedding using PCA to select the components that retain 95% of the total variability leads to a larger decrease in bias while maintaining a competitive overall performance.
Bias Mitigation: We focused the analysis on pleural effusion, due to its clinical significance and prevalence in the datasets. We used the Area Under Precision-Recall Curve (AUPRC) as the primary performance metric, due to its effectiveness in imbalanced data and in balancing precision and recall. We assessed the presence of bias related to the sub-

groups using ΔAUPRC. 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. Each experiment is run 5 times and results are averaged. The XGBoost training parameters are described in the Appendix 4. As shown in Figure 1C, the original CNN model leads to higher performance differences between the subgroups than our novel approach. By incorporating XGBoost, the model leads to fairer and more consistent results. Results show a decrease in bias of 79.2% for sex, 47.1% for age, and 61.7% for race while improving the overall performance by 8.9%.

OOD adaptability: We test our method, using a DenseNet model trained on the CheXpert dataset, on the MIMIC dataset to verify performance and bias mitigation consistency OOD. While the classification performance on the MIMIC dataset is lower than on the CheXpert dataset, our results demonstrate that our approach generalizes well OOD, reducing overall bias by 39.2% and improving overall performance by 8.2%.



B Datasets ratio

	Train-valid-test size	Pleural Effusion ratio
CheXpert	67,263-4,484-40,358	39.4
MIMIC	114,976-1,231-42,781	31.8

	Sex ratio (F-M)	Age ratio (Y-O)	Race ratio (W-B-A)
CheXpert	44.8-55.2	44.4-55.6	80.6-15.5-3.9
MIMIC	42.0-58.0	62.9-37.1	78.2-7.1-14.7

C Results

	AUPRC	ΔAUPRC sex	ΔAUPRC race	ΔAUPRC age
Existing bias CheXpert	64.9	2.4	23.5	6.8
CNN-XGBoost	70.7 ↑ 8.9%	0.5 ↓ 79.2%	9.0 ↓ 61.7%	3.6 ↓ 47.1%
Existing bias MIMIC	53.5	4.9	14.7	10.9
CNN-XGBoost	57.9 ↑ 8.2%	1.9 ↓ 61.2%	10.6 ↓ 27.9%	7.8 ↓ 28.4%

Figure 1: (A) Our CNN-XGBoost pipeline, (B) datasets information, (C) results (AUPRC: higher is better ; ΔAUPRC: lower is better).

4. Conclusion and Future Work

This hybrid bias-reduction method improves performance and mitigates bias related to sex, age, and race in pleural effusion prediction from CXR images. By using a model-agnostic approach, the integration can be applied to existing CNN models without the need for retraining, which is particularly beneficial for already trained models. Future work includes extending this analysis to other model architectures, such as Vision Transformers and Foundation Models, analyzing additional chest medical conditions, and combining other bias mitigation strategies with our method.

References

Alghamdi S.S. Alsuhebany N. et al. Alowais, S.A. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Med Educ*, 2023. doi: 10.1186/s12909-023-04698-z.

Thomas Sounack Shih-Cheng Huang Zhihong Chen Maya Varma Steven QH Truong Chu The Chuong Curtis P. Langlotz Chambon, Jean-Benoit Delbrouck. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. AAAI Press, 2024. URL https://doi.org/10.48550/arXiv.2405.19538.

Joseph Cohen, Joseph Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, and Hadrien Bertrand. Torchxrayvision: A library of chest x-ray datasets and models, 10 2021.

Yazdany J Schmajuk G. Gianfrancesco, Tamang S. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med*, 2018. doi: 10.1001/jamainternmed.2018.3763.

Yousra Hedhoud, Tahar Mekhaznia, and Mohamed Amroune. An improvement of the cnn-xgboost model for pneumonia disease classification. *Polish Journal of Radiology*, 88: 483–493, 2023. ISSN 1899-0967. doi: 10.5114/pjr.2023.132533. URL https://doi.org/10.5114/pjr.2023.132533.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. AAAI Press, 2019. doi: 10.1609/aaai.v33i01.3301590. URL https://doi.org/10.1609/aaai.v33i01.3301590.

Pollard T.J. Berkowitz S.J. et al. Johnson, A.E.W. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data, 2019. URL https://doi.org/10.1038/s41597-019-0322-0.

Vidyasri Shanmugam, Saravanan. A conv-xgb dnn for the detection of lung disease on chest x-ray images using transfer learning. 04 2023.

Endang Sugiharti, Riza Arifudin, Dian Wiyanti, and Arief 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. doi: 10.11591/eei.v11i2.3562.

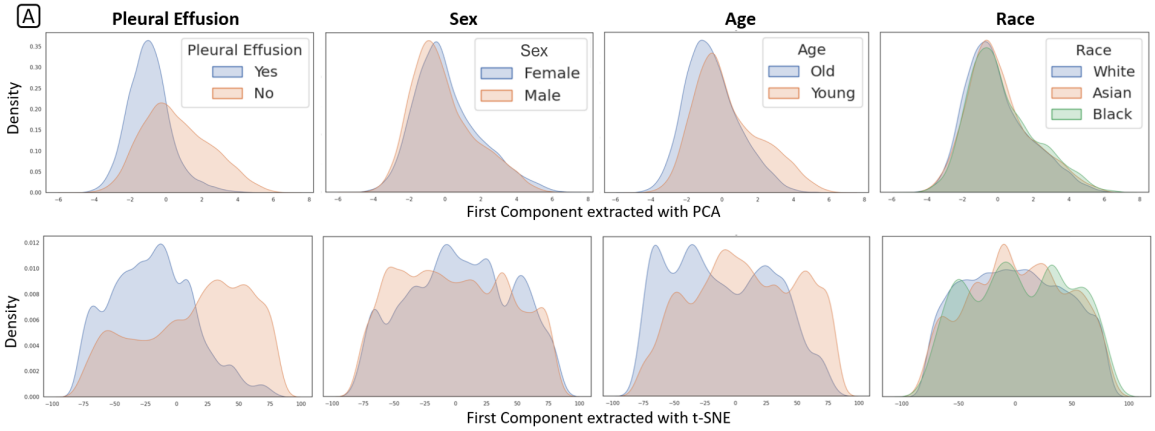
Hinton Van der Maaten. Visualizing data using t-sne. 2008. URL https://doi.org/10.1038/s41597-019-0322-0.

Sendak M et al. Wiens, Saria S. Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine*, 2019. ISSN 1337-1340. doi: 10.1038/s41591-019-0548-6.

Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh 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. URL https://openreview.net/forum?id=ar9IclPk8O.

Gichoya J.W. et al. Yang Y., Zhang H. The limits of fair medical imaging ai in real-world generalization. *Nature Medicine*, 2024. ISSN 2838–2848. URL https://doi.org/10.1038/s41591-024-03113-4.

Appendix



	Sex	Age	Race
B Statistical test p-values			
Kolmogorov–Smirnov	3.6e-78	2.45e-118	6.9e-20
Kruskal Wallis	3.7e-21	1.06e-32	6.9e-20
C Random Forest Classifier AUC	0.86	0.78	0.90

Figure 2: (A) Densities of the embeddings reduced using PCA (top row) and t-SNE (bottom row) according to the different values of pleural effusion and the sensitive attributes, (B) statistical significance (p-values) of the difference between subgroups for all sensitive attributes obtained from different statistical tests, (C) Area Under the Curve (AUC) of a Random Forest Classifier evaluating the ability of embeddings to predict the sensitive attributes.

XGBoost Hyperparameters: The hyperparameters were selected based on model performance evaluated on the validation dataset. The optimized hyperparameters are as follows: `eval_metric = 'logloss'`, `learning_rate = 0.05`, `n_estimators = 150`, and `max_depth = 10`.

Table 1: Classification performance and bias on validation set when reducing the image embeddings using an Encoder-Decoder architecture, a PCA framework, or keeping the original embedding size.

Encoder-Decoder Output dimension	**AUPRC**	**ΔAUPRC sex**	**ΔAUPRC race**	**ΔAUPRC age**
64	70.8	3.2	19.8	6.8
128	71.4	2.2	14.9	6.4
256	71.7	2.6	17.9	5.4
PCA Variance retained (in %)	**AUPRC**	**ΔAUPRC sex**	**ΔAUPRC race**	**ΔAUPRC age**
98	70.8	1.7	16.6	<u>4.8</u>
95	71.1	<u>1.2</u>	**14**	5.1
90	<u>71.9</u>	**0.7**	18	5.5
85	71.1	3.2	<u>14.7</u>	**4.2**
Original size	**AUPRC**	**ΔAUPRC sex**	**ΔAUPRC race**	**ΔAUPRC age**
	72.7	2	17.7	5.9