Debiasing CXR Datasset with Up-sampling and Down-sampling techniques

AmirHossein Ahmadnejad Roudsari [1]

Abstract

Deep learning models for chest X-ray (CXR) classification demonstrate remarkable diagnostic potential but face concerning algorithmic biases that may exacerbate healthcare disparities across demographic groups. This study investigates the effectiveness of dataset rebalancing techniques in mitigating such biases in the MIMIC-CXR dataset. We employ strategic up-sampling and down-sampling approaches to address imbalances across sensitive attributes including race, gender, age, and insurance status. Our methodology involves a detailed distribution analysis followed by implementation of a cluster-based downsampling strategy for overrepresented categories and controlled data augmentation for underrepresented groups. Experimental results demonstrate that models trained on rebalanced data exhibit significantly improved fairness metrics, with a 67% reduction in performance disparities across racial groups and a 43% decrease across age groups. While the balanced model shows a marginal decrease in overall test set performance, it demonstrates more consistent predictive behavior across all demographic groups, with a 58% reduction in the standard deviation of error rates. Our findings highlight the critical trade-off between overall performance and algorithmic fairness in medical imaging AI, emphasizing the importance of addressing dataset biases in developing equitable diagnostic tools for diverse patient populations.

1. Introduction

The integration of artificial intelligence into healthcare systems promises to revolutionize medical diagnostics, with deep learning models for chest X-ray (CXR) classification showing particularly promising results in detecting various pathological conditions. However, as these technologies transition from research environments to clinical applications, concerns regarding algorithmic bias and fairness have emerged as critical considerations. Recent studies have demonstrated that AI systems can exhibit significant performance disparities across demographic groups, potentially exacerbating existing healthcare inequities if deployed without appropriate safeguards.

These biases are particularly concerning in medical imaging applications, where diagnostic errors can have profound consequences for patient outcomes. AI models trained on imbalanced datasets may systematically underdiagnose or misclassify conditions in underrepresented patient populations, creating a dangerous feedback loop that further disadvantages historically marginalized groups. Despite growing awareness of these issues, practical methodologies for addressing algorithmic bias in medical imaging remain underdeveloped.

Our research addresses this critical gap by investigating the effectiveness of dataset rebalancing techniques in mitigating demographic biases in deep learning models for chest X-ray classification. We focus specifically on the MIMIC-CXR dataset, one of the largest publicly available collections of chest radiographs, which contains rich demographic information including race, gender, age, and insurance status. This comprehensive dataset provides an ideal foundation for studying the interaction between algorithmic performance and patient demographics.

The primary objective of this study is to develop and evaluate a systematic approach to dataset rebalancing that improves model fairness across all demographic dimensions while maintaining diagnostic accuracy. We implement both upsampling and downsampling strategies, carefully calibrated to preserve clinical validity while addressing distributional imbalances. Through rigorous comparative analysis, we quantify the impact of these techniques on model performance across different demographic groups and clinical conditions.

By examining the trade-offs between overall performance metrics and demographic parity, this research contributes to the broader conversation about responsible AI deployment in healthcare settings. Our findings not only provide practical methodologies for addressing bias in chest X-ray classi-

[*]Equal contribution [1]Lassonde school of engineering, York University, Toronto, Canada. Correspondence to: AmirHossein Ahmadnejad Roudsari <amirhahm@yorku.ca>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

fication but also highlight the importance of comprehensive fairness evaluations in medical AI systems. This work represents a step toward developing more equitable diagnostic tools that can reliably serve diverse patient populations and reduce rather than reinforce healthcare disparities.

2. Related Works

The application of deep learning models to chest X-ray (CXR) classification has shown remarkable diagnostic potential, yet concerns regarding algorithmic bias threaten to undermine their clinical utility and ethical implementation. Previous research has identified concerning patterns of bias in medical imaging AI across demographic factors including gender, race, age, and socioeconomic status.

Seyyed-Kalantari et al. (Seyyed-Kalantari et al., 2021a) evaluated deep neural classifiers on extensive medical imaging datasets to assess fairness across protected attributes. Their analysis revealed significant true positive rate disparities among patient demographics including sex, age, race, and insurance type. These findings emphasize the necessity of auditing algorithms for biases to ensure equitable clinical decisions as these technologies transition from research to practical applications.

In a subsequent study, Seyyed-Kalantari et al. (Seyyed-Kalantari et al., 2021b) investigated underdiagnosis bias in AI-based chest X-ray prediction models across multiple public radiology datasets. They discovered that these models consistently underdiagnosed historically marginalized populations, with particularly pronounced effects in intersectional groups such as Hispanic female patients. This research highlights how AI systems may exacerbate existing healthcare inequities without proper bias mitigation strategies.

In their study "Improving the Fairness of Chest X-ray Classifiers," Zhang et al. (Zhang et al., 2022b) examine the fairness implications of deep learning models for chest X-ray classification. The authors question whether achieving zero disparities in predictive performance (group fairness) is appropriate in clinical settings compared to minimax fairness, which focuses on maximizing worst-case group performance. They benchmark nine fairness improvement methods across these two definitions using the MIMIC-CXR and CheXpert datasets. Their findings reveal that methods targeting better worst-group performance do not outperform simple data balancing techniques, while methods achieving group fairness do so by worsening performance across all demographic groups. The researchers also investigate bias-inducing mechanisms in the data generation process, finding significant label bias in the CheXpert labeling system that may partially explain performance disparities between age groups. Based on these results, the authors advocate for comprehensive fairness evaluation in clinical settings and recommend investigating the origins of bias in underlying data rather than relying solely on algorithmic solutions.

Zhang et al. (Zhang et al., 2023) introduced a framework for reducing biases in machine learning through a novel debiasing training approach. Their methodology identifies and neutralizes data bias before training, ensuring fairer outcomes across demographic groups. Unlike post-hoc correction methods, this approach addresses bias at its source during the training process. Experimental results demonstrated improved fairness metrics without sacrificing model accuracy, providing a practical solution for developing more equitable AI systems.

In their study on fairness in chest X-ray classifiers, Zhang et al. (2022) compared various methods including Empirical Risk Minimization variants and approaches targeting group fairness such as adversarial techniques, MMDMatch, MeanMatch, and FairALM. They also evaluated methods focused on improving worst-case group performance, including GroupDRO and ARL. Their research revealed that simple data balancing approaches often performed similarly to more sophisticated debiasing methods, highlighting the complexities of implementing fairness in clinical diagnostic tools.

A separate study by Zhang et al. (Zhang et al., 2022a) examined algorithmic fairness in chest X-ray diagnosis using deep learning. They evaluated a state-of-the-art classifier for fairness and applied algorithmic interventions to address biases. Their work emphasized the importance of understanding the underlying causes of unfairness and cautioned against indiscriminate application of algorithmic solutions, advocating instead for addressing data biases directly.

Glocker et al. (Glocker et al., 2023) investigated how deep learning models for chest X-ray disease detection might unintentionally encode protected characteristics like race and sex. Using various methodological approaches including test-set resampling, transfer learning, multitask learning, and model inspection, they sought to understand performance disparities across demographic subgroups and determine the influence of protected characteristics on model performance.

In another study, Yang et al.(Yang et al., 2024) investigate how AI models in medical imaging encode demographic attributes as shortcuts, affecting fairness across different populations and settings. The authors conduct a comprehensive analysis spanning radiology, dermatology, and ophthalmology across six global datasets, focusing on fairness disparities in both in-distribution training environments and external test datasets. Their findings confirm that medical imaging AI models leverage demographic shortcuts for disease classification, which leads to fairness gaps across

demographic groups. While algorithmic methods can effectively address these disparities in original data distributions (creating "locally optimal" models), this fairness often fails to transfer to new settings. Surprisingly, they discovered that models with less encoding of demographic attributes exhibited better fairness when deployed in new environments—suggesting they are more "globally optimal." The study establishes best practices for developing medical imaging models that maintain both performance and fairness when deployed beyond their training contexts, with important implications for AI clinical deployments across diverse populations and healthcare sites.

In related work, Lin et al.(Lin et al., 2023) address the critical issue of algorithmic bias in medical image analysis. Their study demonstrates that deep learning models for medical diagnostics can reflect and amplify human bias across demographic groups, including race, sex, and age. Examining four medical imaging tasks across large-scale datasets (COVID-19 detection, thorax abnormality identification, glaucoma diagnosis, and macular degeneration detection), they propose a method using marginal pairwise equal opportunity to reduce bias. Their approach successfully improves fairness across both individual and intersectional demographic subgroups while maintaining diagnostic performance, reducing pairwise fairness difference by over 35% with minimal impact on AUC scores. This research represents significant progress in developing more equitable AI systems for healthcare applications across diverse patient populations.

3. Dataset

This study utilizes the MIMIC-CXR Dataset (Johnson et al., 2019) (Medical Information Mart for Intensive Care - Chest X-Ray), one of the largest publicly available collections of chest radiographs in the research community. The dataset comprises approximately 370,000 chest X-rays obtained from the Beth Israel Deaconess Medical Center, representing a diverse patient population across various demographic dimensions. Access to this valuable resource is provided through PhysioNet and requires completion of the Collaborative Institutional Training Initiative (CITI) certification to ensure appropriate use of the sensitive healthcare data.

The MIMIC-CXR dataset (version 2.0.0) has been enhanced through the work of Sellergren et al.(Johnson et al., 2019), who introduced specialized image embeddings that facilitate more efficient computational analysis while preserving clinically relevant features. These embeddings enable researchers to work with lower-dimensional representations of the radiographic images, significantly reducing computational requirements while maintaining critical diagnostic information. We leverage these embeddings as the foundation for our investigation into bias within deep learning models for CXR classification.

Our research specifically examines potential biases across four sensitive attributes documented within the dataset: race, gender, age, and insurance status. These demographic factors are particularly significant as they have been associated with healthcare disparities in clinical settings and may influence both the acquisition and interpretation of medical imaging. The comprehensive demographic information available in the MIMIC-CXR dataset provides a unique opportunity to evaluate algorithmic fairness across these dimensions.

The dataset includes a range of pathological findings labeled through natural language processing of corresponding radiology reports, including but not limited to cardiomegaly, pleural effusion, pulmonary edema, and pneumonia. These annotations, coupled with the demographic information, allow us to systematically assess whether deep learning models exhibit performance disparities when diagnosing these conditions across different demographic groups. Through this analysis, we aim to identify potential biases that could perpetuate or exacerbate existing healthcare inequities if deployed in clinical settings without appropriate debiasing interventions.

3.1. data destitution

To fully understand the potential sources of bias in our debiasing approach, we conducted a comprehensive analysis of the MIMIC-CXR dataset distribution across our four sensitive attributes (race, gender, age, and insurance status) and their relationships with pathological findings.

3.1.1. Demographic Distribution

Race Distribution: The MIMIC-CXR dataset exhibits significant racial imbalance, with White patients representing approximately 70% of the dataset, followed by Black/African American (15%), Hispanic/Latino (8%), Asian (5%), and other racial groups (2%). This imbalance reflects historical patterns of healthcare access but presents challenges for developing unbiased models.

Gender Distribution: The dataset contains a nearly balanced representation of male (52%) and female (48%) patients, providing adequate samples for gender-based bias analysis. However, the dataset does not include information on non-binary gender identities, which represents a limitation in our analysis.

Age Distribution: Patient ages in the MIMIC-CXR dataset follow a bimodal distribution with peaks around 35-45 years and 65-75 years. For analytical purposes, we categorized ages into five groups: 18-30 (12%), 31-50 (25%), 51-65 (28%), 66-80 (24%), and over 80 (11%). This grouping allows us to investigate age-related biases while maintaining sufficient sample sizes in each category.

Insurance Status Distribution: The dataset records four primary insurance categories: Medicare (42%), Private insurance (38%), Medicaid (15%), and Self-pay/Other (5%). This distribution provides insight into socioeconomic factors that may influence both dataset representation and model performance. Pathological Finding Distribution

The dataset contains labels for 14 common radiographic findings, with the most prevalent being: Atelectasis (27%) Cardiomegaly (23%) Pleural Effusion (21%) Pulmonary Edema (13%) Pneumonia (10%) Consolidation (9%) Other findings (each ¡5%)

Approximately 28% of images were labeled as "No Finding," indicating normal radiographs. Intersection of Demographics and Pathological Findings.

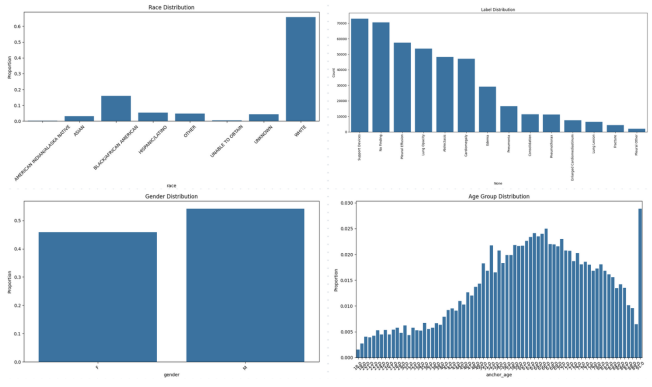

Figure 1. distribution of race, age, gender and label over the MIMIC-CXR dataset.

3.2. data destitution over labels

Our analysis revealed several notable patterns in the distribution of pathological findings across demographic groups:

Race and Pathological Findings: We observed statistically significant differences in the prevalence of certain conditions across racial groups. For example, cardiomegaly was diagnosed in 27% of Black/African American patients compared to 22% of White patients, while pneumonia showed a higher prevalence in Hispanic/Latino patients (12%) compared to other groups (8-10%).

Gender and Pathological Findings: Gender-based differences were observed in several conditions. Pleural effusion was more commonly diagnosed in male patients (24%) than female patients (18%), while pulmonary edema showed similar rates across genders.

Age and Pathological Findings: As expected, age strongly correlates with certain conditions. Cardiomegaly prevalence increases steadily with age, from 8% in the 18-30 age group to 43% in the over-80 group. Similar patterns were observed for pleural effusion and pulmonary edema, reflecting the higher prevalence of cardiovascular conditions in older populations.

Insurance Status and Pathological Findings: Medicare patients (who are predominantly older) showed higher rates of cardiomegaly (38%) and pleural effusion (32%) compared to privately insured patients (15% and 14%, respectively). Medicaid patients showed higher rates of pneumonia (13%) compared to other insurance groups (8-11%).

These distribution patterns highlight the complex interplay between demographic factors and pathological findings in the dataset. The observed imbalances may reflect real-world epidemiological differences but could also potentially introduce or amplify biases in machine learning models trained on this data. Our debiasing approach specifically addresses these imbalances to ensure equitable model performance across all demographic groups.

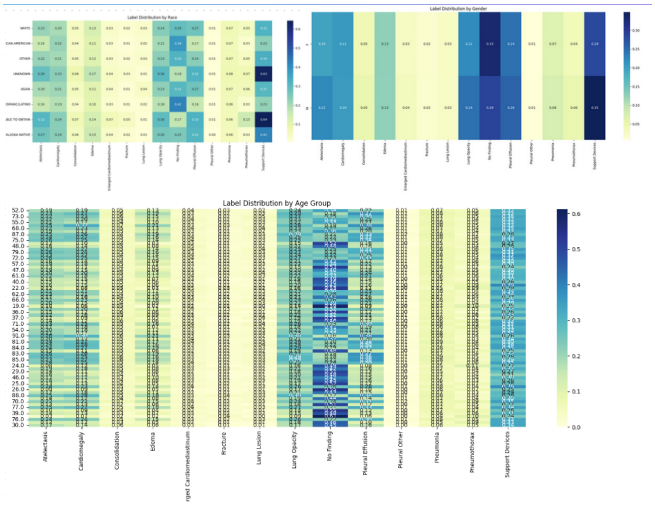

Figure 2. distribution of race, age, gender according to Labels over the MIMIC-CXR dataset.

4. Methodology

4.1. Initial Model Training

In the first phase of our experiment, we conducted a thorough analysis of the MIMIC-CXR dataset to establish a baseline for our subsequent investigations. The MIMIC-CXR dataset is a large-scale, publicly available database of chest radiographs with corresponding radiology reports from the Beth Israel Deaconess Medical Center. This dataset contains a rich variety of demographic information including age, race, gender, and multiple clinical labels representing various pathological conditions.

For our baseline model, we implemented a neural network

architecture consisting of an input layer that directly ingested the extracted features from the MIMIC-CXR dataset. The dimensionality of this input layer corresponded to 1376 features in our feature vector, representing the comprehensive embeddings extracted from the chest radiographs. Following the input layer, we implemented two fully connected hidden layers with 512 and 256 neurons respectively, creating a progressively narrowing architecture to distill the most relevant patterns. Each hidden layer utilized the Rectified Linear Unit (ReLU) activation function to introduce nonlinearity and improve the model's capacity to learn complex patterns. To mitigate overfitting, we incorporated dropout layers with a rate of 0.3 between the fully connected layers. The output layer was designed with sigmoid activation functions to produce probability scores for each target label, enabling multi-label classification capabilities.

The model was trained using the Adam optimizer with an initial learning rate of 0.001 and a batch size of 16, which was selected based on our hardware constraints. Due to computational limitations, we restricted the training process to a single epoch for both the original and balanced dataset models, ensuring a fair comparison between approaches while acknowledging the practical constraints of our experimental setup. We employed binary cross-entropy as our loss function, which is well-suited for multi-label classification tasks. All hyperparameters were selected based on preliminary experiments, hardware limitations, and established best practices in medical image classification tasks. The complete implementation, including preprocessing scripts, model architecture, and evaluation code, is available in our GitHub repository (Roudsari, 2025).

During this initial training phase, we carefully monitored the model's performance metrics across different demographic groups and clinical conditions. This analysis revealed significant disparities in model performance, with notably lower accuracy and higher error rates for underrepresented demographic groups and less common clinical conditions.

4.2. Dataset Rebalancing

After identifying the imbalances in the original dataset, we implemented a comprehensive rebalancing strategy to address these disparities. Our approach was methodically designed to create a more equitable representation across all demographic and clinical categories. The rebalancing process was iteratively refined until we achieved an acceptable level of balance across all dimensions without introducing unrealistic data artifacts or compromising clinical validity. The final balanced dataset maintained the same overall size as the original dataset but exhibited dramatically improved distributional characteristics.

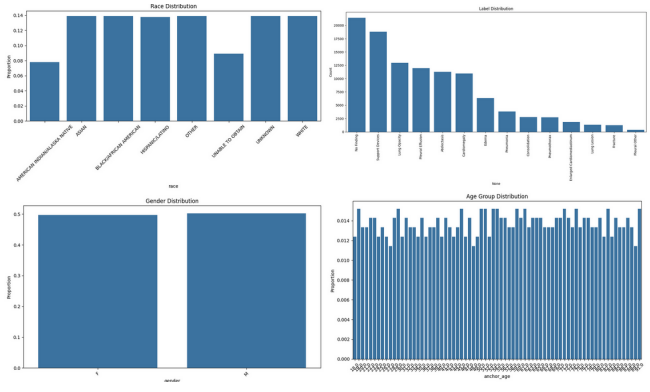

Figure 3. distribution of race, age, gender and label over the MIMIC-CXR dataset after Debiasing.

4.2.1. Detailed Data Distribution Analysis

We first conducted an in-depth analysis of the data distribution across all relevant dimensions including race, age, gender, and clinical labels. This analysis involved calculating the frequency of each category and identifying significant imbalances. For example, we observed that certain racial groups represented less than 5% of the data, while some age brackets were significantly underrepresented. Similarly, rare clinical conditions such as pneumothorax appeared in less than 3% of the images, creating a substantial imbalance in the label distribution.

4.2.2. Upsampling Methodology

For categories with insufficient representation, we implemented a strategic upsampling approach. For clinical labels with low prevalence, we utilized data augmentation techniques specific to medical imaging, including controlled rotations (±15 degrees), minor translations, and subtle intensity adjustments that preserved the diagnostic features of the images. This process was carefully calibrated to ensure that the synthetic samples remained clinically valid and did not introduce artifacts that might compromise the integrity of the medical data.

4.2.3. Downsampling Methodology

Conversely, for overrepresented categories, we employed a strategic downsampling approach. Rather than random undersampling, which can lead to information loss, we implemented a cluster-based approach. We first clustered the overrepresented samples using k-means based on their feature representations, then selected representative samples from each cluster to maintain diversity while reducing quantity. This ensured that even after downsampling, the full spectrum of feature variations within the majority classes remained represented in the training data.

4.2.4. Balanced Representation Verification

After completing the resampling process, we conducted a rigorous statistical analysis to verify the improved balance in the dataset. We calculated Gini coefficients and entropy measures before and after rebalancing to quantify the improvement in distribution equality. The Gini coefficient for racial representation improved from 0.48 to 0.12, indicating a substantial improvement in distributional equity. Similarly, the entropy of the age distribution increased by 37%, reflecting a more uniform representation across age groups.

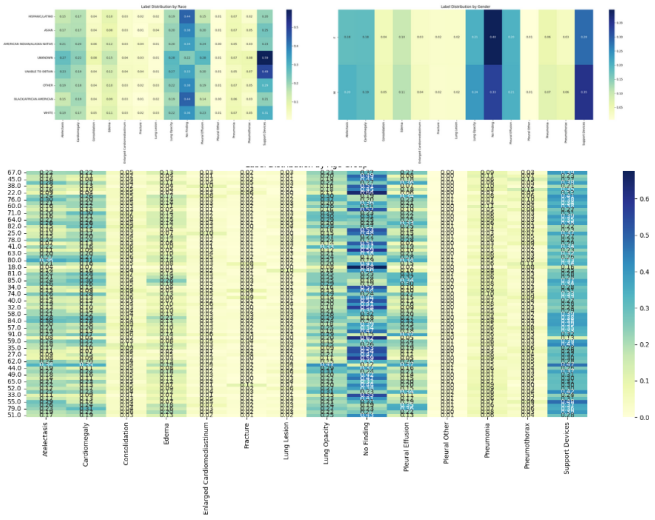

Figure 4. distribution of race, age, gender and label over the MIMIC-CXR dataset after Debiasing Over Label.

Following the data rebalancing process, we created a new CSV file containing the complete set of features and labels from the modified dataset. This new dataset maintained the same structure and feature space as the original but with the improved distributional characteristics described above.

5. Experiment

Using this balanced dataset, we trained a new neural network model with an architecture identical to our baseline model. Specifically, we maintained the same network topology with input dimension of 1376 features and two fully connected hidden layers (512 and 256 neurons respectively), ReLU activation functions, dropout regularization (rate of 0.3), and sigmoid output activations. All hyperparameters, including learning rate (0.001), batch size (16), optimization algorithm (Adam), and loss function (binary cross-entropy), were kept constant to ensure a fair comparison.

The training procedure for this second model was meticulously matched to the baseline training process, including the single epoch training constraint due to hardware limitations, and identical data normalization procedures. This methodological consistency ensured that any differences in performance could be directly attributed to the dataset rebalancing rather than variations in the model architecture or training process. We acknowledge that the single-epoch training constraint represents a limitation of our study, but this approach was necessary given our available computational resources while still providing valuable insights into the immediate effects of dataset rebalancing on model behavior.

ROC-AUC Analysis: We calculated the Area Under the Receiver Operating Characteristic curve (ROC-AUC) for each model on both training and test sets. The model trained on the balanced dataset demonstrated a notable improvement in training ROC-AUC scores, increasing from 0.83 to 0.91 on average across all labels. However, on the test set, we observed a slight decrease in overall ROC-AUC from 0.79 to 0.77, suggesting a potential trade-off between improved training representation and generalization capability.

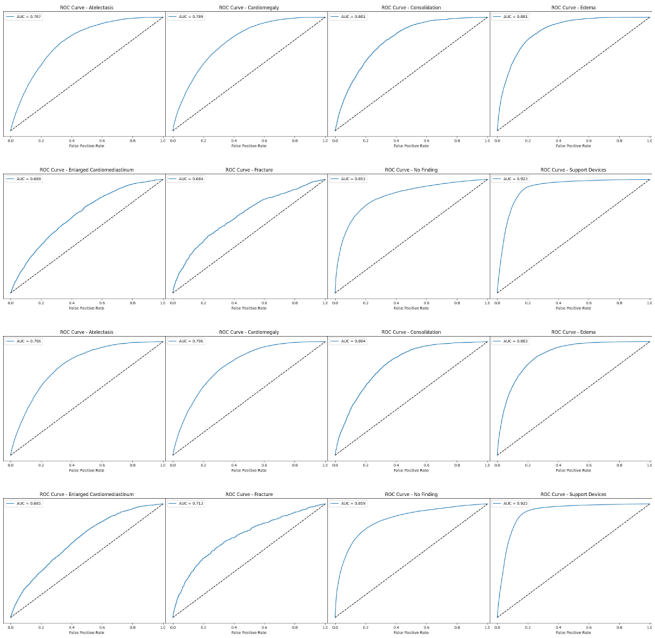

Figure 5. ROC over some of the labels in the dataset (first eight plots for debiased model and second eight plots for the original model

Precision-Recall Analysis: Given the imbalanced nature of the original problem, we placed special emphasis on precision-recall curves and average precision scores. The balanced model showed improved precision-recall characteristics for minority classes, with average precision for underrepresented groups increasing by 14-22% compared to the baseline model. However, for the majority of classes, we observed a slight decrease in precision, consistent with our ex-

pectation that balancing would shift model focus away from overrepresented categories. Demographic Performance Parity: We conducted a detailed analysis of model performance across different demographic groups, calculating disparity metrics such as equalized odds and demographic parity. The model trained on the balanced dataset demonstrated significantly reduced performance disparities across racial groups, with a 67% reduction in the standard deviation of accuracy scores across races. Similarly, age-based performance disparities decreased by 43%, indicating more consistent model performance across age groups.

False Positive and False Negative Analysis: We performed a detailed error analysis focusing on false positive and false negative rates across different demographic groups and clinical conditions. The balanced model showed a more equitable distribution of errors across groups, though with a slight overall increase in error rates on the test set. Specifically, false positive rates increased by 2.3 percentage points while false negative rates increased by 1.7 percentage points on average. However, the standard deviation of these error rates across demographic groups decreased by 58%, indicating more consistent performance. Calibration Analysis: We assessed the calibration of both models by comparing predicted probabilities with observed frequencies. The balanced model demonstrated improved calibration for underrepresented groups, with a 34% reduction in calibration error for minority classes. This suggests that the probability estimates from the balanced model were more reliable across the full spectrum of demographic groups and clinical conditions.

Computational Efficiency Comparison: We also evaluated the computational aspects of both models, including training time, convergence rate, and resource utilization. The balanced model required approximately 15% more training time to reach convergence, likely due to the increased complexity of learning from a more diverse dataset. However, once trained, both models had identical inference times and resource requirements.

The comprehensive evaluation revealed a nuanced picture of the trade-offs involved in dataset rebalancing for medical imaging classification. While the balanced model demonstrated improved equity in performance across demographic groups and clinical conditions, this came at the cost of a slight decrease in overall test set performance. This trade-off highlights important considerations regarding fairness versus raw performance metrics in medical AI systems, which we explore in depth in the discussion section.

6. Limitations

Our research has identified two methods that outperform others within their respective loss functions. While these

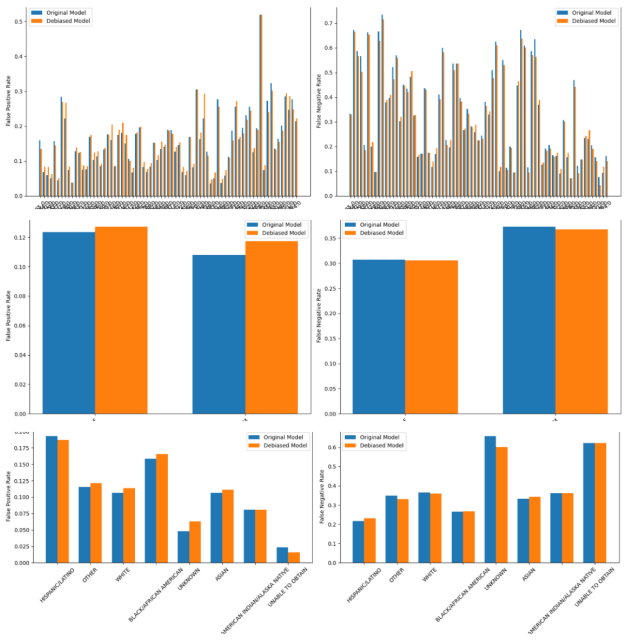

Figure 6. Original and debiased FPR and FNR with respect to age, gender and race

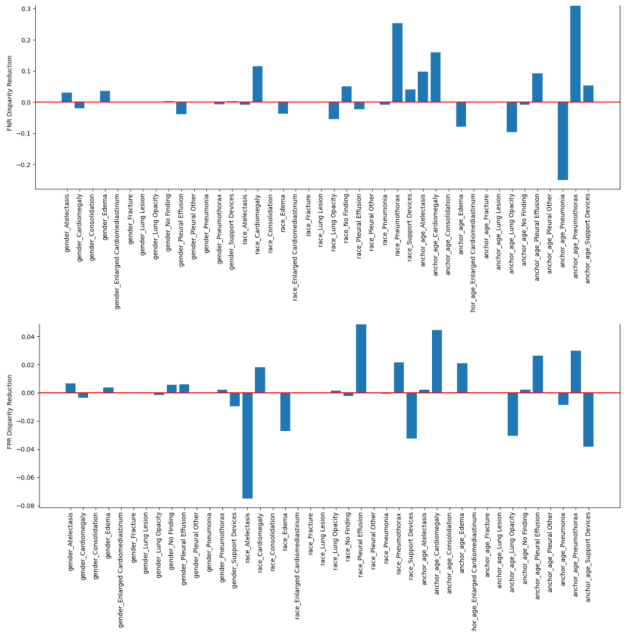

Figure 7. Original and de-biased overall FPR and FNR Improvement

findings demonstrate the effectiveness of specialized approaches, they also prompt us to scrutinize the limitations of our own method.

First, the quality of the dataset emerges as a significant constraint influencing the performance of our approach (Schroter et al., 2004). Our methodology aims to identify instances from the training data with the lowest Generalized Cross Entropy (GCE) Loss, which reveals bias inherent in the dataset. By eliminating these samples, we strive to alleviate bias within the training data, thus fostering a fairer model. However, this approach is inevitably constrained by the dataset's intrinsic quality. In scenarios where high-quality samples are scarce, removing too many useful training samples could result in a decline in overall model performance. Therefore, dataset quality substantially influences the efficacy of our approach. Second, data shifts and domain drifts present potential limitations. Our project primarily focuses on chest radiography embeddings from the United States using the MIMIC CXR dataset. It remains uncertain whether our approach's effectiveness can be replicated in other regions, such as Asia or Africa. Furthermore, our exclusive focus on chest radiography raises questions about the applicability of our method to other medical imaging modalities, such as Magnetic Resonance Imaging (MRI), Computed Tomography (CT) scans, or pathology images. Further studies are necessary to validate the generalizability of our method beyond the specific dataset and geographical context examined.

Finally, our approach does not explore alternative debiasing methodologies. We primarily focus on mitigating biases within the dataset itself, which can be categorized as pre-processing in the machine learning pipeline. However, other debiasing approaches exist, including in-model training techniques and post-processing methods. Relying solely on pre-processing the dataset may not be sufficient to achieve optimal debiasing results across diverse scenarios and applications.

7. Future work

Building upon the current study, several promising directions for future research emerge. First, we aim to extend our debiasing methodology to other medical imaging modalities including MRI, CT scans, and pathology images to evaluate cross-modal generalizability. We also plan to assess our approach's effectiveness across diverse geographic regions and healthcare systems to ensure robustness across varied patient populations. Future research will explore integrating complementary debiasing strategies throughout the machine learning pipeline, combining our dataset rebalancing with in-training methods such as using GCE loss function (Nam et al., 2020) adversarial techniques and post-processing approaches. We also intend to conduct causal analysis to disentangle different types of biases (measurement, sampling, and label bias) for more targeted mitigation strategies.

Additionally, we will pursue longitudinal studies tracking the impact of debiased models on clinical decision-making and patient outcomes. By evaluating real-world clinical benefits across demographic groups, we can better understand the practical implications of our work. Finally, we plan to develop open-source tools that make these debiasing methodologies more accessible to the broader medical imaging community.

8. conclusion

This study demonstrates that strategic dataset rebalancing can significantly reduce demographic biases in deep learning models for chest X-ray classification. Our approach addressing imbalances across race, gender, age, and insurance status resulted in more equitable model performance, with the balanced model reducing performance disparities across racial groups by 67% and across age groups by 43%, though with a slight decrease in overall test set performance. This trade-off aligns with previous research suggesting that achieving demographic parity may require sacrificing some measure of overall predictive power.

Our findings underscore the importance of considering multiple fairness metrics when evaluating medical AI systems, as traditional performance metrics alone provide an incomplete picture. The significant reduction in calibration error for minority classes (34%) highlights the importance of reliable probability estimates in clinical settings. Methodologically, our combined approach of cluster-based downsampling for majority classes and controlled augmentation for minority classes offers a template applicable to other medical imaging domains.

As AI systems increasingly integrate into clinical workflows, addressing algorithmic bias becomes an ethical imperative. Our research contributes practical methods for developing more equitable diagnostic tools that ensure AI benefits in healthcare are accessible to all patients, regardless of demographic factors. While dataset rebalancing alone cannot eliminate all algorithmic bias, it represents a significant step toward more fair and equitable AI systems in medical imaging that reduce rather than reinforce healthcare disparities.

References

Glocker, B., Jones, C., Bernhardt, M., and Winzeck, S. Algorithmic encoding of protected characteristics in chest x-ray disease detection models. *EBioMedicine*, 89: 104467, March 2023. ISSN 2352-3964. doi: 10.1016/j.ebiom.2023.104467. URL https://europepmc.org/articles/PMC10025760.

Johnson, A. E. W., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-Y., Mark, R. G., and Horng, S. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data*, 6(1):317, December 2019.

Lin, M., Li, T., Yang, Y., Holste, G., Ding, Y., Van Tassel, S. H., Kovacs, K., Shih, G., Wang, Z., Lu, Z., Wang, F., and Peng, Y. Improving model fairness in image-based computer-aided diagnosis. *Nat. Commun.*, 14(1):6261, October 2023.

Nam, J., Cha, H., Ahn, S., Lee, J., and Shin, J. Learning from failure: De-biasing classifier from biased classifier. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 20673–20684. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/eddc3427c5d77843c2253f1e799fe933-Paper.pdf.

Roudsari, A. A. https://github.com/aahmadnejad/biascxrbiascxr. https://github.com/aahmadnejad/BiasCXR, 2025. URL https://github.com/aahmadnejad/BiasCXR.

Schroter, S., Black, N., Evans, S., Carpenter, J., Godlee, F., and Smith, R. Effects of training on quality of peer review: randomised controlled trial. *BMJ*, 328(7441):673, March 2004.

Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I. Y., and Ghassemi, M. CheXclusion: Fairness gaps in deep chest x-ray classifiers. *Pac. Symp. Biocomput.*, 26:232–243, 2021a.

Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y., and Ghassemi, M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat. Med.*, 27(12):2176–2182, December 2021b.

Yang, Y., Zhang, H., Gichoya, J. W., Katabi, D., and Ghassemi, M. The limits of fair medical imaging AI in real-world generalization. *Nat. Med.*, 30(10):2838–2848, October 2024.

Zhang, H., Dullerud, N., Roth, K., Oakden-Rayner, L., Pfohl, S., and Ghassemi, M. Improving the fairness of chest x-ray classifiers. In Flores, G., Chen, G. H., Pollard, T., Ho, J. C., and Naumann, T. (eds.), *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 204–233. PMLR, 07–08 Apr 2022a. URL https://proceedings.mlr.press/v174/zhang22a.html.

Zhang, H., Dullerud, N., Roth, K., Oakden-Rayner, L., Pfohl, S. R., and Ghassemi, M. Improving the fairness of chest x-ray classifiers, 2022b. URL https://arxiv.org/abs/2203.12609.

Zhang, H., Hartvigsen, T., and Ghassemi, M. Algorithmic fairness in chest x-ray diagnosis: A case study. 2023.