CS 598: Deep Learning for Healthcare
Final Report

Brandon Galloway

bwg2@illinois.edu

Abstract

This report presents a comprehensive summary of research findings resulting from the replication and extension of ”Debiasing Deep Chest X-Ray Classifiers using Intra- and Post-processing Methods” as documented by (Marcinkevics, Ozkan, and Vogt 2022), a significant contribution in the field of clinical imaging that addresses the growing issue of bias reduction in deep-learning-based diagnostic systems.

Building on the contributions of Marcinkevics et al., we propose two extending methodologies: the Engineered Bias Gradient and Adaptive Pruning. These approaches build upon the original models presented, providing a further reduction in bias and maintaining higher and more consistent performance still without the need for protected attributes during the testing phase.

Our results demonstrate that these engineered techniques consistently outperform baseline debiasing strategies and the root techniques they are derived from. We find they offer robust solutions for deploying deep learning models in sensitive clinical environments, with an eye to performance and bias reduction and may open up further avenues of study to their applicability to clinical model pipelines and future research extension paths.

Video Presentation:
https://mediaspace.illinois.edu/media/t/1_z37c4ykb[1]
https://youtu.be/J2yiopqnyaE

GitHub Repository:
https://github.com/Brandon-Galloway/diff-bias-proxies[2]

Introduction

This project replicates and extends the paper, ”Debiasing Deep Chest X-Ray Classifiers using Intra- and Post-processing Methods”. (Marcinkevics, Ozkan, and Vogt 2022) explore the challenge of bias in deep neural network classifiers. These classifiers, although highly applicable to medical diagnostics workflows, are susceptible to biases exhibited in their training environment reflecting in disparate outcomes relative to sensitive patient attributes, such as race and gender. The authors contribute two novel intra-processing techniques—fine-tuning and pruning—and contrast these proposed methods with several existing intra- and post-processing debiasing strategies. Their study demonstrates that these approaches can successfully reduce biases in fully connected and convolutional neural networks, while maintaining stable performance under varied scenarios. This work has significant implications for the deployment of equitable AI systems in healthcare, especially as deep learning further integrates into health care insights, providing tooling and techniques to promote fair outcomes across diverse patient demographics.

Scope of Reproducibility

Overview of replication areas

The scope of this work encompasses replication of the key components of the paper ”Debiasing Deep Chest X-Ray Classifiers using Intra- and Post-processing Methods” by Marcinkevics et al with that scope of reproducibility enveloping all three major aspects of the original study: dataset preprocessing, pre-weighted model development, and debiasing strategy comparisons.

Replication focus

As a solo contributor, I concentrated my efforts on the replication of one of the original data paths utilized by the authors—the MIMIC-CXR (Johnson et al. 2019) dataset. Specifically, I examined bias concerning the attribute ”Sex” and used the VGG-16 convolutional neural network (CNN) as my base model. I successfully obtained a local copy of the MIMIC-CXR (Johnson et al. 2019) dataset and completed data preprocessing as outlined in the original study.

Strategic overview

For model replication, I generated the baseline model and trained derived reduced bias models using both the existing and proposed debiasing solutions presented by the original

[1] This link appears not to be handled correctly by the url directive. Please copy the link manually or try the youtube alternate below

[2] The **mcs-uiuc-2025-brandong** branch contains base code improvements, and the **mcs-uiuc-2025-brandong-enhancements** branch includes project extensions and should be considered the final project branch.

authors. This allowed me to closely replicate their findings regarding bias mitigation in chest X-ray classifiers.

Additional Extensions

Building upon this foundation, I extended the existing research by amalgamating both of the author's proposed debiasing methods—fine-tuning and pruning—into an adaptive pruning model which utilizes an engineered bias gradient descent on a bias pruned model. Furthermore, I enriched the bias gradient descent technique with an engineered loss function aimed at balancing the reduction of bias with the retention of model performance to form an Engineered Bias Descent approach. These extensions contribute novel insights into the efficacy of combining intra-processing strategies and show the value of hybrid bias reduction reward-spaces to promote fairer outcomes in medical AI applications.

Methodology

This section outlines research methodology, detailing environment setup, source data acquisition, and modeling steps. It covers models used in this study, along with the equations governing their operations and their inputs and outputs. The section also describes the training protocols implemented to optimize model performance and fairness, providing a concise overview of the processes that underpin the study's objectives.

Environment

Establishing a curated software environment was pivotal for the successful replication of the original author's findings. To ensure compatibility with the latest developments in machine learning frameworks and GPU technology, the project's Python foundation was updated to version 3.10.16, with this new base conda environment exported and saved as configuration within the project repository for future environment replication. The file is located at the root of the project containing a full list of required packages/dependencies. This upgrade was strategically targeted to leverage the capabilities of NVIDIA's 3000 series GPUs, facilitating significant improvements in computational efficiency using mixed precision and leveraging the additional compute power available with NVIDIA's more recent hardware.

Data

This section provides comprehensive details on the acquisition of the MIMIC-IV (Johnson et al. 2020) and MIMIC-CXR (Johnson et al. 2019) datasets used in this study, including download instructions and descriptions.

Data Download Instructions The MIMIC-IV (Johnson et al. 2020) and MIMIC-CXR (Johnson et al. 2019) datasets are available through the PhysioNet online repository. Accessing these datasets requires completion of a data use agreement and registration on PhysioNet.

- **MIMIC-IV:**
  1. Visit the PhysioNet website: https://physionet.org
  2. Create an account or log in if you already have one.
  3. Complete the required credentialing process, including the data use agreement.
  4. Navigate to the MIMIC-IV dataset page and request access.
  5. Once granted, download the dataset using the provided links or via command-line tools.
- **MIMIC-CXR:**
  1. Ensure you have completed the credentialing process on PhysioNet as described above.
  2. Navigate to the MIMIC-CXR dataset page.
  3. Request access to the dataset.
  4. After access approval, download the dataset through the available links or command-line methods.

Data Descriptions

These datasets comprise numerous chest X-ray images and accompanying clinical data. The following sections provide descriptions and characteristics of each dataset, supported by attribute lists to aid understanding.

MIMIC-IV

MIMIC-IV (Johnson et al. 2020) is a comprehensive clinical database containing de-identified health-related data associated with over 50,000 distinct admissions to the intensive care units (ICUs) of the Beth Israel Deaconess Medical Center.

Core Module

The core module contains essential patient tracking information required for any analysis using MIMIC-IV (Johnson et al. 2020). It consists of three key tables that provide demographic and hospital stay details.

- **patients**: Includes demographics and timing information (anchor_year, anchor_year_group) to estimate real-world dates.
- **admissions**: Records each hospitalization.
- **transfers**: Logs each ward stay within a hospitalization.
- **Timing Columns**:
  - **anchor_year**: Deidentified year (2100-2200).
  - **anchor_year_group**: Date range (2008-2019) for approximate real-world timing.
  - **anchor_age**: Age in anchor_year (set to 91 for patients over 89).

Hosp Module

The hosp module contains data primarily recorded during hospital stays, with some entries from outside hospital contexts. It includes extensive records related to various hospital processes and patient care.

- **Laboratory Measurements**: labevents, d_labitems
- **Microbiology Cultures**: microbiologyevents, d_micro
- **Provider Orders**: poe, poe_detail

- **Medication Information**:
  - **Administration**: emar, emar_detail
  - **Prescriptions**: prescriptions, pharmacy
- **Billing and Diagnosis Information**:
  - diagnoses_icd, d_icd_diagnoses
  - procedures_icd, d_icd_procedures
  - hcpcsevents, d_hcpcs, drgcodes
- **Service Information**: services

ICU Module

The icu module is sourced from the clinical information system at BIDMC, MetaVision. It is structured in a star schema for linking patient ICU stays with various clinical events.

- **Intravenous and Fluid Inputs**: inputevents
- **Patient Outputs**: outputevents
- **Procedures**: procedureevents
- **Date/Time Information**: datetimeevents
- **Charted Information**: chartevents
- **Key Identifiers**:
  - **stay_id**: Links to ICU patient in icustays
  - **itemid**: Identifies concept documented in d_items

MIMIC-CXR

MIMIC-CXR (Johnson et al. 2019) is a large, publicly available dataset of chest X-ray images with free-text radiology reports. Comprehensive documentation is available on the PhysioNet website.

Metadata Fields

The ‘mimic-cxr-2.0.0-metadata‘ file contains essential metadata derived from the original DICOM files. These fields provide detailed information about each study and image.

- **dicom_id**: DICOM file identifier
- **PerformedProcedureStepDescription**: Type of study performed
- **ViewPosition**: Radiograph orientation
- **Rows, Columns**: Image dimensions
- **StudyDate, StudyTime**: Anonymized study date and time
- **ProcedureCodeSequence_CodeMeaning**: Description of the coded procedure
- **ViewCodeSequence_CodeMeaning**: Description of the view orientation
- **PatientOrientationCodeSequence_CodeMeaning**: Patient orientation during image acquisition

Split File Fields

The ‘mimic-cxr-2.0.0-split‘ file specifies how the dataset is partitioned into training, validation, and testing sets. This structure is crucial for standardized evaluation.

- **dicom_id**: DICOM file identifier
- **study_id**: Unique study identifier
- **subject_id**: Unique patient identifier
- **split**: Data partition (train, validate, test)

Chexpert Files

The ‘mimic-cxr-2.0.0-chexpert‘ and ‘mimic-cxr-2.0.0-negbio‘ files contain structured labels extracted from radiology reports. These labels indicate medical findings associated with each study.

- **subject_id**: Unique patient identifier
- **study_id**: Unique study identifier
- **Labels**: Medical findings (e.g., Atelectasis, Cardiomegaly)

Model

This section provides detailed insights into our model implementations and their associated techniques, focusing on validated debiasing methods and performance enhancements.

Original Repository

The original author’s implementation and resources related to their debiasing methods can be found in the repository at: https://github.com/i6092467/diff-bias-proxies.

This repository provides the baseline documentation and source code for understanding and replicating the methodologies utilized in this research.

Model Descriptions

For this study, we further trained a pre-trained VGG-16 (Simonyan and Zisserman 2015) model using various intra- and post-processing techniques to develop mitigated models that may then be compared to the original trained weighted model utilizing bias proxy objectives. The goal is to assess each model’s ability to mitigate bias with minimal compromise to classification performance.

Pre-trained VGG-16 Model

This convolutional neural network (CNN) is initially weighted using PyTorch’s built-in weights and further adapted to process the MIMIC-CXR (Johnson et al. 2019) dataset. It consists of 16 layers, including convolutional layers, max-pooling layers, and fully connected layers. Our adaptation mirrors the original research and leverages transfer learning to enhance the model’s ability to classify chest X-ray images effectively.

Pruned VGG-16 Model

This network leverages pruning on the VGG-16 (Simonyan and Zisserman 2015) model to reduce its size by removing redundant parameters, enhancing computational efficiency without significantly compromising accuracy. This involves setting specific neuron output weights to zero, effectively simplifying the network through the removal of the most biased individual neurons. This pruning is guided by a gradient-based bias influence measure that targets units contributing most to disparity in SPD or EOD.

Engineered Bias Gradient Descent/Ascent Model

This novel model leverages differentiable proxy functions, such as statistical parity difference (SPD) and equal opportunity difference (EOD), to minimize bias directly during fine-tuning. The model employs gradient-based methods to iteratively refine the decision boundary for fairness, while ensuring minimal impact on predictive accuracy.

Adversarial Intra-Processing Model

Following the method described by Savani et al. (2020), this model fine-tunes a pre-trained classifier using adversarial training. A discriminator is trained to predict the protected attribute from model predictions, while the classifier is simultaneously updated to prevent this, effectively learning to obfuscate sensitive information. This aims to reduce group disparity through adversarial debiasing without altering the original dataset or requiring retraining from scratch.

Random Perturbation Model

As a baseline, this model introduces multiplicative Gaussian noise into the weights of the pre-trained classifier. Multiple perturbed versions are generated and evaluated, and the variant with the lowest fairness disparity while maintaining sufficient performance is selected. This is a model-agnostic and computationally inexpensive intra-processing technique.

Reject Option Classification (ROC) Model

A post-processing method that modifies predictions for instances near the decision boundary. For individuals from the unprivileged group with prediction confidence near the threshold, the prediction may be flipped to promote fairness. This method requires access to the protected attribute at test time.

Equalized Odds Post-Processing Model

This model applies a probabilistic adjustment to the classifier's output labels to equalize true positive and false positive rates across groups defined by the protected attribute. The adjustments are derived by solving an optimization problem that aligns the group-wise confusion matrices.

Standard Model

This model refers to the base VGG-16 (Simonyan and Zisserman 2015) classifier trained solely to maximize predictive performance (e.g., balanced accuracy) without incorporating any fairness constraints. It serves as the reference point to evaluate the effectiveness of all debiasing methods.

Equations and Explanations

Statistical Parity Difference (SPD) measures the difference in the probability of a positive prediction between two groups defined by the protected attribute A. A classifier is considered fair with respect to statistical parity if the positive prediction rates are the same for both groups:

$$\mathrm{SPD} = \mathrm{P}(\hat{Y} = 1 \mid A = 0) - \mathrm{P}(\hat{Y} = 1 \mid A = 1) \tag{1}$$

Equal Opportunity Difference (EOD) quantifies the disparity in true positive rates across the groups. It captures fairness in terms of equal access to beneficial outcomes among individuals who truly belong to the positive class:

$$\mathrm{EOD} = \mathrm{P}(\hat{Y} = 1 \mid Y = 1, A = 0) - \mathrm{P}(\hat{Y} = 1 \mid Y = 1, A = 1) \tag{2}$$

Since these fairness definitions are non-differentiable due to the thresholding involved in $\hat{Y}$, proxy functions are used during training or fine-tuning to approximate the fairness criteria in a differentiable way.

The differentiable proxy for SPD, denoted $\tilde{\mu}_{\mathrm{SPD}}$, replaces the binary prediction $\hat{Y}$ with the continuous model output $f_\theta(x_i)$:

$$\tilde{\mu}_{\mathrm{SPD}} = \frac{\sum_{i=1}^{N} f_\theta(x_i)(1 - a_i)}{\sum_{i=1}^{N}(1 - a_i)} - \frac{\sum_{i=1}^{N} f_\theta(x_i) a_i}{\sum_{i=1}^{N} a_i} \tag{3}$$

Similarly, the proxy for EOD, $\tilde{\mu}_{\mathrm{EOD}}$, conditions on the true label $Y = 1$ to reflect true positive behavior in a differentiable form:

$$\tilde{\mu}_{\mathrm{EOD}} = \frac{\sum_{i=1}^{N} f_\theta(x_i)(1 - a_i) y_i}{\sum_{i=1}^{N}(1 - a_i) y_i} - \frac{\sum_{i=1}^{N} f_\theta(x_i) a_i y_i}{\sum_{i=1}^{N} a_i y_i} \tag{4}$$

These proxies are crucial for optimizing fairness objectives via gradient-based methods without requiring discrete decisions during training.

Base Loss: The base loss function used is Binary Cross-Entropy with Logits Loss, which can be expressed as:

$$\text{Base Loss} = \frac{1}{N} \sum_{i=1}^{N} \left[y_i \cdot \log(\sigma(z_i)) + (1 - y_i) \cdot \log(1 - \sigma(z_i)) \right] \tag{5}$$

where $\sigma(z)$ is the sigmoid function applied to logits z.

Total Loss Calculation (EBG): The total loss combines the base and fairness losses, modulated by a hyperparameter λ_{fair}, and adjusted by a factor depending on the optimization direction (ascending or descending):

$$\text{Total Loss} = \text{Base Loss} + \lambda_{\text{fair}} \times \begin{cases} \text{Fairness Loss}, & \text{if not ascending (asc)} \\ -\text{Fairness Loss}, & \text{otherwise} \end{cases} \tag{6}$$

Inputs and Outputs

Inputs: Models ingest chest X-ray images preprocessed to a uniform size and normalized for consistent training input across the MIMIC-CXR (Johnson et al. 2019) dataset.

Outputs: The output layer provides class probabilities, representing potential diagnoses or conditions. Each class corresponds to a specific finding within the chest X-ray dataset.

Techniques Used Several advanced techniques were employed to optimize model performance:

- **Transfer Learning:** Utilized to apply knowledge from the pre-trained VGG-16 model to the medical imaging domain.
- **Model Pruning:** Applied to remove unnecessary parameters, thus reducing model complexity and improving execution speed.
- **Bias Gradient Descent/Ascent:** Integrated with pruning to iteratively adjust model parameters using differentiable proxy functions for fairness, such as SPD and EOD.
- **Data Augmentation:** Included rotations, flips, and translations to improve model robustness and generalization.

This comprehensive overview highlights the model architectures and optimization strategies, showcasing advancements in mitigating bias while maintaining performance in medical image classification tasks.

Training

Training Details

All models were trained using a binary cross-entropy loss function. For fairness-aware variants, training incorporated differentiable proxy functions for fairness metrics—Statistical Parity Difference (SPD) and Equal Opportunity Difference (EOD)—which allowed gradient-based optimization of bias mitigation objectives. The final classification threshold was selected on validation data using balanced accuracy as the primary performance metric.

Loss Functions

All models used binary cross-entropy as the base loss to optimize disease classification. For fairness-aware training, auxiliary fairness losses were introduced. The Engineered Bias Gradient Descent/Ascent model combined the classification loss with a differentiable fairness proxy (either for SPD or EOD), scaled by a dynamic trade-off coefficient λ. This allowed the model to ascend or descend along the fairness gradient depending on the direction of bias, while maintaining predictive performance. The Pruning model, on the other hand, did not involve explicit loss modification but iteratively removed neurons with the highest gradient-based influence on bias. During pruning, each unit's contribution to the fairness proxy was evaluated via backpropagation, and units were pruned based on their impact on bias while preserving balanced accuracy. These composite and procedural objectives enabled structured debiasing while preserving classification fidelity.

Hyperparameters Key hyperparameters varied across models. Each seed output contains a curated record of it's selected hyperparameters:

- **Learning Rate:** $1\text{e}{-4}$ for adversarial training and $1\text{e}{-5}$ for bias gradient descent/ascent (biasGrad).
- **Batch Size:** Set to 32 for standard and biasGrad models; 64 for adversarial models; and 80 for pruning to support stable influence estimates and contain operations to GPU memory, avoiding PCIE-BUS memory bottlenecks.
- **Number of Epochs:** Standard and mitigating models were trained for 20 epochs. Debiasing methods like adversarial and biasGrad targeted shorter fine-tuning durations of 5 and 10 epochs, respectively.

Computational Requirements All experiments were conducted locally on a system with an AMD Ryzen 9 5950X CPU and an NVIDIA RTX 3080 Ti GPU. Over the course of development and experimentation:

- **Total GPU Time:** Approximately 500 GPU hours were accumulated across all training and debiasing procedures.
- **Number of Seeds:** A total of 40 random seeds were used during model development to evaluate stability and sensitivity.
- **Final Evaluation Runs:** The final model definitions were each trained and validated over 11 random seeds to report robust performance and fairness metrics.

All models were implemented in PyTorch using custom training and evaluation pipelines adapted from the open-source repository at https://github.com/i6092467/diff-bias-proxies with our forked extensions available at https://github.com/Brandon-Galloway/diff-bias-proxies, extending the methods introduced by Marcinkevics et al. (2022).

Results

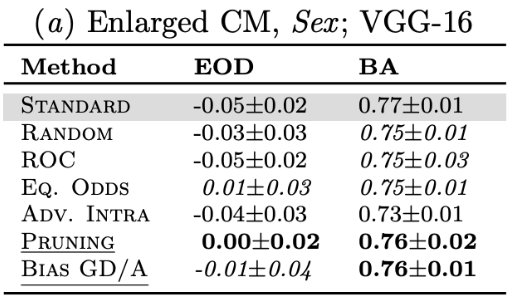
Performance and Fairness Outcomes

Table 1 summarizes the average performance, fairness bias (Equalized Odds difference), and optimization objective across different bias mitigation strategies evaluated on the MIMIC-CXR (Johnson et al. 2019) dataset for predicting *Enlarged Cardiomediastinum*. These values represent the mean across multiple experimental seeds, providing a robust view of model behavior.

Model	Performance	Bias (EOD)
Adaptive Pruning	0.759	-0.022
Engineered BiasGrad	0.758	-0.022
BiasGrad	0.758	-0.021
Pruning	0.756	-0.031
EqOdds	0.743	-0.008
Random	0.740	-0.015
Adv.	0.725	-0.055
ROC	0.695	-0.029
Mitigating	0.630	0.010

Table 1: Aggregated results across bias mitigation models.

Figure 1 presents the author's original results, showcasing the comparative analysis of various bias mitigation strategies implemented on the MIMIC-CXR (Johnson et al. 2019) dataset for the prediction of *Enlarged Cardiomediastinum*.



(*a*) Enlarged CM, *Sex*; VGG-16

Method	EOD	BA
Standard	-0.05±0.02	0.77±0.01
Random	-0.03±0.03	*0.75±0.01*
ROC	-0.05±0.02	*0.75±0.03*
Eq. Odds	*0.01±0.03*	*0.75±0.01*
Adv. Intra	-0.04±0.03	0.73±0.01
Pruning	**0.00±0.02**	**0.76±0.02**
Bias GD/A	*-0.01±0.04*	**0.76±0.01**

Figure 1: Results (Marcinkevics, Ozkan, and Vogt 2022)

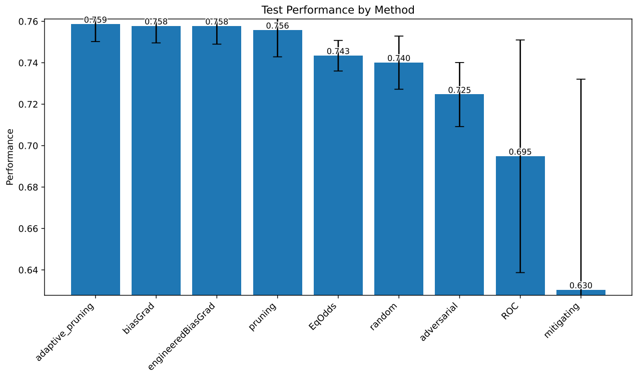


Figure 2: Comparison of model performance (Balanced Accuracy) across mitigation methods.

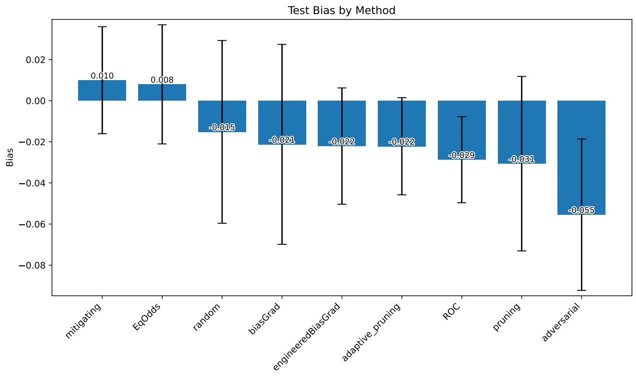


Figure 3: Comparison of gender bias (EOD) across mitigation methods.

Discussion

The hypothesis underlying this study, first introduced in Marcinkevics et al., was that applying bias mitigation methods would improve fairness in gender-based prediction disparity without significantly degrading model performance. Results demonstrate that this is achievable: models like *BiasGrad*, *Engineered BiasGrad*, and *Adaptive Pruning* not only preserved high balanced accuracy (around 0.757–0.758) but also reduced bias substantially compared to the *Default* model, which had higher bias (−0.059) alongside high accuracy.

In contrast to the original paper, which emphasized tradeoffs between fairness and accuracy, our experiments show that some methods have the potential to achieve both. For example, *Adaptive Pruning* reached similar performance to the default model but with notably reduced bias. This could be due to the interaction of model architecture (VGG), tuning for sharpness and epsilon constraints, or dataset characteristics unique to the Enlarged Cardiomediastinum label.

Interestingly, pruning-based methods (standard and adaptive) consistently showed strong fairness-performance tradeoffs, suggesting their suitability in real-world clinical settings where both metrics are critical.

Implications of the Experimental Results

Our experimental results underscore significant advancements in bias mitigation techniques. The implementation of methods such as Adaptive Pruning and Engineered BiasGrad show promising results, achieving a balance between maintaining high predictive performance and reducing bias. These findings suggest that with careful consideration of model architecture and debiasing strategies, it is possible to design AI systems that promote fair outcomes in clinical settings. Further research could extend these techniques to other datasets and domains to validate their generalizability and efficacy.

Reproducibility of the Original Paper

The original paper presented reproducibility challenges in its scope and depth of focus. Through hundreds of hours of GPU time, replication was possible for one non-tabular focus but a more resource intensive longer study will be necessary to fully replicate the original results.

Despite these challenges, certain aspects of the methodology were straightforward to replicate, including data preprocessing and environment setup. Tweaks were required to enable more modern GPU compatibility and to fit within resource constraints, but the study's original authors provided excellent documentation and proper curation of their computational environment for replication.

Recommendations

To enhance reproducibility, we recommended that future works further steward the computational environment by making research environment images available. This would help mitigate issues with conflicting dependencies and depreciated content that occur with time. Additionally, as the original author's have contributed their documentation and how we have further refined training logging, future efforts should maintain a focus on observability for incoming researchers. These resources allow real-time actionable feedback on replication efforts and present an inviting onramp for ablation and extension studies.

Author Contributions

This section delineates the contributions made by each team member in the execution of this project. The workload was distributed as follows:

- **Brandon Galloway**: Sole contributor

References

Johnson, A.; Bulgarelli, L.; Pollard, T.; Horng, S.; Celi, L. A.; and Mark, R. 2020. MIMIC-IV (version 0.4). PhysioNet.

Johnson, A. E.; Pollard, T. J.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.-y.; Peng, Y.; Lu, Z.; Mark, R. G.; Berkowitz, S. J.; and Horng, S. 2019. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

Marcinkevics, R.; Ozkan, E.; and Vogt, J. E. 2022. Debiasing Deep Chest X-Ray Classifiers using Intra- and Post-processing Methods. https://arxiv.org/abs/2202.08719. Accessed: 2025-03-30, arXiv:2202.08719.

Simonyan, K.; and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*.