

Eliminating race-related shortcuts in deep neural networks for chest X-ray analysis

Ryan Wang^{1*}, Li-Ching Chen^{1*}, Pei-Chuan Lin¹, Judy Wawira Gichoya², Leo Anthony Celi^{3,4}, Po-Chih Kuo¹

¹Department of Computer Science, National Tsing Hua University

²Department of Radiology, Emory University ³Institute for Medical Engineering and Science, Massachusetts Institute of Technology ⁴Department of Medicine, Beth Israel Deaconess Medical Center

Abstract

Current deep learning (DL) models are capable of exceptional performance in the detection of anomalies in chest X-ray (CXR) images. However, some diseases are more prevalent within specific racial groups, such that an imbalance in the number of patients of a specific race can bias the training process, with corresponding disparities in prediction accuracy among racial groups. DL models have proven highly effective in classifying races based on CXR images; however, the ease with which machine learning models identify racial features raises the possibility that such features are used as shortcuts in the identification of pathological features; i.e., introducing bias. In this study, we investigated the underlying methods by which DL models identify race in order to prevent the use of shortcuts based on irrelevant racial features, which could potentially bias detection results. Our use of image processing to emphasize or enhance specific image characteristics revealed that the outline of the lung contributed significantly to race identification. We accordingly modified the CXR images with the aim of reducing the emphasis on these racial characteristics while maintaining the ability to detect important radiological anomalies. The performance of anomaly detection of the DL model trained using processed CXR images was less biased than that of the model trained using the original CXR images. Our results demonstrate the efficacy of the proposed scheme in identifying disease-related features without shortcut solutions resulting from race-related information hidden within CXR images.

Introduction

Computer-aided diagnosis (CAD) and deep learning (DL) have proven to be highly effective in disease detection (Rajpurkar et al. 2017; Baltruschat et al. 2019) and anatomical segmentation (Minaee et al. 2021; Ronneberger, Fischer, and Brox 2015); however, DL models are prone to taking shortcuts based on oversimplified characteristics in the data. For example, a machine learning model may detect a cow in an image based not on the presence of a cow but rather on the grassland in the background (Geirhos et al. 2020). In such situations, an extraordinary setting (e.g., a cow standing on a beach) would result in identification failure in the absence of the shortcut (grassland). Machine learning systems

are prone to taking the same kinds of shortcuts in the analysis of medical images (Zech et al. 2018). For example, if training were performed using the image dataset from a hospital with a high prevalence of pneumonia, then the machine might prioritize the metal token placed by the technologists in the corner of the CXR image over the more complex shapes and patterns indicative of pneumonia. Similar situations might arise when a machine learning model focuses on features that are typical of race rather than pathologic phenomena. These “shortcuts” enable the machine learning system to obtain impressive performance, not in the actual detection or identification of disease. The shortcut problem can also seriously compromise prediction accuracy in real-world settings.

One recent study (Banerjee et al. 2021) reported that machine learning models are highly effective in differentiating among individuals of different races, based on CXR images, cervical spine X-rays, and computed tomography (CT) scans of the chest. In that study, machine learning models achieved high AUC scores (0.8-0.99) even when trained using images of low quality, segmented regions, or other perturbations. The ease with which machine learning models identify race from CXR images raises the possibility that such features are used as shortcuts in the identification of pathological features with potential to introduce bias. In a previous study (Seyyed-Kalantari et al. 2020), a model generated disparate results as a function of race, such that some racial subgroups faced an elevated risk of false predictions. Obviously, the use of metal tokens as shortcuts is easily recognized and rectified by clinicians; however, the racial attributes in CXR images can be very subtle. Researchers are largely at a loss when it comes to understanding the means by which machine learning models identify race from CXR images, thereby making it nearly impossible to eliminate race-related bias from diagnostic results (Banerjee et al. 2021). In the current study, we first apply a variety of image processing methods to CXR images and assess the model’s performance of race identification with the aim of identifying racial features that could potentially bias the results of machine learning. We then propose augmentation methods to modify CXR images for eliminating race-related shortcuts. The performance of anomaly detection of the model trained using the modified CXR images was less biased than that of the model trained using the original images.

*These authors contributed equally.

Data

This study was based on de-identified CXR images and clinical data in the MIMIC-CXR (Johnson et al. 2019a,b) and MIMIC-IV (Goldberger et al. 2000) databases. The institutional review boards of Massachusetts Institute of Technology (No. 0403000206) and Beth Israel Deaconess Medical Center (2001-P-001699/14) both approved the creation of the database for research. We used only frontal CXR images in MIMIC-CXR, each of which had a corresponding subject_id, study_id, and 14 disease classes, which were derived from free-text radiological reports. MIMIC-IV lists the demographic information of intensive care patients with a unique subject_id. We extracted White, Black, and Asian patients as our patient cohort, in accordance with the methods described in our recent study (Banerjee et al. 2021). We matched the subject_id in MIMIC-CXR and MIMIC-IV in order to link the demographic labels to the CXR images. We obtained a total of 194,359 frontal images from patients of three major races: White (150,663), Black (36,282), and Asian (7,414). The dataset was randomly split into training (60%), validation (10%), and testing (30%) datasets with no patients shared across the splits. Histogram equalization was applied to all images in MIMIC-CXR, and all images were resized to (224, 224). CXR images from Emory University were used as an additional cohort for external validation: White (5,281), Black (6,067), Asian (484).

Method

All DL models were implemented on the Densenet121 architecture equipped with ImageNet pretrained weights obtained from the TensorFlow Keras library. For race classification, we added a Softmax classification layer with 3 nodes. The race labels were mutually exclusive; therefore, we used the Adam optimizer to optimize categorical cross-entropy loss. For the detection of radiological anomalies, we added a sigmoid classification layer with 14 nodes. The 14 radiological features were independent; therefore, we used the Adam optimizer to optimize binary cross-entropy loss. For both tasks, the number of epochs depended on the specifics of the training process. Training was stopped when validation loss did not improve through 4 consecutive epochs. The batch size was set to 128. The learning rate decayed by 5% per 2 epochs with an initial learning rate of 1e-3. The area under the curve (AUC) was used as the primary metric, and the 95% confidence interval (CI) was calculated over 5 runs of model training using identical hyperparameters. All images were written into TFrecords to ensure data consistency for every run of the training process. The global random seed was set to 2021 to ensure reproducibility. For model explanation, we compared saliency maps of various DL models to determine whether a given DL model focused on true disease-related characteristics or a shortcut (Selvaraju et al. 2017). The study was implemented in four consecutive phases. Phase 1 involved the assessment of race disparity in the dataset. The chi-square test and permutation test were used to identify race-induced deviations in radiological findings. Phase 2 involved image processing aimed at elucidating the means by which the DL models recog-







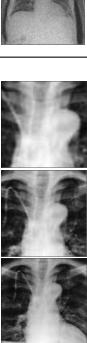
Methods	Procedures	Examples
Original	N/A	
Sobel filter	Extract edges with kernel size of 3 and scale factor of 4.	
Binary Otsu's thresholding	Extract the outline of lungs and trunk using binarizing pixel values.	
Binary Otsu's thresholding with erosion and dilation	Extract the outline and then smooth the edges by adding erosion and dilation (filter size = 3x3). We used 1 iteration for erosion and 3 iterations for dilation.	
Gaussian blurring	Image blurring via Gaussian filter with standard deviation of the kernel set to 10.	
Gaussian noise addition	Add Gaussian noise with the mean of the normal distribution set to 0 and the standard deviation set to 0.1.	
Central cropping (large/medium/small)	Retain central area of specific size (128, 128)/ (96, 96)/ (64, 64) and resize the image to the original size.	

Table 1: Proposed methods for identification of race-related characteristics in Phase 2.

nized the race of patients via CXR image analysis. Phase 3 involved image transformations aimed at reducing the effect of race classification while maintaining accuracy in the detection of radiological anomalies. Phase 4 involved the use of true positive rate (TPR) disparity metrics to evaluate the accuracy of the proposed models in the detection of 14 disease classes. We also compared the saliency map derived from each model to determine whether it focused more on the 14 disease classes.

Phase 1: Detection of shortcut learning in CXR

Potential shortcuts in the future model training were detected using the chi-square test to evaluate the independence of race (White vs. Black) and the 14 radiological features. Logistic regression (LR) was used to predict 14 radiological features based on race and evaluate the significance using a permutation test under a significance level of 0.001. The

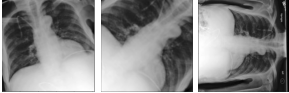
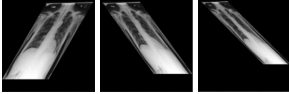
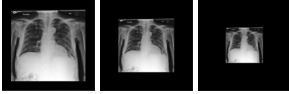
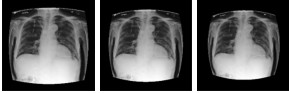
Methods	Procedures	Examples
Rotating transformation (light/medium/heavy)	Rotate images by a random amount ranging from -20° to 20° / -45° to 45° or -90° to 90°.	
Shear transformation (light/medium/heavy)	Apply shear transformation to images with a random radian ranging from $-\frac{\pi}{6}$ to $\frac{\pi}{6}$ / $-\frac{\pi}{5}$ to $\frac{\pi}{5}$ / $-\frac{\pi}{4}$ to $\frac{\pi}{4}$.	
Scaling transformation (light/medium/heavy)	Scale the side length of images to a randomly selected size ranging from 0.8 to 1 / 0.6 to 1 / 0.4 to 1.	
Fisheye distortion (light/medium/heavy)	Apply fisheye effect with distortion coefficient set to 0.2/0.3/0.4.	

Table 2: Proposed methods for mitigating effects of race-related characterization in Phase 3.

weighted average f1-scores over 100,000 times were computed by shuffling the image labels.

Phase 2: Identification of race-related characteristics

Our primary objective here was to determine the means by which the DL models recognized race in CXR images through the use of images processed using the methods listed in Table 1. Based on the performance of models in terms of race classification, we then speculated as to the distinctive racial characteristics recognized by the DL models.

Phase 3: Mitigating effects of race-related characterization

Table 2 lists the methods used to mitigate the effects of race-related characterization and presents example images.

Phase 4: Evaluation of the proposed method

The efficacy of the proposed method was evaluated using TPR disparity as an evaluation metric (Seyyed-Kalantari et al. 2020). We first calculated the TPR disparity between White and Black patients in order to measure bias in the detection of the 14 radiological features. We selected binary classification thresholds for each class to maximize the weighted f1 score for the validation set. We then compared saliency maps generated using gradient weighted class activation mapping (GradCAM) (Selvaraju et al. 2017) of various models in order to visualize the accuracy of predictions for specific diseases. The heavy rotation transformation was selected as the final method, for two reasons. First, the implementation of random rotation is simple and efficient. Second, this approach made it possible to fill the frame without losing as much information as would have been the case using other methods. Note that we were concerned that black areas in the frame could leave clues, which would allow the

DL models to reconstruct the original images, thereby rendering the proposed modifications useless. Subsequent external validation was performed using the Emory dataset.

Result

Phase 1: Detection of shortcut learning

The results of the chi-square test revealed a dependency between detection findings and race ($p < 0.01$). LR results revealed significantly high f1 scores for all but 1 (“No Finding”) of the 14 radiological features ($p < 0.01$ in the permutation test).

Phase 2: Identification of race-related characteristics

Table 3 lists the AUCs for race identification. The good performance obtained using a Sobel filter revealed that the DL models were able to identify the race of the patient based on the outlines of CXR images. Binary Otsu’s thresholding reduced the performance of the model to below that of the original due to the inclusion of binary pixel values. Nonetheless, an AUC score exceeding 0.8 indicates that the given DL model was still able to recognize the race of a patient based on the strengthened outlines. Erosion and dilation were shown to reduce race identification performance; i.e., a smooth outline undermined race-related characteristics. Adding Gaussian blurring reduced race identification performance whereas adding Gaussian noise had no influence, thereby demonstrating the importance of lung outlines in race detection. In central cropping experiments, model performance was inversely correlated with the size of the lung region that was examined. Note that smaller regions tended to exclude the outer outline of the lungs. Taken together, these results indicate that the outer outline of the lung is crucial to race identification.

Image Processing Methods	AUC-White	AUC-Black	AUC-Asian	Findings
Original	0.946 ± 0.010	0.956 ± 0.008	0.943 ± 0.013	N/A
Sobel filter	0.929 ± 0.012	0.940 ± 0.009	0.921 ± 0.025	Outlines matter.
Binary Otsu's thresholding	0.813 ± 0.019	0.829 ± 0.012	0.801 ± 0.013	Pixel intensity does not matter.
Binary Otsu's thresholding with erosion and dilation	0.779 ± 0.013	0.790 ± 0.011	0.772 ± 0.013	Details of outlines are important.
Gaussian blurring	0.824 ± 0.107	0.831 ± 0.110	0.818 ± 0.077	Outlines are important.
Gaussian noise addition	0.894 ± 0.034	0.907 ± 0.030	0.891 ± 0.031	Pixel values do not matter.
Central cropping (light/medium/heavy)	0.929 ± 0.003 0.859 ± 0.131 0.757 ± 0.072	0.929 ± 0.004 0.858 ± 0.140 0.764 ± 0.076	0.913 ± 0.018 0.854 ± 0.089 0.727 ± 0.048	Less information equates to lower performance. Excluding the outer outline of the lungs decreases performance.

Table 3: Results of image manipulation methods in race identification.

Phase 3: Mitigating the effects of race-related features

Table 4 lists the results for race identification and radiological feature detection. Note that the degree of influence on the two tasks is summarized in the last two columns (L: Low; M: Medium; H: High). All methods performed well in the detection of radiological features. The use of heavily rotated images for training had the most pronounced effect on mitigating the effects of race-related features, and the low CI for the detection of radiological features demonstrated that the training process was stable.

Phase 4: Evaluation of proposed image manipulation scheme

The results of TRP disparity analysis are presented in Figure 1, where bar length is proportional to the degree of disparity. In MIMIC-CXR and Emory dataset, TPR disparity was lower for the model trained using rotated images (8 out of 14 classes and 9 out of 14 classes, respectively) than for the model trained using the original images.

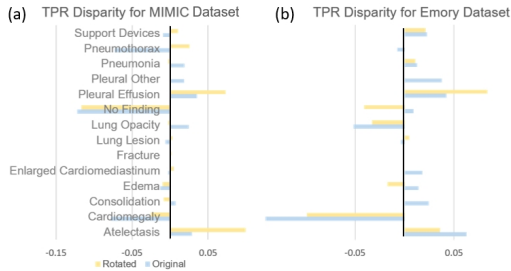


Figure 1: Disparities in the accuracy in the identification of radiological features between the original model and the proposed model: (a) Results obtained using MIMIC-CXR; (b) External validation results obtained using the Emory dataset.

In the comparison of saliency maps, consolidation was selected as the target disease. The model trained using rotated CXR images focused more on the disease characteristics than did the model trained using the original CXR images. The results in Figure 2 illustrate how the removal of race-related characteristics indeed facilitated the learning of

true radiological features instead of taking shortcuts. Despite the fact that this did not enhance performance in the detection of radiological features, it greatly reduced the influence of shortcuts.

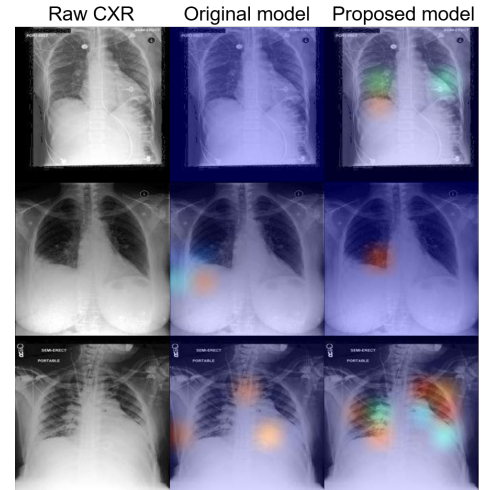


Figure 2: GradCAM of the various models, in which the original includes cues outside the lungs and the proposed models shift the focus to cues within the lungs where the findings are supposed to be.

Discussion

Our results revealed that the shape of lungs is an important factor in the identification of race and is therefore a likely shortcut for the detection of radiological features. We also determined that image rotation can be used to decrease the weight assigned to race-related features without compromising the performance of anomaly detection. Unlike shear, scaling, and fisheye transformations, rotation preserved most of the important features in the CXR images. The proposed training scheme was also shown to mediate the disparities in detection performance among races. Furthermore, the efficacy of the rotation transformation proved effective across multiple architectures and datasets. To the best of our knowledge, this is the first study to introduce a

Methods	AUC-White	AUC-Black	AUC-Asian	Averaged AUC-radiological features	Influence on race identification	Influence on detection of radiological features
Original	0.946 \pm 0.010	0.954 \pm 0.008	0.943 \pm 0.013	0.769 \pm 0.023	N/A	N/A
Rotating transformation (light/medium/heavy)	0.932 \pm 0.008 0.844 \pm 0.130 0.816 \pm 0.108	0.940 \pm 0.006 0.856 \pm 0.129 0.825 \pm 0.111	0.915 \pm 0.023 0.823 \pm 0.123 0.781 \pm 0.010	0.737 \pm 0.077 0.751 \pm 0.024 0.752 \pm 0.005	L/ M/ H	L/ L/ L
Shear transformation (light/medium/heavy)	0.871 \pm 0.132 0.862 \pm 0.154 0.771 \pm 0.179	0.881 \pm 0.137 0.869 \pm 0.154 0.783 \pm 0.183	0.860 \pm 0.123 0.841 \pm 0.137 0.730 \pm 0.174	0.749 \pm 0.024 0.702 \pm 0.065 0.743 \pm 0.027	M/ M/ H	L/ M/ L
Scaling transformation (light/medium/heavy)	0.938 \pm 0.015 0.879 \pm 0.127 0.806 \pm 0.210	0.948 \pm 0.013 0.915 \pm 0.061 0.817 \pm 0.212	0.934 \pm 0.023 0.865 \pm 0.148 0.770 \pm 0.214	0.760 \pm 0.037 0.751 \pm 0.036 0.744 \pm 0.045	L/ M/ H	L/ L/ L
Fisheye distortion (light/medium/heavy)	0.926 \pm 0.008 0.916 \pm 0.010 0.907 \pm 0.012	0.939 \pm 0.005 0.926 \pm 0.010 0.919 \pm 0.009	0.919 \pm 0.015 0.909 \pm 0.007 0.890 \pm 0.031	0.767 \pm 0.015 0.757 \pm 0.030 0.747 \pm 0.034	L/ L/ L	L/ L/ L

Table 4: Results of image manipulation methods. L:Low; M: Medium; H: High.

method by which to mediate the influence of race-related physiological features in the interpretation of CXR images.

Building a fair DL model for diagnosis and prognosis requires that practitioners understand the means by which DL models formulate their decisions (Singh, Sengupta, and Lakshminarayanan 2020). Nonetheless, the actual operation of a DL model remains largely shrouded, despite knowledge of the underlying algorithm. Some DL applications (e.g., classifying handwritten digits) can be elucidated from a purely visual perspective (Selvaraju et al. 2017); however, saliency maps are notoriously unreliable (Adebayo et al. 2018; ?; Kindermans et al. 2019). GradCAM works well on object localization and mapping; however, it often overlooks fine-grained features, such as uneven boundaries and subtle anatomical anomalies (Lei et al. 2020; Jung and Oh 2021). The extreme complexity of radiographic medical images renders many of the explanations provided by machines opaque to human comprehension. We also found that GradCAM produced inconsistent mappings when applied to race-classification tasks

One previous study reported that DL models are able to detect subtle patterns (undetectable to humans) that are highly predictive of image classes (Ilyas et al. 2019). This makes it possible for many DL models to extract non-robust features, i.e., those that appear as nothing more than noise to human observers. In many cases, non-robust features are sufficient to train well-generalized models, even in cases involving semantic alteration to the dataset. An adversarial network could be used to facilitate model interpretation by humans, thereby revealing the features on which the model relies in order to determine whether they are legitimate disease features or interpretive shortcuts of little or no clinical relevance. However, our use of an adversarial attack to mediate the effects of race was not particularly effective; i.e., race-related traits were not entirely removed.

In the future, we will focus on applying the proposed approach to other imaging modalities (e.g., magnetic resonance imaging and computer tomography) and demographic

attributes (e.g., gender and age). We will also evaluate other image manipulation techniques in the context of shortcut reduction.

References

- Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; and Kim, B. 2018. Sanity checks for saliency maps. *arXiv preprint arXiv:1810.03292*.
- Baltruschat, I. M.; Nickisch, H.; Grass, M.; Knopp, T.; and Saalbach, A. 2019. Comparison of deep learning approaches for multi-label chest X-ray classification. *Scientific reports*, 9(1): 1–10. Publisher: Nature Publishing Group.
- Banerjee, I.; Bhimoreddy, A. R.; Burns, J. L.; Celi, L. A.; Chen, L.-C.; Correa, R.; Dullerud, N.; Ghassemi, M.; Huang, S.-C.; Kuo, P.-C.; Lungren, M. P.; Palmer, L.; Price, B. J.; Purkayastha, S.; Pyrros, A.; Oakden-Rayner, L.; Okechukwu, C.; Seyyed-Kalantari, L.; Trivedi, H.; Wang, R.; Zaiman, Z.; Zhang, H.; and Gichoya, J. W. 2021. Reading Race: AI Recognises Patient’s Racial Identity In Medical Images. *arXiv:2107.10356 [cs, eess]*. ArXiv: 2107.10356.
- Geirhos, R.; Jacobsen, J.-H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; and Wichmann, F. A. 2020. Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence*, 2(11): 665–673. ArXiv: 2004.07780.
- Goldberger, A. L.; Amaral, L. A.; Glass, L.; Hausdorff, J. M.; Ivanov, P. C.; Mark, R. G.; Mietus, J. E.; Moody, G. B.; Peng, C.-K.; and Stanley, H. E. 2000. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *circulation*, 101(23): e215–e220. Publisher: Am Heart Assoc.
- Ilyas, A.; Santurkar, S.; Tsipras, D.; Engstrom, L.; Tran, B.; and Madry, A. 2019. Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175*.
- Johnson, A. E.; Pollard, T. J.; Berkowitz, S. J.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.-y.; Mark, R. G.; and Horng, S. 2019a. MIMIC-CXR, a de-identified publicly available

database of chest radiographs with free-text reports. *Scientific data*, 6(1): 1–8. Publisher: Nature Publishing Group.

Johnson, A. E. W.; 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. 2019b. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv:1901.07042 [cs, eess]*. ArXiv: 1901.07042.

Jung, H.; and Oh, Y. 2021. Towards Better Explanations of Class Activation Mapping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1336–1344.

Kindermans, P.-J.; Hooker, S.; Adebayo, J.; Alber, M.; Schütt, K. T.; Dähne, S.; Erhan, D.; and Kim, B. 2019. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, 267–280. Springer.

Lei, Y.; Tian, Y.; Shan, H.; Zhang, J.; Wang, G.; and Kalra, M. K. 2020. Shape and margin-aware lung nodule classification in low-dose CT images via soft activation mapping. *Medical image analysis*, 60: 101628. Publisher: Elsevier.

Minaee, S.; Boykov, Y. Y.; Porikli, F.; Plaza, A. J.; Kehtarnavaz, N.; and Terzopoulos, D. 2021. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Publisher: IEEE.

Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.; and Shpan-skaya, K. 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Seyyed-Kalantari, L.; Liu, G.; McDermott, M.; Chen, I. Y.; and Ghassemi, M. 2020. CheXclusion: Fairness gaps in deep chest X-ray classifiers. In *BIOCOMPUTING 2021: Proceedings of the Pacific Symposium*, 232–243. World Scientific.

Singh, A.; Sengupta, S.; and Lakshminarayanan, V. 2020. Explainable deep learning models in medical image analysis. *Journal of Imaging*, 6(6): 52. Publisher: Multidisciplinary Digital Publishing Institute.

Zech, J. R.; Badgeley, M. A.; Liu, M.; Costa, A. B.; Titano, J. J.; and Oermann, E. K. 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15(11): e1002683. Publisher: Public Library of Science.