Eliminating race-related AI shortcuts from chest radiography analysis

Ryan Wang[1*], Li-Ching Chen[1*], Pei-Chuan Lin[1] , Judy Wawira Gichoya[2], Leo Anthony Celi[3,4], Po-Chih Kuo[1]

[1]Department of Computer Science, National Tsing Hua University

[2]Department of Radiology, Emory University

[3]Institute for Medical Engineering and Science, Massachusetts Institute of Technology

[4]Department of Medicine, Beth Israel Deaconess Medical Center

Introduction

- 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.
- The ease with which deep learning (DL) models identify racial features could potentially bias detection results since such features may be used as shortcuts.
- It is nearly impossible to eliminate race-related bias from diagnostic results.
- 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.

Methodology and Result

Phase 1. The chi-square test and logistic regression (LR) permutation test were used to identify race-induced deviations in radiological findings. 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. The image processing aimed at elucidating the means by which the DL models recognized the race of patients via chest x-ray (CXR) image analysis. Table 1 indicates that the outer outline of the lung is crucial to race identification.

Phase 3. The image transformations aimed at reducing the effect of race classification while maintaining accuracy in the detection of radiological anomalies. Table 2 shows the use of heavily rotated images for training had the most pronounced effect on mitigating the effects of race-related features.

Phase 4. Using true positive rate (TPR) disparity metrics to evaluate the accuracy of the proposed model in the detection of 14 disease classes. The results of TRP disparity analysis are presented in Figure 1 and the TPR disparity was lower for the proposed model (8 out of 14 classes and 9 out of 14 classes in MIMIC and Emory datasets, respectively) than original model. Figure 2 illustrates the heatmaps of the models and the removal of race-related characteristics indeed facilitated the learning of true radiological features instead of taking shortcuts.

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.
- The image rotation can be used to decrease the weight assigned to race-related features without compromising the performance of anomaly detection.
- The proposed training scheme was also shown to mediate the disparities in detection performance among races.
- To the best of our knowledge, this is the first study to introduce a method by which to mediate the influence of race-related physiological features in the interpretation of CXR images.

Table 1. The results of **Phase 2.**

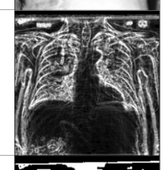
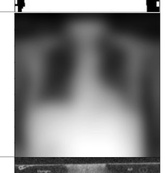
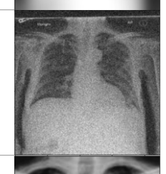
Image Processing Methods	AUC White	AUC Black	AUC Asian	Findings	Examples
Original	0.946	0.956	0.943	N/A	
Sobel filter	**0.929**	**0.940**	**0.921**	**Outlines matter.**	
Binary Otsu's thresholding	0.813	0.829	0.801	Pixel intensity does not matter.	
Binary Otsu's thresholding with erosion and dilation	0.779	0.790	0.772	Details of outlines are important.	
Gaussian blurring	**0.824**	**0.831**	**0.818**	**Outlines are important.**	
Gaussian noise addition	0.894	0.907	0.891	Pixel values do not matter.	
Central cropping (light/medium/heavy)	0.929 0.859 0.757	0.929 0.858 0.764	0.913 0.854 0.727	Less information equates to lower performance. **Excluding the outer outline of the lungs decreases performance**.	

Figure 1. The results of TPR disparity analysis, where bar length is proportional to the degree of disparity.

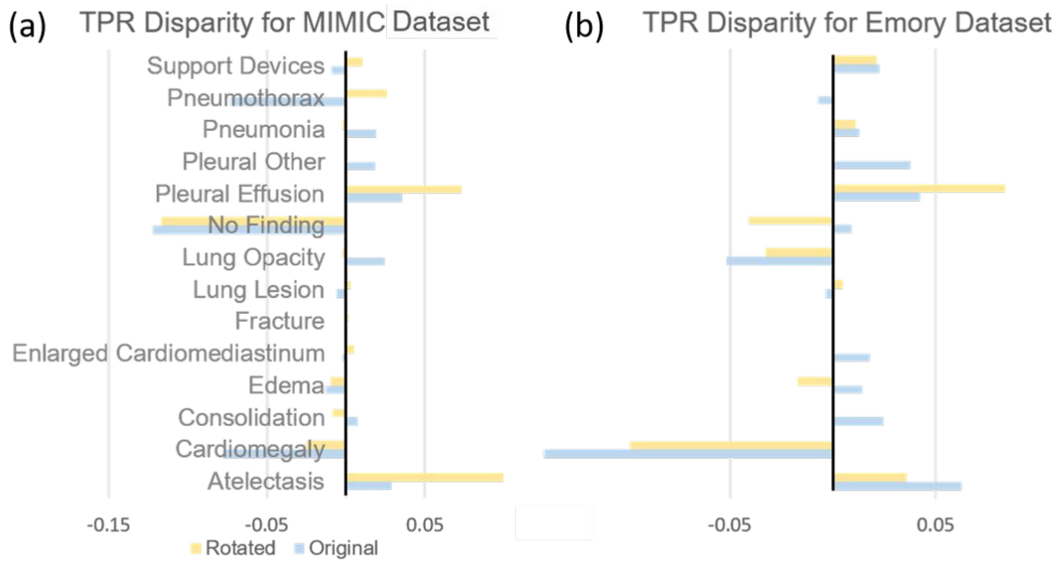


Figure 2. The comparison of heatmaps. The proposed model focused more on the disease characteristics than did the original model.

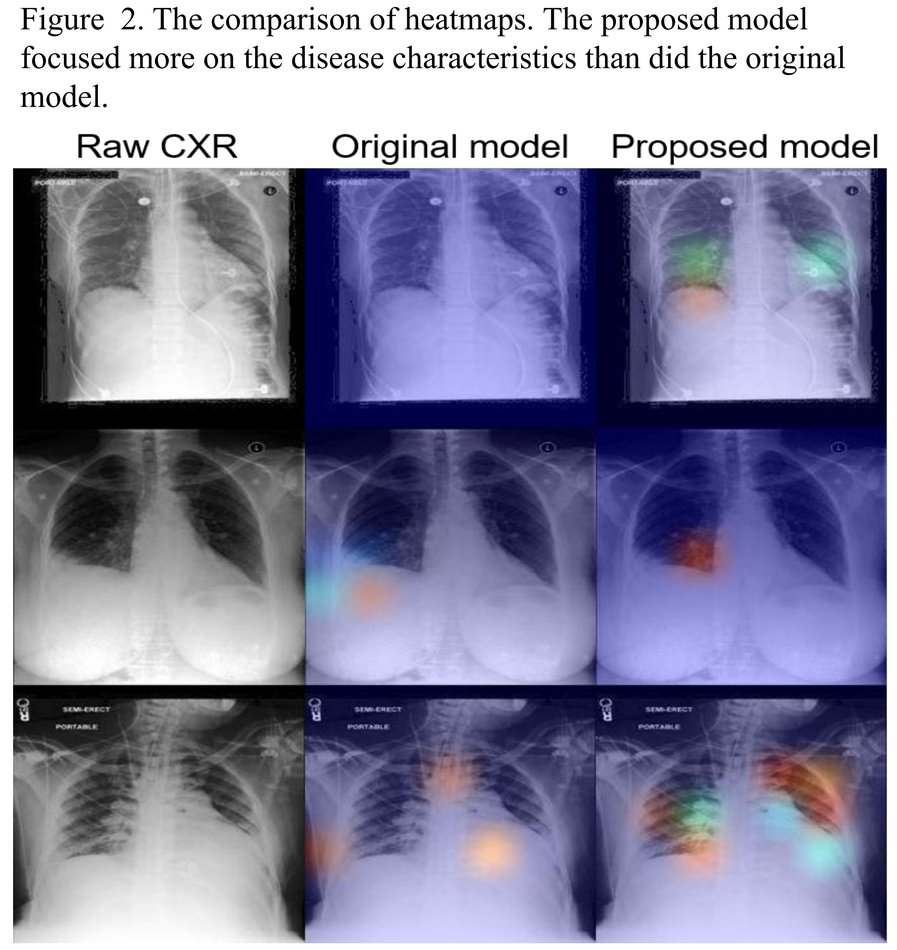
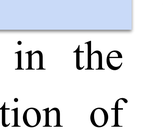
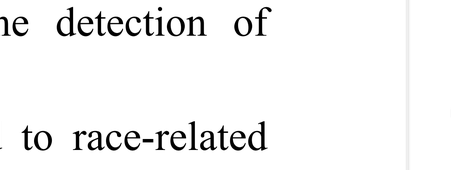
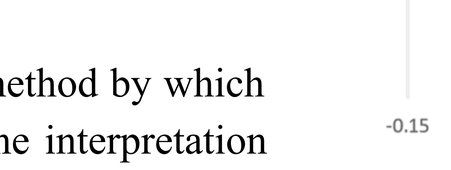


Table 2. The results of **Phase 3**.

Methods	AUC-White	AUC-Black	AUC-Asian	Averaged AUC-radiological findings	Examples
Original	0.946	0.954	0.943	0.769	
Rotating transformation (light/medium/**heavy**)	0.932/0.844/**0.816**	0.940/0.856/**0.825**	0.915/0.823/**0.781**	0.737/0.751/**0.752**	
Shear transformation (light/medium/heavy)	0.871/0.862/0.771	0.881/0.869/0.783	0.860/0.841/0.730	0.749/0.702/0.743	
Scaling transformation (light/medium/heavy)	0.938/0.879/0.806	0.948/0.915/0.817	0.934/0.865/0.770	0.760/0.751/0.744	
Fisheye distortion (light/medium/heavy)	0.926/0.916/0.907	0.939/0.926/0.919	0.919/0.909/0.890	0.767/0.757/0.747	