The racial effect in radiographs interpretation

Machine Learning In Healthcare (2022)

Sharon Peled

Introduction

Background

Previous studies in medical imaging have shown disparate abilities of deep learning models to detect a person's race from X-ray images.

Yet there is no known correlation for race on medical imaging that would be obvious to human experts when interpreting X-ray images.

Our work aims to understand the impact race has on X-ray medical imaging diagnostics in deep learning models.

Data

CheXpert	MIMIC-CXR
224,316 chest radiographs from Stanford Hospital	371,920 chest radiographs from BIDMC of Boston

Outline

- Performance gaps across protected groups
- Race prediction abilities across protected groups
- Race encoding through the network
- Validation on out-of-distribution dataset
- Short label comparison across datasets

Research

Perfomance per Protected Group

	Atelectasis	Cardiomegaly	Consolidation	Edema	Pleural Effusion	Mean
Asian	0.75, 0.84	0.95, 0.80	0.91, 0.79	0.89, 0.92	0.93, 0.91	0.89, 0.85
Black	0.74, 0.82	0.88, 0.80	0.81, 0.89	0.90, 0.92	0.92, 0.92	0.85, 0.87
Hispanic	0.74, 0.82	0.86, 0.83	0.83, 0.81	0.85, 0.93	0.94, 0.93	0.84, 0.86
White	0.76, 0.83	0.81, 0.83	0.81, 0.85	0.89, 0.91	0.90, 0.94	0.83, 0.87

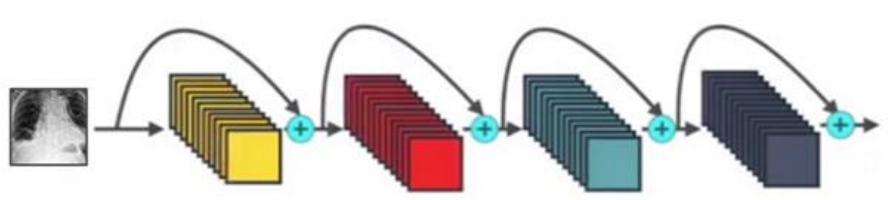
Table 1: Each cell presents the AUC score on Validation Set 2 (CXP, left) and Validation Set 3 (MXR, right). There isn't a clear bias across any racial group, even though the training data composed mostly from white patients.

Race Prediction

	Asian	Black	Hispanic	White	Mean
CXP	0.95	0.95	0.77	0.90	0.90
MXR	0.90	0.87	0.66	0.87	0.83

Table 2: AUC scores for race detection. Performance on MXR validation declined compared to the CXP validation. The model struggles with detecting Hispanics.

Race Encoding Through the Network



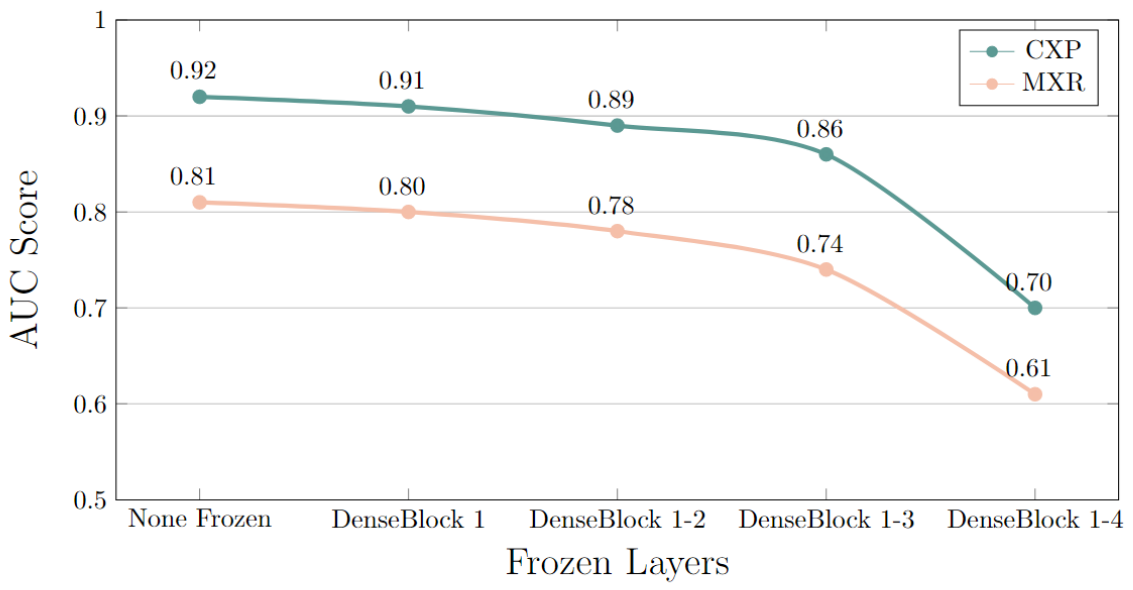


Figure 1 (left): DenseNet121 architecture. It consists of 4 denseblocks, whereas each block composed of multiple feature maps. If a denseblock is frozen, the gradients won't propagate through its layers, leaving it unchanged.

Figure 2 (right): AUC score for predicting race, such that an increasing prefix of the network is being frozen.

Conclusion

Findings

- No unreasonable bias across race, gender, or age.
- Elderly patients are harder to diagnose.
- Race information is encoded through the network, although its influence might not be significnt.
- When predicting race, Hispanics are harder to classify.
- CXP has significantly more uncertainties than MXR.

Limitations

- Using self-reported race as racial identity
- Genetic ancestry as potential confounder for detection of diseases
- Validation sets that had been automatically labeled

Discussion

We analyzed potential pitfalls of deploying deep neural networks.

Our finding that racial information propagates through the entire network, a network which was trained to find X-ray pathologies, a task that should have little to no correlation to race, creates a risk for deep models' deployment in real life settings.