Description of the problem you are tackling  
Background and related work  
Description of the data  
Method  
Results  
Limitations of the study  
Discussion

The racial effect in X-ray medical images  
Introduction  
The performance of deep learning models in the field of medical imaging has reached or even exceeded human-level performance, especially when it comes to diagnosing disease using chest X-rays. However, neural networks often learn to make predictions that overly rely on spurious correlation existing in the dataset, which causes the model to be biased. This kind of bias is often difficult to identify, due to the lake of explainability of such classifiers. As computer vision systems are deployed at scale in variety of settings, it becomes increasingly important to be aware of such drawbacks, especially in the medical domain.  
Previous studies in medical imaging have shown disparate abilities of deep learning models to detect a person’s race, yet there is no known correlation for race on medical imaging that would be obvious to human experts when interpreting the images.   
Our work aims to understand the impact race has on X-ray medical imaging diagnostics in deep learning models. To this extent, we’re utilizing two popular large-scale X-ray datasets: CheXpert and MIMIC-CXR.

More on problem [MISSING]  
More on experiments and methods [MISSING]

# Background and related work

**Bias.** Biased machine learning models is a topic of increasing attention, classifier biases have been discovered in various ranging domains, such as racial bias in criminal defendant risk assessments [17 MISSING] and gender bias in online recruiting [16 MISSING]. Studies in the medical domain have shown similar results in many health-care applications, such as mortality prediction [7 MISSING], and melanoma detection [5,6 MISSING].  
Sources of bias may originate in different points along the classical machine learning pipeline. Considering minority groups, collected features may not be as indicative compared to the rest of the population. This phenomenon of deep learning models to learn spurious correlations existing in the data is often a major obstacle in promising generalization. This is evident in the struggle of deep learning models to generalize on out-of-distribution data [MISSING].

The focus in our work [MISSING]

**Fairness.** Fairness has been conceptualized mathematically and philosophically in a variety of ways, such as error rate balance [MISSING], worst-case group accuracy [MISSING], and fairness through unawareness. There are several conflicting definitions of fairness, many of which are not simultaneously achievable, the appropriate choice of a disparity metric is generally task dependent. With the advancement of deep learning technology, artificial intelligence and decision support systems become more popular, the idea of unintentionally relying on protected attributes (such as race, gender, age, etc.) could be alarming, especially, in sensitive environments such as hospitals.

In our contest [MISSING]

**Protected Attributes Detection.** It’s been shown that deep learning algorithms can identify various patient demographic attributes, even when these do not form part of the input. For example, DK et al. [MISSING] utilized such abilities to improve a radiologist performance in skeletal age assessment. PH et al. [MISSING] Illustrated how such classifiers able to determine the age and sex of patients from chest radiographs. Furthermore, a study by Gichoya at el. demonstrates how race could be accurately identified from X-ray images alone, an ability that is unexplainable by physicians. Moreover, Gichoya at el. study shows that even with severe data augmentation the race identification abilities almost didn’t decrease.

The results from our study emphasise that the ability of AI deep learning models to predict self-reported race is itself not the issue of importance. However, our finding that AI can accurately predict self-reported race, even from corrupted, cropped, and noised medical images, often when clinical experts cannot, creates an enormous risk for all model deployments in medical imaging [MISSING]

Our work takes one step further, beside identify race, we wanted to see if diagnosis prediction models are encoding racial features when making predictions. [MISSING]

**Self-Reported Race.** Race identity often conflated with biological constructs, in our study we define race as a social construct that pertains to how we interact with each other and how others perceive us. To this end, we use self-reported race as the racial identity of patients throughout the study.

# The Data

**CheXpert.** (Irvin et al., 2019) [MISSING] An X-ray dataset from Stanford Hospital that contains 224,316 frontal and lateral chest radiographs of 65,240 patients.   
**MIMIC-CXR.** (Johnson et al., 2019) [MISSING] An X-ray dataset sourced from the Beth Israel Deaconess Medical Center between 2011 – 2016. The dataset consists of 371,920 chest X-rays associated with 227,943 imaging studies from 65,079 patients.  
**Demographics.** Both datasets include demographic data about patients, such as their gender, age, and self-reported race. Statistical aggregates, comparisons, and other statistics can be found in Appendix A [MISSING].  
**Labeled.** For both datasets a rule-based labeler (Irvin et al., 2019) [MISSING] was used to extract observations from a free text radiology report to create structured labels for the images. The labeler was design to automatically detect the presence of 14 observations in radiology reports, capturing uncertainties inherent in radiograph interpretation. Each of the defined 14 observations were categorized by the labeler into 4 classes: confidently present (1), confidently absent (0), uncertainly present (u), or not mentioned (blank). The labeler was evaluated on 1000 distinct randomly sampled patient studies that were annotated by two board-certified radiologists. Disagreements were resolved by consensus discussion.

# Experiments

Uncertainty imputation

Challenge labels

# References

[1]

16. M. De-Arteaga, A. Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik,K. Kenthapadi and A. T. Kalai, Bias in bios: a case study of semantic representation bias in a Pacific Symposium on Biocomputing242high-stakes setting (2019), Atlanta, GA.

17. A. Chouldechova, Fair prediction with disparate impact: A study of bias in recidivism predictioninstruments,Big data5, 153 (2016).

18. J. Buolamwini and T. Gebru, Gender Shades: Intersectional Accuracy Disparities in CommercialGender Classification,81, p. 15 (2018).

19. I. Chen, F. D. Johansson and D. Sontag, Why Is My Classifier Discriminatory?, inNIPS’31,2018 pp. 3539–3550.20. J. Kleinberg, S. Mullainathan and M. Raghavan, Inherent Trade-Offs in the Fair Determinationof Risk Scores,arXiv:1609.05807(2016).21. M. Srivastava, H. Heidari and A. Krause, Mathematical notions vs. human perception of fairness:a descriptive approach to fairness for machine learning,arXiv preprint arXiv:1902.04783(2019).22. D. S. Char, N. H. Shah and D. Magnus, Implementing machine learning in health care —addressing ethical challenges,New England Journal of Medicine378, 981 (2018).23. Z. Obermeyer and S. Mullainathan, Dissecting racial bias in an algorithm that guides healthdecisions for 70 million peoples, p. 89 (2019), Atlanta, GA.24. M. A. Gianfrancesco, S. Tamang, J. Yazdany and G. Schmajuk, Potential biases in machinelearning algorithms using electronic health record data.,JAMA internal medicine178, 1544(2018).25. S. Akbarian, L. Seyyed-Kalantari, F. Khalvati and E. Dolatabadi, Evaluating knowledge transferin neural network for medical images,arXiv preprint arXiv:2008.13574(2020).26. G. Huang, Z. Liu, L. v. d. Maaten and K. Q. Weinberger, Densely Connected ConvolutionalNetworks, 2261 (2017).27. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, ImageNet: A Large-Scale HierarchicalImage Database (2009).28. D. P. Kingma and J. Ba, Adam: A method for stochastic optimization,arXiv:1412.6980v9(2017).29. A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone and E. Ferrante, Gender imbalance inmedical imaging datasets produces biased classifiers for computer-aided diagnosis,Proceedingsof the National Academy of Sciences117, 12592 (2020).30. T. B. Hashimoto, M. Srivastava, H. Namkoong and P. Liang, Fairness Without Demographicsin Repeated Loss Minimization,arXiv:1806.08010(2018).31. R. G. J. Miller,Simultaneous Statistical Inference(Springer-Verlag New York, 1981).32. J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano and E. K. Oermann, Confoundingvariables can degrade generalization performance of radiological deep learning models,PLOSMedicine15, p. e1002683 (2018).33. N. Tomaˇsev, X. Glorot, J. W. Rae, M. Zielinski, H. Askham, A. Saraiva, A. Mottram, C. Meyer,S. Ravuri, I. Protsyuket al., A clinically applicable approach to continuous prediction of futureacute kidney injury,Nature572, p. 116 (2019

Yi PH, Wei J, Kim TK, et al. Radiology “forensics”: determination of age and sex from chest radiographs using deep learning. Emerg Radiol 2021; 28: 949–54.

12

Eng DK, Khandwala NB, Long J, et al. Artificial intelligence algorithm improves radiologist performance in skeletal age assessment: a prospective multicenter randomized controlled trial. Radiology 2021; 301: 692–99. 13 Rim TH, Lee G, Kim Y, et al. Prediction of systemic biomarkers from retinal photographs: development and validation of deeplearning algorithms. Lancet Digit Health 2020; 2: e526–36.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al., 2019. Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. arXiv:1901.07031.