JMI Journal of Medical Imaging

RESEARCH PAPER

Ability of artificial intelligence to identify self-reported race in chest x-ray using pixel intensity counts

John Lee Burns,[a,*] Zachary Zaiman,[b] Jack Vanschaik,[a] Gaoxiang Luo,[c] Le Peng,[c] Brandon Price,[d] Garric Mathias,[a] Vijay Mittal,[a] Akshay Sagane,[a] Christopher Tignanelli,[c] Sunandan Chakraborty,[a] Judy Wawira Gichoya,[b] and Saptarshi Purkayastha[a]

[a]Indiana University, School of Informatics, Indianapolis, Indiana, United States
[b]Emory University, School of Medicine, Department of Radiology, Atlanta, Georgia, United States
[c]University of Minnesota Twin Cities, Medical School, Minneapolis, Minnesota, United States
[d]University of Florida College of Medicine, College of Medicine, Gainesville, Florida, United States

ABSTRACT. **Purpose:** Prior studies show convolutional neural networks predicting self-reported race using x-rays of chest, hand and spine, chest computed tomography, and mammogram. We seek an understanding of the mechanism that reveals race within x-ray images, investigating the possibility that race is not predicted using the physical structure in x-ray images but is embedded in the grayscale pixel intensities.

Approach: Retrospective full year 2021, 298,827 AP/PA chest x-ray images from 3 academic health centers across the United States and MIMIC-CXR, labeled by self-reported race, were used in this study. The image structure is removed by summing the number of each grayscale value and scaling to percent per image (PPI). The resulting data are tested using multivariate analysis of variance (MANOVA) with Bonferroni multiple-comparison adjustment and class-balanced MANOVA. Machine learning (ML) feed-forward networks (FFN) and decision trees were built to predict race (binary Black or White and binary Black or other) using only grayscale value counts. Stratified analysis by body mass index, age, sex, gender, patient type, make/model of scanner, exposure, and kilovoltage peak setting was run to study the impact of these factors on race prediction following the same methodology.

Results: MANOVA rejects the null hypothesis that classes are the same with 95% confidence (F 7.38, $P < 0.0001$) and balanced MANOVA (F 2.02, $P < 0.0001$). The best FFN performance is limited [area under the receiver operating characteristic (AUROC) of 69.18%]. Gradient boosted trees predict self-reported race using grayscale PPI (AUROC 77.24%).

Conclusions: Within chest x-rays, pixel intensity value counts alone are statistically significant indicators and enough for ML classification tasks of patient self-reported race.

 [DOI: 10.1117/1.JMI.10.6.061106]

Keywords: machine learning; bias; population imaging; x-ray

Paper 23030SSR received Jan. 30, 2023; revised May 15, 2023; accepted Jul. 18, 2023; published Aug. 4, 2023.

*Address all correspondence to John Lee Burns, jolburns@iu.edu

2329-4302/2023/$28.00 © 2023 SPIE

1 Introduction

It is trivial for convolutional neural networks (CNN) to predict self-reported race within medical imaging. Radiologists are not trained or proven capable of performing this task; what is learned by CNN is not clear.[1,2] Artificial intelligence (AI) can detect race from medical images, and computer vision-based AI models can unknowingly integrate racial biases into prognostic or treatment algorithms.[3] There is potential for discriminatory harm if we assume that AI models are agnostic to race—understanding the relationship between race and medical imaging AI models is important.[4–8]

There are no known imaging biomarker correlates for racial identity; however, medical imaging AI models produce racial disparities.[9,10] Prior work sought to answer how AI systems could produce disparities in multiple medical imaging modalities. Within chest x-rays, AI models can predict self-reported race with an area under the receiver operating characteristic (AUROC) of 0.974.[1] Gichoya et al.[1] showed that the features learned appear to involve all regions of the image and frequency spectrum, suggesting that mitigation efforts will be challenging.

We seek an understanding of the mechanism that reveals race within medical imaging by investigating the possibility that race predicting features may be embedded within the individual grayscale pixel intensities of an x-ray image. We remove all image structures by counting how many times each grayscale value appears, testing for statistical differences between the pixel intensities within race groups, and training machine learning models to predict race using these grayscale counts. Although this method removes the structure of the image, the presence of body habitus can remain encoded in this representation. We investigate possible confounders of body habitus using body mass index (BMI) as well as modality configuration settings by limiting the device to a single make/model and controlling for kilovoltage peak (KVP) and exposure.

2 Approach

The dataset consists of three academic health centers (AHC) and one publicly available dataset, MIMIC-CXR.[11] Dataset population descriptions are described in Table 1; all use self-reported race, are front-view AP/PA chest x-rays, and were collected between 1/1/2021 and 12/31/2021 (except MIMIC-CXR[11]). AHC 1, Indiana University School of Medicine in Indianapolis, has two datasets—uncontrolled hospital W (1.1) and one year at hospitals X, Y, and Z (1.2) limited to the top 10% of diverse x-ray devices, defined as the devices with the largest percent of non-White patients. AHC 2, Emory University in Atlanta, has five datasets—uncontrolled (2.1) and four limited to one device make and model (Carestream DRX-Revolution[12]) categorized by BMI—underweight (2.2), normal (2.3), overweight (2.4), and obese (2.5). AHC 3, University of Minnesota in Minneapolis, has one uncontrolled dataset.[3] Overall, 298,827 images are included in the analysis. All institutions acquired IRB approval with waiver of consent and de-identified datasets prior to processing. All institutional data were collected retrospectively without control to pathologies present.

KVP, exposure, and modality information are extracted from DICOM headers, and then the images are converted from DICOM format to 8-bit grayscale PNG format. 8-bit grayscale format was chosen to match the MIMIC-CXR format.[11] No windowing, leveling, or grayscale normalization are applied to images during conversion. When photometric interpretation equals "MONOCHROME1," images are grayscale inverted. The conversion of local DICOM files was done with a function of [(pixel_grayscale_value/overall_image_max_grayscale_value)*255]. Images are then converted into a data frame, with columns of grayscale values from 0 to 255 and race and row values being the number of pixels appearing in the image with that value. The zero-grayscale value is dropped as this value has high variance and often only appears due to postprocessing, such as image rotations. Grayscale pixel counts are converted to percent per image (PPI), normalizing for resolution of the image. The code for this process is included in the linked Github repository.

2.1 Statistical Methods

Multivariate analysis of variance (MANOVA) and subsampled class balanced MANOVA are run on all datasets and combined datasets. The test hypothesis is that groups contain differences in pixel values. Results are analyzed for significance of 95% ($p < 0.05$) and F-value > 2. MANOVA results include Bonferroni multiple-comparison adjustment at an $\alpha = 0.05$, and

Table 1 Dataset population characteristics. Female (F) and male (M) are presented when possible, and total image count for race (T), where total is different than persons (MIMIC). Age is presented as (mean, median, standard deviation).

Dataset	Asian F|M|T	Asian age	Black F|M|T	Black age	Hispanic F|M|T	Hispanic age	White F|M|T	White age
1.1	118|191	35.8, 27, 28.6	860|822	40.4, 41, 25.7	323|399	30.1, 25, 25.7	851|1007	53.2, 58, 23.5
1.2	316|288	57.3, 62, 19.2	981|849	49.5, 49, 18.7	329|467	46.9, 44, 19.4	967|1032	60.5, 63, 17.1
2.1	4180	Not available	5208	Not available	0|0	Not applicable	5207	Not available
2.2	77|132	55.9, 70, 25.1	913|1362	58.9, 70, 19.52	19|64	58.7, 70, 19.3	1114|1196	63.1, 70, 20.6
2.3	393|753	66.1, 70, 15.5	3546|5107	59.9, 70, 18.7	172|251	62.8, 70, 19.1	4711|6308	68.3, 70, 17.7
2.4	202|121	66.5, 70, 26.3	7206|4577	61.7, 70, 16.7	269|194	64.2, 70, 18.8	5369|7643	68.3, 70, 15.7
2.5	265|427	63.4, 70, 15.7	3559|4033	59.3, 70, 15.8	182|238	64.4, 70, 13.7	3961|7824	64.9, 70, 15.1
3	570|474	54.7, 54, 20.1	1467|1048	46.6, 46, 17.2	0|0	Not applicable	2464|1838	60.8, 62, 20.2
MIMIC	1002|879|7106	58.8, 61, 18.9	5378|3194|34,238	54.5, 55, 17.5	1722|1379|11,166	50.3, 50, 16.9	16,220|16,536|141,873	62.5, 64, 18

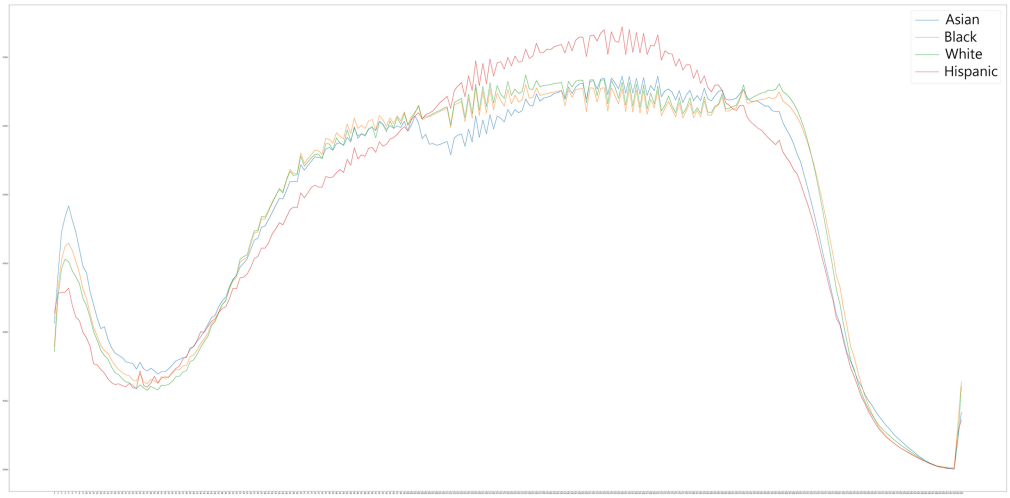

Fig. 1 Grayscale histogram for the combined–no MIMIC dataset. *X* represents grayscale values between 1 and 255, and *Y* represents the PPI.

significant p values must be <0.0038. Histogram line charts were created describing the grayscale PPI by race (Fig. 1).[13,14]

2.2 Visualizing Grayscale Value Presentations in Chest X-ray

To visualize the statistical differences in distributions of grayscale pixels between the groups, we plotted a grouped histogram using the D3.js v3 library.[15] We used a random sample of 500 images from each race at AHC 1. Filters are utilized to segment data by race, sex, and grayscale range. When filtering by range, the chest x-ray image embedded within the page highlights in yellow the current area of the image. Figure 2 highlights regions of interest within the histogram. The raw individual image data are plotted in a multi-line plot, with a filter for how many lines are shown. ANOVA test results are listed by pixel value, and bar charts represent the filtered population age and sex by race.

2.3 Machine Learning Methods

KerasTuner is used on the combined dataset to determine the best hyperparameters of feed-forward networks (FFN) classifying race.[16] 10% of data is randomly withheld as a test dataset. The tuning process uses the validation AUROC on a validation set consisting of 20% of the training data as the metric to tune on. The tuning process trials 500 models of dense layer (DL) 2 to 10 depth, DL width (512 to 4096), activation functions (relu, tanh, and sigmoid), regularization layers (dropout, l2, and batch normalization), and Adam optimizer run at 60 epochs each. Multi-class, binary Black or White, and binary Black or other classification models were tuned. Multi-class classification failed to achieve over 55% validation set AUROC in any model and was not utilized for further tests. Black or White achieved a validation set AUROC of 68.47%, and Black or other achieved the highest AUROC of 69.51%. Model descriptions, package versions, and performance metrics are included in Appendix A.

The resulting model is retrained on each dataset separately, with a random data split of 10%/80%/20% for test/training/validation. Categorical cross-entropy is used for multi-class and binary cross-entropy for binary classification. Early stopping for minimum validation loss is utilized to stop training. Each dataset is trained and evaluated once for each classification problem using binary accuracy over all samples and AUROC.

Random forest (RF), gradient boosted trees (GBT), and cart models were trained on each dataset, with 80% training data and 20% testing. RF and GBT utilized the Keras hyperparameter template "benchmark_rank1," and cart utilized the default Keras cart settings.[17]

Using the combined single modality datasets (2.2 through 2.5), controls are applied for KVP (KVP = 125 and $n = 38{,}102$) and exposure (mAs = 1 to 4 and $n = 39,795$) with a combined $n = 28{,}381$ samples.[18] This dataset includes bucketed age, bucketed BMI, gender, and patient type (emergency, inpatient, and outpatient). Using this controlled dataset, we tested the race

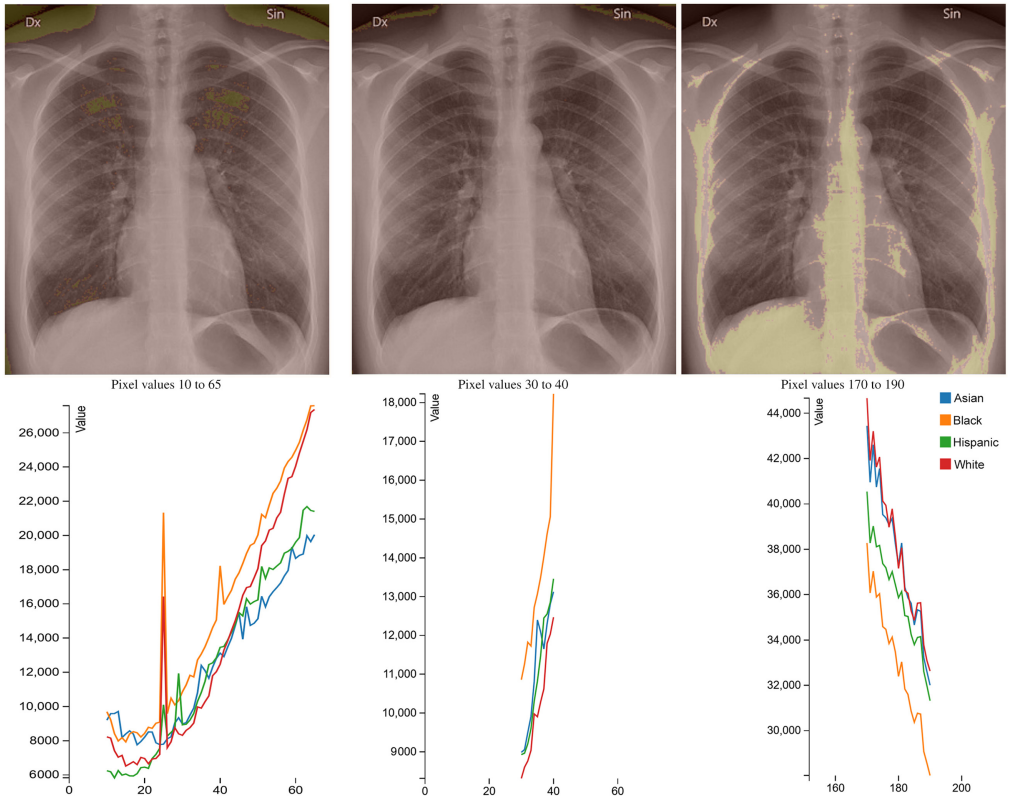

Fig. 2 Regions of interest within the histogram, visualized in a sample chest x-ray. Yellow highlighting in images shows current pixel range selected. Pixel values 10 to 65 have more pixels for Black patients and appear to correlate with background, skin/muscle, and some lung areas. Values 30 to 40 are minimal soft tissue, though that does vary within any given chest x-ray. Values 170 to 190 correlating to bone and some organ systems and less pixels on average for Black patients. Note: due to the overlaid nature of x-ray, there is no direct correlation between grayscale value and body regions as in computed tomography with houndsfield units. Chest x-ray image sourced from Wikimedia Commons under Creative Commons CCO 1.0 Universal Public Domain Dedication.

prediction tasks as well as age, BMI, gender, and patient type prediction. Additionally, we applied the controls (KVP = 120, mAs = 1 to 4, age = 60 to 80, gender = male, patient type = inpatient, and $n = 5718$) with the race prediction tasks. A final test was done on the full images using these controls and prediction tasks, following the methodology of Ref. 1.

3 Results

3.1 Statistical Results

ANOVA assumes that variables are uncorrelated, and a correlation matrix is created and assessed for correlations (Fig. 3). Many pixel counts appear to be highly correlated with other pixel counts. MANOVA is more appropriate as it accounts for correlations between variables. To ensure the validity of the MANOVA test in this setting, we conducted tests against several random splits of population subgroups for each dataset. None of these were significant after multiplicity correction (Sec. 5.3), implying that each race group followed a consistent distribution, so the following across-group tests will detect differences due to race and not due to sampling. All MANOVA tests have significant P values (with Bonferroni multiple-comparison adjustment significant P values < 0.0038), indicating that for all source datasets, the pixel percentage distribution is significantly different across different races. Balanced MANOVA tests have significant P values except datasets 1.1, 2.2, and 2.3. Table 2 describes dataset MANOVA results.

Single make/model modality controlled for KVP/exposure MANOVA results are listed in Table 3 (unbalanced) and Table 4 (balanced). MANOVA results show that all results are significant ($P < 0.05$).

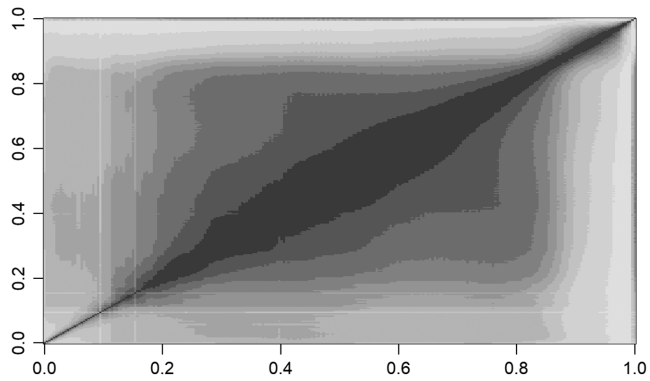

Fig. 3 Pixel correlation matrix, generated using R, showing the highly correlated nature of grayscale values and reasoning for MANOVA testing. *X* and *Y* scales are grayscale values (0 to 255) normalized to between 0 and 1. Darker regions are highly correlated pixel values.

Table 2 MANOVA results by dataset. DF, degrees of freedom.

		MANOVA		MANOVA balanced	
Dataset	DF	*F*-value	*P*-value	*F*-value	*P*-value
1.1	762	1.49	<0.0001	1.14	0.0119
1.2	762	1.22	<0.0001	1.16	0.0031
1–all	762	1.64	<0.0001	1.30	<0.0001
2.1	508	3.23	<0.0001	2.99	<0.0001
2.2	762	1.38	<0.0001	0.93	0.7726
2.3	762	2.65	<0.0001	1.13	0.0140
2.4	762	2.88	<0.0001	1.36	<0.0001
2.5	762	2.73	<0.0001	1.18	0.0015
2–all	762	7.38	<0.0001	2.02	<0.0001
3	508	2.58	<0.0001	1.67	<0.0001
MIMIC	762	7.04	<0.0001	2.90	<0.0001
Combined–no MIMIC	762	8.63	<0.0001	3.37	<0.0001
Combined–all	762	35.64	<0.0001	11.07	<0.0001

Table 3 MANOVA *N*, *F*, and *P* values for unbalanced and controlled tests. Single make/model of modality, KVP = 120, exposure mAs = 1 to 4.

		KVP			Exposure			Both			Uncontrolled		
Task	DF	*N*	*F*	*P*	*N*	*F*	*P*	*N*	*F*	*P*	*N*	*F*	*P*
Black or all	762	26,925	2.48	<0.0001	70410	5.16	<0.0001	26,387	2.44	<0.0001	72,188	5.31	<0.0001
Black or White	254	25,215	3.28	<0.0001	66743	8.46	<0.0001	24,692	3.20	<0.0001	68,429	8.59	<0.0001
Age	762	26,890	3.52	<0.0001	70078	6.05	<0.0001	26,352	3.45	<0.0001	71,849	6.22	<0.0001
Gender	254	26,925	13.79	<0.0001	70410	23.85	<0.0001	26,387	13.59	<0.0001	72,188	24.00	<0.0001
BMI category	762	26,925	17.15	<0.0001	70410	39.34	<0.0001	26,387	16.57	<0.0001	72,188	40.67	<0.0001
Patient type	508	26,925	6.42	<0.0001	70410	13.71	<0.0001	26,387	6.29	<0.0001	72,188	14.01	<0.0001

Table 4 MANOVA *N*, *F*, and *P* values for balanced and controlled tests. Single make/model of modality, KVP = 120, exposure mAs = 1 to 4.

		KVP			Exposure			Both			Uncontrolled		
Task	DF	*N*	*F*	*P*	*N*	*F*	*P*	*N*	*F*	*P*	*N*	*F*	*P*
Black or all	762	2820	1.42	<0.0001	5444	1.86	<0.0001	2788	1.41	<0.0001	5556	1.90	<0.0001
Black or White	254	15,038	2.70	<0.0001	59,040	7.65	<0.0001	14,660	2.57	<0.0001	60,606	7.76	<0.0001
Age	762	6720	1.93	<0.0001	27,372	3.93	<0.0001	6616	1.91	<0.0001	28,108	4.04	<0.0001
Gender	254	23,002	11.87	<0.0001	62354	21.45	<0.0001	22,680	12.15	<0.0001	63,754	21.39	<0.0001
BMI category	762	5900	4.95	<0.0001	19,196	11.84	<0.0001	5868	5.03	<0.0001	19,508	12.62	<0.0001
Patient type	508	4530	1.88	<0.0001	8946	2.88	<0.0001	4476	1.95	<0.0001	9099	2.68	<0.0001

3.2 Visualizing Results

Grayscale histograms were created for each dataset, and a subsample is visualized and available for browsing in Ref. 19. Features of the visualization website are shown in Fig. 4.

3.3 Machine Learning Results

FFN and decision tree results are listed in Table 5. In general, model performance follows dataset size. For binary Black or White classification, the best model is RF on dataset 3 with an accuracy of 70.5 and AUROC of 74.1. The full dataset GBT performs better than all other datasets and models, with an accuracy of 75.6 and AUROC of 70.4. For binary Black or all classification, the best model is GBT on the full dataset with an accuracy of 68.5 and AUROC of 77.2.

Single modality/body habitus models show better results than the combined models in some cases for FFN; however, for decision trees, this does not happen. Both Black or White/Black or all FFN experiments on Institution 2 data show that the overall combined dataset performs slightly worse (FFN AUROC 64.5/63.4) than some of the single modality (FFN AUROC 65.3/64.6). However, we see the opposite relationship with a better overall performance with decision trees on the full dataset (RF AUROC 69.6/68.8) compared with the single modality best performance (RF AUROC 67.3/67.6).

Single make/model modality controlled for KVP/exposure MANOVA FFN results are listed in Table 6. For race prediction tasks, controlling for KVP significantly improves model performance, whereas controlling for exposure has a similar performance, and controlling for both decreases model performance. Of the other tasks, the gender prediction performs best (AUROC 76.5) when controlled for KVP. All other tasks failed to accurately predict. Race prediction when fully controlled (single make/model of modality, KVP, exposure, patient type, gender, and age), listed in Table 7, shows improved performance compared with dataset size.

The full image CNN tests are listed in Table 6 and have high AUROC (0.99) in predicting race and gender. Age and patient type are predictable, whereas BMI is not. In these tasks, there does not appear to be any variation when controlling for KVP, exposure, or both when utilizing the full image.

4 Conclusions

4.1 Overall Conclusion

MANOVA results show a statistically significant relationship between grayscale PPI and race. Visualization of this data proved critical for analysis and idea generation. Presenting the PPI average alongside a chest x-ray image and controls for filtering by grayscale value allowed us to quickly communicate with radiologists in a format that they understood. The interpretation

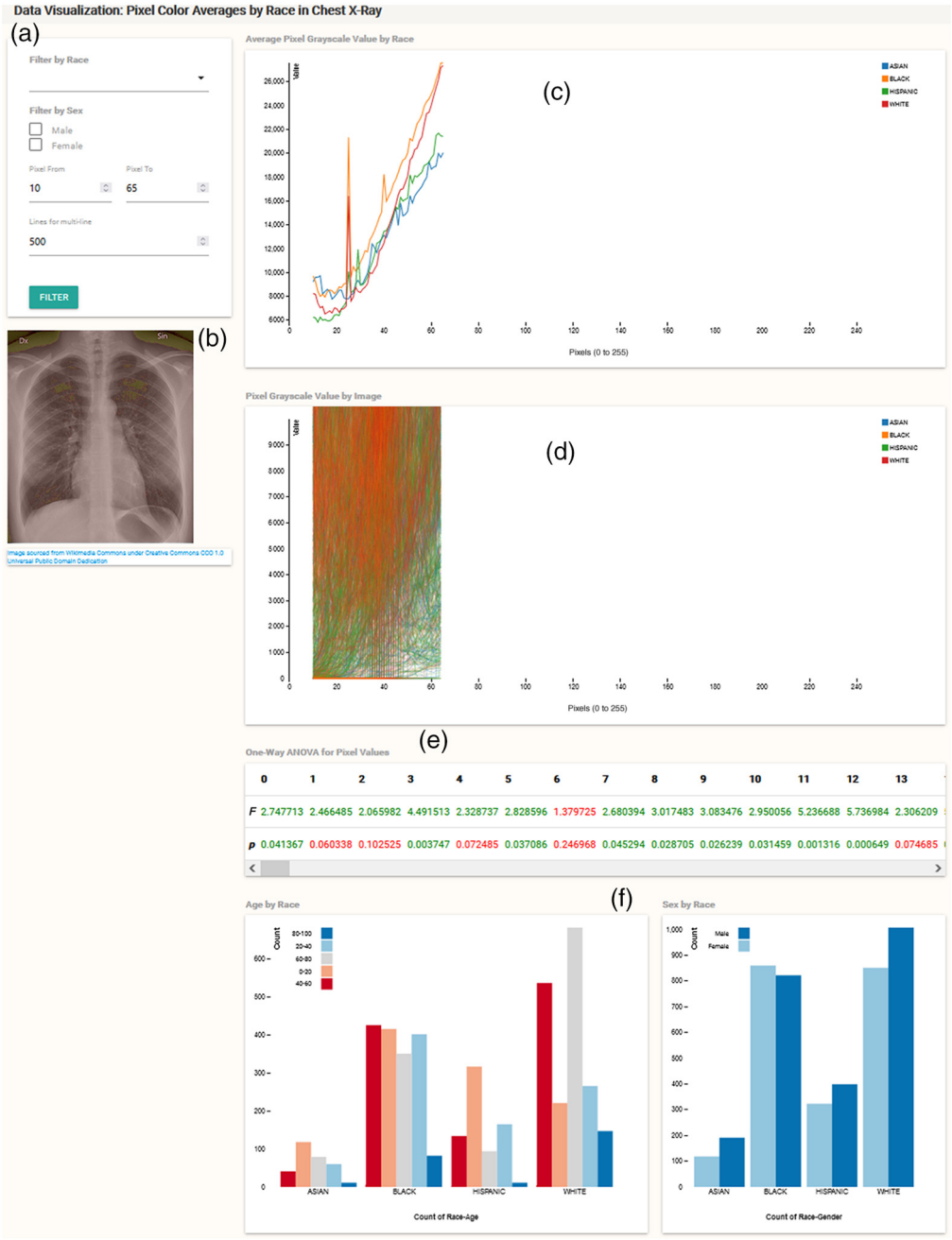

Fig. 4 Web visualization hosted in Ref. 19 of a subsampled dataset. (a) Image filters affect all charts and images. (b) Chest x-ray image filtered to show highlighted pixel range in yellow. (c) Average line graph by race, (d) multi-line by image colored by race, (e) ANOVA results for dataset, and (f) age/sex of dataset by race.

of where race data may exist, specifically areas linked to body habitus and BMI, informed model building decisions.

FFN were unable to accurately predict self-reported race from uncontrolled PPI. The best model achieved an AUROC of 69.18% using the full dataset. There is a possibility that additional data would increase AUROC and accuracy as these metrics generally went up as data size increased. Decision trees had better success in predicting self-reported race from PPI, having a higher AUROC than FFN in all but three cases. Utilizing the full dataset, GBT achieved an AUROC of 77.24% on a withheld test set.

Table 5 FFN/decision tree test set accuracy and AUROC by dataset and classification type. Bold cells indicate best performing model in each type.

	Binary Black or White					Binary Black or all				
	FFN		Decision tree [RF|GBT|cart]			FFN		Decision tree [RF|GBT|cart]		
Dataset	Accuracy	AUROC	Top model	Accuracy	AUROC	Accuracy	AUROC	Top model	Accuracy	AUROC
1.1	60.9	63.2	RF	57.3	60.6	62.9	57.9	RF	65.3	58.3
1.2	58.8	59.8	GBT	57.1	61.5	63.6	54.1	RF	65.9	62.4
1–All	57.1	58.2	RF	60.5	66.8	64.5	58.0	RF	66.3	63.2
2.1	60.6	63.5	RF	63.4	67.7	67.5	64.6	RF	66.3	66.6
2.2	60.6	64.6	RF	64.0	67.3	54.5	52.9	RF	61.4	65.1
2.3	63.2	62.0	RF	63.5	67.3	62.5	64.4	RF	65.1	66.8
2.4	59.7	62.2	RF	65.2	65.5	65.3	64.6	RF	66.0	67.6
2.5	62.2	65.3	RF	61.6	64.7	60.4	62.9	RF	62.9	66.4
2–All	61.7	64.5	RF	64.7	69.6	62.5	63.4	RF	65.9	68.8
3	67.4	67.9	**RF**	**70.5**	**74.1**	68.5	66.1	RF	71.9	72.6
MIMIC	80.5	61.2	GBT	80.4	61.7	82.4	60.2	GBT	82.3	60.0
Combined–no MIMIC	58.4	62.5	GBT	63.0	66.8	61.2	62.7	GBT	64.3	65.8
Combined–all	**75.0**	**69.2**	GBT	75.6	70.4	**77.0**	**68.4**	**GBT**	**68.5**	**77.2**

Table 6 Using pixel PPI—FFN test set accuracy and AUROC or macro *F*1 (age, BMI, and patient type) for controlled tests. Using full image, CNN test set accuracy and AUROC or macro *F*1 (age, BMI, and patient type) for controlled tests. Single make/model of modality, KVP = 120, exposure mAs = 1 to 4.

	KVP		Exposure		Both		Uncontrolled	
Task–PPIFFN	Accuracy	AUROC/*F*1	Accuracy	AUROC/*F*1	Accuracy	AUROC/*F*1	Accuracy	AUROC/*F*1
Black or all	72.1	75.2	63.3	68.1	42.7	39.8	62.7	66.8
Black or White	69.8	73.5	62.2	66.2	58.8	61.8	60.4	65.2
Age	19.2	10.4	15.8	9.3	15.2	9.8	49.7	21.6
Gender	69.2	76.5	32.4	25.4	64.8	70.6	68.0	74.6
BMI category	42.1	32.2	35.6	26.5	35.0	27.4	14.8	13.4
Patient type	67.5	39.1	26.5	17.7	71.3	41.7	27.0	17.2
Task–full image CNN								
Black or all	97	99	96	99	97	99	96	99
Black or White	97	99	96	99	96	99	96	99
Age	73	72	74	72	73	72	74	72
Gender	99	99	99	99	99	99	99	99
BMI category	48	43	48	43	47	43	49	44
Patient type	83	63	84	61	83	63	84	61

Table 7 FFN/decision tree test set accuracy and AUROC for controlled tests. Single make/model of modality, KVP = 120, exposure mAs = 1 to 4, patient type: inpatient, gender: male, and age: 60 to 80.

	FFN		Decision tree [RF|GBT|cart]		
Task	Accuracy	AUROC	Top model	Accuracy	AUROC
Black or all	64.0	67.1	RF	63.9	64.7
Black or White	63.8	68.6	RF	63.7	64.4

4.2 Controlled/Alternate Bias Factors Tests

There is some evidence in this data that modality configurations or BMI are correlated to model performance. Single institution models do appear to perform better, with 1.1 (single hospital in network) outperforming 1.2 (three other hospitals in same network) and 3 (single site AHC) performing well in comparison with the multi-site AHCs. Potentially, there is less effect of specific modality configurations and more effect toward hospital specific protocols and population.

Controlling for KVP within the single make/model of the modality dataset significantly improved FFN performance in race prediction tasks and controlling for exposure has no effect on model performance. Controlling for both KVP and exposure reduces performance; however, this follows the pattern of smaller dataset sizes having a reduction in performance seen across all datasets. This pattern is broken when comparing the fully controlled (KVP, exposure, patient type, gender, age, and $n = 5{,}718$) versus the uncontrolled ($n = 72{,}188$). The fully controlled FFN perform as well as or better than the uncontrolled.

We are unable to predict age, BMI category, or patient type following this methodology. We had success in predicting gender in the controlled dataset tests. We did not have these fields individually annotated in the other datasets and were unable to test this on a larger scale.

4.3 Limitations and Comparison with Prior Work

This study is limited to retrospective analysis of patients blinded to present disease. It should be noted that, although we removed the image structure, we did not entirely removed the physical structure—aspects of body habitus remaining embedded within this information. There is a chance that we are not picking up on a feature like skin tone but population metrics such as obesity that are observable within chest x-ray. Future research in this area following a prospective methodology, controlling patient factors such as BMI, disease, and limiting to a single modality, configured the same for each scan, and operated under the same protocol may be warranted. Additional tests with other body parts and modalities (CT/MRI/etc) are warranted and could limit the effect of other confounders. For example, using CT imaging could allow for segmentation of regions of interest, such as skin, and performing similar analysis.

Prior work utilized CNN and the full image to achieve high AUROC in race prediction.[1] Following their methodology, we found similar success in classifying race/gender and could classify age/patient type with less accuracy. We were unable to identify the BMI category using this methodology.

Our intent was to investigate the low-pass/high-pass filter and resolution reduction findings of AI recognition of patient race in medical imaging as a modeling study.[1] The low- and high-pass findings indicate that racial information existed on both ends of the grayscale spectrum, whereas the resolution reduction showed that the image structure could play less of a role than average grayscale values. In both cases, it was demonstrated that race was still predictable, even when humans no longer could identify the image as an x-ray. Our work expands on this—completely removing the image structure and attempting to predict race from simple grayscale value counts.

We are unable to predict self-reported race using grayscale values alone with the same accuracy as prior full-image work. However, CNN utilize features of the image, and it was expected that performance would decrease when the image structure was removed. GBT can interpret this data, showing that there is predictive value in grayscale PPI for self-reported race. It is not clear

that grayscale PPI is a factor in what the prior study CNNs learned for the race prediction task, but we have shown that it is possible to remove the image structure entirely and perform this task.

For both internal and publicly available datasets, race is deeply embedded in chest x-ray images in ways that are not obvious to human observers.

5 Appendix A: Model and Computational Setup Details

5.1 Keras Tuner Hypermodel Outcomes

The best models created using the Keras Tuner process for each task are described below. Both used the Adam optimizer with learning rate = 0.0001, beta_1 = 0.9, beta_2 = 0.999, and epsilon $= 1 \times 10^{-7}$.

The black or White model is as follows: input size 255, DL 1024, DL 2048 activation tanh, DL 2048 activation relu, dropout value 0.01, kernel regularizer l2 value 0.0001, and DL size 1 with activation sigmoid.

The black or all model is as follows: input size 255, DL 1024, 2× (DL 1024 activation relu), DL 1024 activation tanh, dropout value 0.01, kernel regularizer l2 value 0.0001, DL 1024, and DL size 1 with sigmoid activation.

5.2 Computational Setup

All analysis was completed on a system consisting of an Intel Xeon E5-2609 v4 CPU, 128 GB RAM, 4× GeForce RTX 2080Ti, and 4× GeForce GTX 1080. Python 3.9.7 and libraries Numpy 1.19.2, Pandas 1.1.3, Pillow 8.0.1, Pydicom 2.1.2, Scipy 1.5.2, and Matplotlib 3.3.2 are utilized for conversion of images and histogram plotting. R 4.1.1 was utilized for correlation plots and MANOVA. Model training and evaluation utilized Python 3.8.10 and libraries Scikit-learn 0.23.1, Pandas 1.3.1, Numpy 1.19.5, Keras 2.6.0, and GPUtil 1.4.0.

Training and evaluation run-time varies between 9 and 205 s, using a maximum of 5.05 GB RAM, 20% of up to 6 processor cores, and a single 2080ti GPU.

5.3 MANOVA Random Subsampling

For each dataset, data are split into race subgroups, and each subgroup is randomly split in half via a dummy variable. Then MANOVA is performed against the dummy variable. This is repeated 5 times for each subgroup. Results of this subsampling analysis are included in Table 8. After multiplicity correction (alpha = 0.000208), there were no significant tests. This is exactly what we would expect, confirming that the MANOVA tests were indeed reliable.

Table 8 MANOVA random subsampling analysis to determine if random patient groupings could produce significant results. No significant results were found with random groupings.

DF	Pillai	Approx F	Num DF	Den DF	Pr (>*F*)	Dataset	Race	Replication
1	0.022262	1.289666	254	14,387	0.001391	Combined–all	Asian	4
1	0.074876	1.250691	254	3925	0.005367	2.1	Asian	2
1	0.891255	1.742418	254	54	0.007912	1.1	Asian	3
1	0.004084	1.201707	254	74,432	0.015351	Combined–all	Black	3
1	0.851154	1.530898	254	68	0.0191	2.5	Asian	2
1	0.616741	1.317772	254	208	0.019357	2.5	Hispanic	5
1	0.008821	1.19074	254	33,983	0.020347	MIMIC	Black	4
1	0.400541	1.22849	254	467	0.029149	1.1	Hispanic	4
1	0.056385	1.164972	254	4952	0.040539	2.1	White	1
1	0.006795	1.160335	254	43,078	0.04077	2–all	White	1
1	0.055885	1.154261	254	4953	0.050675	2.1	Black	5

Table 8 (*Continued*).

DF	Pillai	Approx F	Num DF	Den DF	Pr (>*F*)	Dataset	Race	Replication
1	0.02456	1.142726	254	11,528	0.060841	2.5	Black	4
1	0.024313	1.131142	254	11,530	0.07644	2.4	White	4
1	0.073933	1.132154	254	3602	0.080353	1–all	White	2
1	0.005677	1.123355	254	49,980	0.086573	Combined–no MIMIC	White	1
1	0.019397	1.120414	254	14,387	0.093037	Combined–all	Asian	5
1	0.200981	1.12299	254	1134	0.111955	2–all	Hispanic	4
1	0.183696	1.120741	254	1265	0.113492	1–all	Hispanic	2
1	0.023799	1.106649	254	11,530	0.119698	2.4	White	2
1	0.005578	1.103712	254	49980	0.123697	Combined–no MIMIC	White	2
1	0.007888	1.10357	254	35,256	0.124283	2–all	Black	4
1	0.079472	1.107378	254	3258	0.125223	1–all	Black	3
1	0.824549	1.258162	254	68	0.130637	2.5	Asian	1
1	0.396111	1.128516	254	437	0.136087	2.4	Asian	4
1	0.066368	1.098466	254	3925	0.14311	2.1	Asian	4
1	0.071818	1.097255	254	3602	0.146657	1–all	White	3
1	0.001955	1.092265	254	14,1616	0.149697	MIMIC	White	2
1	0.120897	1.093688	254	2020	0.16199	2.2	Black	4
1	0.05284	1.087643	254	4952	0.167952	2.1	White	4
1	0.036472	1.085065	254	7281	0.172629	Combined–no MIMIC	Asian	5
1	0.006322	1.079064	254	43,078	0.185396	2–all	White	2
1	0.148969	1.085417	254	1575	0.187187	1.2	Black	3
1	0.446117	1.106683	254	349	0.190352	1.2	Asian	5
1	0.021001	1.077382	254	12,757	0.192078	2.5	White	1
1	0.036198	1.0766	254	7281	0.196288	Combined–no MIMIC	Asian	1
1	0.00628	1.071754	254	43,078	0.207029	2–all	White	4
1	0.019091	1.072136	254	13,992	0.207446	Combined–all	Hispanic	4
1	0.001918	1.071236	254	14,1616	0.208087	MIMIC	White	4
1	0.575387	1.109675	254	208	0.217655	2.5	Hispanic	4
1	0.632288	1.117011	254	165	0.221142	2.4	Hispanic	4
1	0.234567	1.074987	254	891	0.229071	2.3	Asian	2
1	0.087471	1.066482	254	2826	0.233725	Combined–no MIMIC	Hispanic	4
1	0.041185	1.064548	254	6295	0.234089	2–all	Asian	5
1	0.007592	1.061916	254	35,256	0.238722	2–all	Black	1
1	0.001402	1.060815	254	19,1851	0.241698	Combined–all	White	1
1	0.018902	1.061282	254	13,992	0.242179	Combined–all	Hispanic	2

Table 8 (*Continued*).

DF	Pillai	Approx F	Num DF	Den DF	Pr (>*F*)	Dataset	Race	Replication
1	0.020647	1.058858	254	12,757	0.250588	2.5	White	3
1	0.06421	1.0603	254	3925	0.250912	2.1	Asian	1
1	0.159095	1.063664	254	1428	0.252187	1.1	Black	2
1	0.076224	1.058386	254	3258	0.258811	1–all	Black	5
1	0.628234	1.097745	254	165	0.259007	2.4	Hispanic	5
1	0.018297	1.055701	254	14,387	0.261242	Combined–all	Asian	3
1	0.024307	1.055735	254	10,764	0.261806	2.3	White	1
1	0.29135	1.065064	254	658	0.267222	1–all	Asian	2
1	0.063166	1.055438	254	3976	0.267338	3	White	1
1	0.43876	1.074164	254	349	0.267606	1.2	Asian	1
1	0.007525	1.05246	254	35,256	0.271545	2–all	Black	5
1	0.069196	1.054216	254	3602	0.272294	1–all	White	1
1	0.03506	1.049541	254	7337	0.285118	2.4	Black	5
1	0.260937	1.057803	254	761	0.285135	3	Asian	4
1	0.007495	1.048225	254	35,256	0.287011	2–all	Black	3
1	0.132859	1.051996	254	1744	0.287579	1.2	White	5
1	0.005289	1.04631	254	49,980	0.293929	Combined–no MIMIC	White	4
1	0.436124	1.062717	254	349	0.298751	1.2	Asian	4
1	0.142443	1.048277	254	1603	0.301735	1.1	White	5
1	0.114601	1.047193	254	2055	0.302411	2.2	White	4
1	0.001379	1.042854	254	191,851	0.306674	Combined–all	White	2
1	0.805751	1.110495	254	68	0.308879	2.5	Asian	3
1	0.116	1.043573	254	2020	0.315713	2.2	Black	3
1	0.115935	1.04291	254	2020	0.31814	2.2	Black	5
1	0.617944	1.069786	254	168	0.319754	2.3	Hispanic	5
1	0.037043	1.037585	254	6851	0.33051	MIMIC	Asian	4
1	0.074853	1.037801	254	3258	0.333227	1–all	Black	2
1	0.007407	1.035811	254	35,256	0.33485	2–all	Black	2
1	0.07474	1.036114	254	3258	0.339735	1–all	Black	1
1	0.838169	1.101105	254	54	0.34353	1.1	Asian	5
1	0.006476	1.031427	254	40,194	0.352484	Combined–no MIMIC	Black	1
1	0.327024	1.038836	254	543	0.356443	1.2	Hispanic	2
1	0.114861	1.031994	254	2020	0.35938	2.2	Black	2
1	0.020086	1.029495	254	12,757	0.361475	2.5	White	4
1	0.023716	1.02947	254	10,764	0.361864	2.3	White	5

Table 8 (*Continued*).

DF	Pillai	Approx F	Num DF	Den DF	Pr (>*F*)	Dataset	Race	Replication
1	0.326354	1.035673	254	543	0.367113	1.2	Hispanic	3
1	0.171537	1.031196	254	1265	0.367516	1–all	Hispanic	5
1	0.256316	1.032612	254	761	0.37037	3	Asian	2
1	0.112747	1.028099	254	2055	0.374471	2.2	White	2
1	0.001356	1.025638	254	191,851	0.376092	Combined–all	White	3
1	0.039743	1.025739	254	6295	0.378417	2–all	Asian	2
1	0.61695	1.046271	254	165	0.37868	2.4	Hispanic	3
1	0.359772	1.033182	254	467	0.379352	1.1	Hispanic	5
1	0.001354	1.023956	254	19,1851	0.383184	Combined–all	White	4
1	0.022048	1.023236	254	11,528	0.387609	2.5	Black	5
1	0.104503	1.024556	254	2230	0.387971	3	Black	5
1	0.049871	1.023313	254	4952	0.389171	2.1	White	3
1	0.15419	1.024892	254	1428	0.390373	1.1	Black	3
1	0.112236	1.022852	254	2055	0.39539	2.2	White	5
1	0.798385	1.060143	254	68	0.39629	2.5	Asian	5
1	0.141569	1.022609	254	1575	0.398446	1.2	Black	4
1	0.358199	1.026139	254	467	0.403222	1.1	Hispanic	1
1	0.083975	1.019957	254	2826	0.405313	Combined–no MIMIC	Hispanic	3
1	0.00597	1.018554	254	43,078	0.406533	2–all	White	3
1	0.023154	1.018178	254	10,911	0.40913	MIMIC	Hispanic	1
1	0.13879	1.017066	254	1603	0.420505	1.1	White	3
1	0.103723	1.01602	254	2230	0.422759	3	Black	2
1	0.224483	1.015395	254	891	0.43224	2.3	Asian	4
1	0.25308	1.015159	254	761	0.434874	3	Asian	3
1	0.066593	1.011729	254	3602	0.4389	1–all	White	5
1	0.022996	1.011079	254	10,911	0.439793	MIMIC	Hispanic	3
1	0.281118	1.013033	254	658	0.444557	1–all	Asian	4
1	0.128305	1.010629	254	1744	0.446289	1.2	White	4
1	0.28101	1.012489	254	658	0.446591	1–all	Asian	5
1	0.049218	1.009219	254	4952	0.448938	2.1	White	2
1	0.029607	1.00877	254	8398	0.450133	2.3	Black	4
1	0.017475	1.007436	254	14,387	0.455552	Combined–all	Asian	1
1	0.021708	1.007286	254	11,530	0.456356	2.4	White	3
1	0.049119	1.007307	254	4953	0.457212	2.1	Black	1
1	0.829176	1.031948	254	54	0.459522	1.1	Asian	1

Table 8 (*Continued*).

DF	Pillai	Approx F	Num DF	Den DF	Pr (>*F*)	Dataset	Race	Replication
1	0.01794	1.006295	254	13,992	0.460599	Combined–all	Hispanic	5
1	0.137505	1.006149	254	1603	0.4652	1.1	White	4
1	0.003415	1.00415	254	74,432	0.469692	Combined–all	Black	2
1	0.048921	1.003029	254	4953	0.475828	2.1	Black	4
1	0.060097	1.000878	254	3976	0.485468	3	White	2
1	0.021503	0.997401	254	11,528	0.500148	2.5	Black	3
1	0.366983	0.99742	254	437	0.505296	2.4	Asian	3
1	0.082129	0.995524	254	2826	0.509012	Combined–no MIMIC	Hispanic	5
1	0.126498	0.994333	254	1744	0.514421	1.2	White	1
1	0.182074	0.993831	254	1134	0.516836	2–all	Hispanic	1
1	0.033492	0.993318	254	7281	0.518386	Combined–no MIMIC	Asian	4
1	0.03848	0.991842	254	6295	0.52493	2–all	Asian	1
1	0.033166	0.990893	254	7337	0.529143	2.4	Black	3
1	0.006221	0.990549	254	40,194	0.530759	Combined–no MIMIC	Black	2
1	0.135478	0.988991	254	1603	0.53688	1.1	White	1
1	0.081608	0.988657	254	2826	0.538681	Combined–no MIMIC	Hispanic	2
1	0.149363	0.987171	254	1428	0.544353	1.1	Black	5
1	0.546099	0.985237	254	208	0.546476	2.5	Hispanic	3
1	0.007299	0.983735	254	33,983	0.561323	MIMIC	Black	3
1	0.035179	0.983467	254	6851	0.561979	MIMIC	Asian	3
1	0.134371	0.979659	254	1603	0.575903	1.1	White	2
1	0.022308	0.980148	254	10,911	0.576936	MIMIC	Hispanic	4
1	0.028779	0.979732	254	8398	0.578593	2.3	Black	5
1	0.033009	0.978506	254	7281	0.583853	Combined–no MIMIC	Asian	3
1	0.006145	0.978499	254	40,194	0.584704	Combined–no MIMIC	Black	4
1	0.273615	0.975811	254	658	0.586438	1–all	Asian	1
1	0.346068	0.972999	254	467	0.593625	1.1	Hispanic	3
1	0.147777	0.974876	254	1428	0.595186	1.1	Black	4
1	0.147744	0.974621	254	1428	0.596234	1.1	Black	1
1	0.070493	0.972772	254	3258	0.607227	1–all	Black	4
1	0.001743	0.973226	254	141,616	0.608197	MIMIC	White	1
1	0.592479	0.961611	254	168	0.6133	2.3	Hispanic	1
1	0.037724	0.971591	254	6295	0.613748	2–all	Asian	3
1	0.028511	0.97031	254	8398	0.619707	2.3	Black	1
1	0.123676	0.969025	254	1744	0.620293	1.2	White	2

Table 8 (*Continued*).

DF	Pillai	Approx F	Num DF	Den DF	Pr (>*F*)	Dataset	Race	Replication
1	0.596139	0.958883	254	165	0.620582	2.4	Hispanic	2
1	0.018943	0.969793	254	12,757	0.622406	2.5	White	5
1	0.007194	0.969516	254	33,983	0.624185	MIMIC	Black	1
1	0.106851	0.967908	254	2055	0.625792	2.2	White	3
1	0.058227	0.967816	254	3976	0.62883	3	White	5
1	0.07993	0.966551	254	2826	0.632923	Combined–no MIMIC	Hispanic	1
1	0.35819	0.960184	254	437	0.638082	2.4	Asian	2
1	0.106484	0.96418	254	2055	0.641089	2.2	White	1
1	0.134362	0.962473	254	1575	0.646159	1.2	Black	2
1	0.814145	0.931299	254	54	0.649381	1.1	Asian	2
1	0.007131	0.960984	254	33,983	0.660773	MIMIC	Black	2
1	0.308762	0.954911	254	543	0.660807	1.2	Hispanic	1
1	0.107389	0.956791	254	2020	0.67073	2.2	Black	1
1	0.063234	0.957262	254	3602	0.672212	1–all	White	4
1	0.098241	0.956479	254	2230	0.672679	3	Black	4
1	0.046778	0.956935	254	4953	0.674775	2.1	Black	2
1	0.001712	0.956352	254	14,1616	0.680532	MIMIC	White	3
1	0.034216	0.955573	254	6851	0.681276	MIMIC	Asian	1
1	0.133266	0.953414	254	1575	0.681823	1.2	Black	1
1	0.003245	0.953917	254	74,432	0.690442	Combined–all	Black	5
1	0.004815	0.952	254	49,980	0.698123	Combined–no MIMIC	White	5
1	0.021952	0.951142	254	10,764	0.700161	2.3	White	2
1	0.016971	0.95103	254	13,992	0.701026	Combined–all	Hispanic	3
1	0.211531	0.941097	254	891	0.719426	2.3	Asian	1
1	0.033879	0.94585	254	6851	0.71994	MIMIC	Asian	5
1	0.582427	0.922538	254	168	0.720275	2.3	Hispanic	2
1	0.530633	0.925789	254	208	0.721565	2.5	Hispanic	1
1	0.582274	0.921957	254	168	0.721788	2.3	Hispanic	4
1	0.337284	0.935731	254	467	0.722126	1.1	Hispanic	2
1	0.158833	0.940408	254	1265	0.727645	1–all	Hispanic	1
1	0.005931	0.944071	254	40,194	0.729414	Combined–no MIMIC	Black	5
1	0.005527	0.942659	254	43,078	0.734884	2–all	White	5
1	0.020314	0.941081	254	11,528	0.739451	2.5	Black	1
1	0.004761	0.941301	254	49,980	0.740135	Combined–no MIMIC	White	3
1	0.303141	0.929965	254	543	0.74503	1.2	Hispanic	4

Table 8 (*Continued*).

DF	Pillai	Approx F	Num DF	Den DF	Pr (>*F*)	Dataset	Race	Replication
1	0.04592	0.93834	254	4952	0.747068	2.1	White	5
1	0.021604	0.93575	254	10,764	0.759043	2.3	White	4
1	0.027426	0.932354	254	8398	0.770577	2.3	Black	2
1	0.001669	0.932308	254	141,616	0.773504	MIMIC	White	5
1	0.208357	0.92326	254	891	0.778986	2.3	Asian	5
1	0.763306	0.863347	254	68	0.790281	2.5	Asian	4
1	0.021271	0.921007	254	10,764	0.809475	2.3	White	3
1	0.016425	0.919922	254	13,992	0.813491	Combined–all	Hispanic	1
1	0.030768	0.91698	254	7337	0.820984	2.4	Black	2
1	0.094213	0.913182	254	2230	0.824584	3	Black	3
1	0.343278	0.899316	254	437	0.825326	2.4	Asian	5
1	0.154411	0.909441	254	1265	0.827402	1–all	Hispanic	3
1	0.342869	0.897686	254	437	0.829436	2.4	Asian	1
1	0.093917	0.910007	254	2230	0.833788	3	Black	1
1	0.127577	0.906756	254	1575	0.838463	1.2	Black	5
1	0.393006	0.889625	254	349	0.839499	1.2	Asian	2
1	0.055552	0.908924	254	3925	0.841683	2.1	Asian	5
1	0.055497	0.907974	254	3925	0.844352	2.1	Asian	3
1	0.005717	0.909855	254	40,194	0.845061	Combined–no MIMIC	Black	3
1	0.035362	0.908533	254	6295	0.845256	2–All	Asian	4
1	0.513706	0.865059	254	208	0.864654	2.5	Hispanic	2
1	0.020357	0.892641	254	10,911	0.887605	MIMIC	Hispanic	5
1	0.290399	0.874875	254	543	0.888625	1.2	Hispanic	5
1	0.226706	0.878354	254	761	0.891409	3	Asian	5
1	0.006619	0.8915	254	33,983	0.891784	MIMIC	Black	5
1	0.226528	0.877461	254	761	0.893204	3	Asian	1
1	0.019227	0.889754	254	11,528	0.894229	2.5	Black	2
1	0.562719	0.835951	254	165	0.900238	2.4	Hispanic	1
1	0.017347	0.886613	254	12,757	0.90121	2.5	White	2
1	0.113688	0.880723	254	1744	0.901854	1.2	White	3
1	0.015379	0.884674	254	14,387	0.905453	Combined–all	Asian	2
1	0.043234	0.881161	254	4953	0.909346	2.1	Black	3
1	0.002999	0.881537	254	74,432	0.91305	Combined–all	Black	1
1	0.382689	0.851794	254	349	0.913326	1.2	Asian	3
1	0.162958	0.869173	254	1134	0.917074	2–all	Hispanic	3

Table 8 (*Continued*).

DF	Pillai	Approx F	Num DF	Den DF	Pr (>*F*)	Dataset	Race	Replication
1	0.025854	0.877482	254	8398	0.918232	2.3	Black	3
1	0.018901	0.874504	254	11,530	0.924287	2.4	White	5
1	0.001156	0.874261	254	191,851	0.92653	Combined–all	White	5
1	0.197021	0.860703	254	891	0.926595	2.3	Asian	3
1	0.248164	0.855084	254	658	0.928479	1–all	Asian	3
1	0.029266	0.870853	254	7337	0.929392	2.4	Black	4
1	0.029404	0.868397	254	7281	0.93333	Combined–no MIMIC	Asian	2
1	0.002931	0.86154	254	74,432	0.945917	Combined–all	Black	4
1	0.772906	0.723571	254	54	0.947649	1.1	Asian	4
1	0.018538	0.857419	254	11,530	0.949976	2.4	White	1
1	0.54218	0.783292	254	168	0.960506	2.3	Hispanic	3
1	0.028475	0.846631	254	7337	0.961584	2.4	Black	1
1	0.019209	0.841298	254	10,911	0.967449	MIMIC	Hispanic	2
1	0.029904	0.83144	254	6851	0.974848	MIMIC	Asian	2
1	0.050287	0.828852	254	3976	0.97548	3	White	4
1	0.14105	0.817827	254	1265	0.977234	1–all	Hispanic	4
1	0.04992	0.822484	254	3976	0.979715	3	White	3
1	0.150124	0.78863	254	1134	0.99026	2–all	Hispanic	5
1	0.146995	0.769364	254	1134	0.994984	2–all	Hispanic	2

Disclosures

J.W.G. and S.P. are funded by the US National Science Foundation (Grant No. 1928481) from the Division of Electrical, Communication and Cyber Systems. All other authors have no relevant financial interests in the manuscript and no other potential conflicts of interest to disclose.

Code, Data, and Materials Availability

Code is available at https://github.com/iupui-soic/cxr-pixel-bias/. Institutional data remains internal. MIMIC-CXR can be accessed in Ref. 11.

References

1. J. W. Gichoya et al., "AI recognition of patient race in medical imaging: a modelling study," *Lancet Digit. Health* **4**(6), e406–e414 (2022).
2. J. Adleberg et al., "Predicting patient demographics from chest radiographs with deep learning," *J. Am. Coll. Radiol.* **19**(10), 1151–1161 (2022).
3. L. Seyyed-Kalantari et al., "Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations," *Nat. Med.* **27**(12), 2176–2182 (2021).
4. B. Glocker et al., Eds., "Algorithmic encoding of protected characteristics and its implications on performance disparities," (2021).
5. A. Tariq et al., "Current clinical applications of artificial intelligence in radiology and their best supporting evidence," *J. Am. Coll. Radiol.* **17**(11), 1371–1381 (2020).
6. M. A. Ricci Lara, R. Echeveste, and E. Ferrante, "Addressing fairness in artificial intelligence for medical imaging," *Nat. Commun.* **13**(1), 4581 (2022).

7. C. E. Kahn, Jr., "Hitting the mark: reducing bias in AI systems," *Radiol. Artif. Intell.* **4**(5), e220171 (2022).
8. K. Zhang et al., "Mitigating bias in radiology machine learning: 2. Model development," *Radiol. Artif. Intell.* **4**(5), e220010 (2022).
9. E. Pierson et al., "An algorithmic approach to reducing unexplained pain disparities in underserved populations," *Nat. Med.* **27**(1), 136–140 (2021).
10. L. Seyyed-Kalantari et al., "CheXclusion: fairness gaps in deep chest x-ray classifiers," *Biocomputing* **26**, 232–243 (2021).
11. A. E. W. Johnson et al., "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Sci. Data* **6**(1), 317 (2019).
12. Carestream, "Carestream DRX-revolution mobile imaging system," https://www.carestream.com/en/us/medical/dr-systems/mobile-x-ray/carestream-drx-revolution (accessed 7 January 2022).
13. R. A. Armstrong, "When to use the Bonferroni correction," *Ophthalmic Physiolog. Opt.* **34**(5), 502–508 (2014).
14. K. P. Weinfurt, "Multivariate analysis of variance," in *Reading and Understanding Multivariate Statistics*, L. G. Grimm and P. R. Yarnold, Eds., pp. 245–276, American Psychological Association, Washington, DC (1995).
15. M. Bostock, "D3 data-driven documents 2021," https://d3js.org (accessed 7 January 2022).
16. Keras, "KerasTuner 2022," https://keras.io/keras_tuner/ (accessed 7 January 2022).
17. TensorFlow, "TensorFlow Decision Forests," https://www.tensorflow.org/decision_forests (accessed 7 January 2022).
18. Association NEM, "Table C.8-27. X-ray acquisition module attributes," (2016) https://dicom.nema.org/medical/Dicom/2016e/output/chtml/part03/sect_C.8.7.2.html (accessed 7 January 2022).
19. J. L. Burns et al., "Data visualization: pixel color averages by race in chest x-ray," (2022) https://ai-vengers.web.app (accessed 28 July 2023).

John Lee Burns is a doctoral student at Indiana University's Health and Bioinformatics Program, minoring in data science. He received his MS degree in health informatics, his BS degree in computer science, and is Project Management Professional (PMP) certified. He works as an informatics director at the IU School of Medicine in the Department of Radiology and Imaging Sciences. His team develops innovative web applications supporting research, education, and clinical projects as well as the physicians' PACS/RIS environment, among other informatics needs. His research interests include real-time clinical decision support, natural language processing, and bias in medical imaging.

Zachary Zaiman, BS, is a recent graduate from Emory University in the Department of Computer Science where his research focused on data science in healthcare. He now is a software engineer at Microsoft.

Gaoxiang Luo, a computer science undergraduate at the University of Minnesota – Twin Cities, conducts research in machine learning and computer vision, focusing primarily on their applications within the healthcare sector.

Le Peng is a PhD candidate in computer science and engineering at the University of Minnesota, under the guidance of Dr. Ju Sun. His research interests encompass a wide spectrum of machine learning, including computer vision, natural language processing, and AI for healthcare.

Christopher Tignanelli is the current dyad director of the University of Minnesota Center for Outcomes, Quality, Delivery and Evaluation (C-QODE). He is the current co-director of the Federated Computer Vision in Healthcare U.S. Collaborative and faculty in the UMN Institute for Health Informatics' Natural Language Processing research lab. He is an AHRQ-funded K12 Learning Health System Scholar.

Sunandan Chakraborty is an assistant professor at the Luddy School of Informatics, Computing, and Engineering. His research centers around data science for social good, where he develops computational models using extensive datasets to address a wide range of problems in health, education, social sciences, and environmental sciences. He utilizes various data sources, including news, social media, and time-series data, to convert raw information into usable knowledge for practical applications.

Judy Wawira Gichoya, MD, MS, is an assistant professor at Emory University in Interventional Radiology and Informatics. Her career focus is on validating machine learning models for health in real clinical settings, exploring explainability, fairness, and a specific focus on how algorithms

fail. She is heavily invested in training the next generation of data scientists through multiple high school programs, serving as the program director for the *Radiology:AI* trainee editorial board and the medical students machine learning elective.

Saptarshi Purkayastha is associate professor of Health Informatics and Data Science at Indiana University Purdue University Indianapolis. He is the program director for health informatics with research interests in combining human and machine learning. He participates in open-source development of EHR systems and mHealth apps. He has industrial R&D experience in logistics and manufacturing. He works in global health through consulting work with the World Health Organization in eHealth architecture and health systems evaluation.

Biographies of the other authors are not available.