Radiology: Artificial Intelligence

ORIGINAL RESEARCH

Improving Fairness of Automated Chest Radiograph Diagnosis by Contrastive Learning

Mingquan Lin, PhD • Tianhao Li, MSc • Zhaoyi Sun, MSc • Gregory Holste, BA • Ying Ding, PhD • Fei Wang, PhD • George Shih, MD • Yifan Peng, PhD

From the Departments of Population Health Sciences (M.L., Z.S., F.W., Y.P.) and Radiology (G.S.), Weill Cornell Medicine, 425 E 61st St, New York, NY 10065; Department of Surgery, University of Minnesota, Minneapolis, Minn (M.L.); and School of Information (T.L., Y.D.) and Department of Electrical and Computer Engineering (G.H.), The University of Texas at Austin, Austin, Tex. Received August 23, 2023; revision requested October 5; revision received July 21, 2024; accepted August 8. **Address correspondence to** Y.P. (email: *yip4002@med.cornell.edu*).

Supported by the National Library of Medicine under award number 4R00LM013001, the National Science Foundation Faculty Early Career Development (CAREER) award number 2145640, the Intramural Research Program of the National Institutes of Health (NIH), and the Amazon Research Award. The Medical Imaging and Data Resource Center (MIDRC) is funded by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) of the NIH under contract 75N920202D00021 and through The Advanced Research Projects Agency for Health (ARPA-H).

Conflicts of interest are listed at the end of this article.

See also the commentary by Johnson in this issue.

Radiology: Artificial Intelligence 2024; 6(5):e230342 • https://doi.org/10.1148/ryai.230342 • Content codes: AI CH

Purpose: To develop an artificial intelligence model that uses supervised contrastive learning (SCL) to minimize bias in chest radiograph diagnosis.

Materials and Methods: In this retrospective study, the proposed method was evaluated on two datasets: the Medical Imaging and Data Resource Center (MIDRC) dataset with 77 887 chest radiographs in 27 796 patients collected as of April 20, 2023, for COVID-19 diagnosis and the National Institutes of Health ChestX-ray14 dataset with 112 120 chest radiographs in 30 805 patients collected between 1992 and 2015. In the ChestX-ray14 dataset, thoracic abnormalities included atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, pneumothorax, consolidation, edema, emphysema, fibrosis, pleural thickening, and hernia. The proposed method used SCL with carefully selected positive and negative samples to generate fair image embeddings, which were fine-tuned for subsequent tasks to reduce bias in chest radiograph diagnosis. The method was evaluated using the marginal area under the receiver operating characteristic curve difference (ΔmAUC).

Results: The proposed model showed a significant decrease in bias across all subgroups compared with the baseline models, as evidenced by a paired *t* test ($P < .001$). The ΔmAUCs obtained by the proposed method were 0.01 (95% CI: 0.01, 0.01), 0.21 (95% CI: 0.21, 0.21), and 0.10 (95% CI: 0.10, 0.10) for sex, race, and age subgroups, respectively, on the MIDRC dataset and 0.01 (95% CI: 0.01, 0.01) and 0.05 (95% CI: 0.05, 0.05) for sex and age subgroups, respectively, on the ChestX-ray14 dataset.

Conclusion: Employing SCL can mitigate bias in chest radiograph diagnosis, addressing concerns of fairness and reliability in deep learning–based diagnostic methods.

Supplemental material is available for this article.

In recent years, artificial intelligence (AI) has been extensively used in image-based disease diagnosis (1–6). Although these models have attained or exceeded expert-level performance, the concern of fairness has emerged in various medical domains and populations (7). In the AI algorithm, fairness denotes the absence of bias or favoritism toward an individual or group based on their inherent or acquired characteristics (8). In medical domains, certain groups, such as those defined by race, sex, and age, have been identified as subject to unfair or biased decisions made by AI models (9–11).

A chest radiograph is a quick and convenient diagnostic tool that uses a low dose of ionizing radiation to produce images of the chest, including the lungs, heart, and chest wall. This imaging technique can shed light on the underlying cause of shortness of breath, persistent cough, chest pain, and injury. Additionally, chest radiographs help diagnose and monitor lung conditions such as pneumonia, emphysema, and cancer. Several studies have focused on automating disease diagnosis based on chest radiograph imaging to achieve accurate results (12–15). Although these efforts have achieved high accuracy in detecting abnormalities in chest radiographs, exploring AI model fairness and bias reduction has been relatively limited. Therefore, there is a need to develop methods to minimize bias in automated chest radiograph diagnosis.

Three primary methods exist to reduce bias in medical image classification. Preprocessing methods work to reduce bias through dataset resampling or augmentation (10,16). In-processing methods typically incorporate an adversarial component into the baseline model. This component predicts sensitive attributes derived from the input image and emphasizes the loss function selection (17,18). Last, postprocessing techniques can address unfairness by introducing perturbations to input images (19). These techniques prevent the model from relying on biased features and can be achieved without necessitating model retraining. Despite these strategies, they have two limitations. First, the changes might inadvertently affect overall performance. This type of degradation, where fairness is achieved by deteriorating the performance of one or more groups, is quite problematic (20,21). Furthermore, the testing and development of these methods are predominantly conducted on relatively small datasets. This limitation can impede their ability to be generalized or applied to more extensive, real-world scenarios.

Abbreviations

ADV = adversarial learning, AI = artificial intelligence, AUC = area under the receiver operating characteristic curve, BS = Brier score, DICOM = Digital Imaging and Communications in Medicine, ΔBS = difference between maximum and minimum BS values, ΔFPR = difference between maximum and minimum FPR values, ΔmAUC = difference between maximum and minimum marginal AUC values, ΔTPR = difference between maximum and minimum TPR values, ΔwAUC = difference between maximum and minimum wAUC values, FPR = false-positive rate, MIDRC = Medical Imaging and Data Resource Center, OR = odds ratio, SCL = supervised contrastive learning, TPR = true-positive rate, wAUC = within-group AUC

Summary

A proposed artificial intelligence model based on supervised contrastive learning effectively minimized bias in automated chest radiograph diagnosis.

Key Points

- A new supervised contrastive learning pretraining method was used to generate fair image embeddings from chest radiographs.
- The proposed method significantly reduced algorithmic bias for subgroups spanning race, sex, and age in automated chest radiograph diagnosis across two large chest radiograph datasets when compared with the baseline (paired *t* test, $P < .001$).
- Although the study focused on bias in COVID-19 and chest abnormality diagnosis, extensive experiments showed that the proposed method also reduced bias across subgroups for the detection of other thorax diseases when compared with the baseline (paired *t* test, $P < .001$).

Keywords

Thorax, Diagnosis, Supervised Learning, Convolutional Neural Network (CNN), Computer-aided Diagnosis (CAD)

The current study aims to investigate fairness issues in employing AI for chest radiograph diagnosis and to mitigate biases related to race, sex, and age. One potential reason for bias in AI models is the presence of nonneglectable subgroup information in image embeddings. For example, in the race subgroup, the image embeddings may contain race-related information that could lead to biased predictions by the models. Supervised contrastive learning (SCL) is a pretraining technique that uses label information to draw embeddings from the same class closer and push those from different classes further apart (22). Benefiting from well-trained embedding, it achieves superior performance on downstream classification tasks. Inspired by this method, we propose using SCL with carefully selected positive and negative samples to generate fair image embedding. Subsequently, the model is fine-tuned for downstream tasks. In our approach, we define images with the same label from different subgroups as positive samples and images with different labels from the same subgroup as negative samples. The evaluation focuses on the model's ability to reduce bias across subgroups.

Materials and Methods

The protocol for this retrospective study was approved by the institutional review board at each clinical center and Weill Cornell Medicine. Due to the publicly available nature of both datasets used in this study, the requirement for obtaining written informed consent from all patients was waived by the institutional review board.

Dataset Acquisition

Our proposed method was designed and assessed using two chest radiograph imaging datasets. The first dataset is a repository created for COVID-19 diagnosis, hosted at the University of Chicago as part of the Medical Imaging and Data Resource Center (MIDRC) (23). The MIDRC is a collaborative initiative funded by the National Institute of Biomedical Imaging and Bioengineering under contracts 75N92020C00008 and 75N92020C00021 and jointly led by the American College of Radiology, the Radiological Society of North America, and the American Association of Physicists in Medicine. The MIDRC accepts images using Digital Imaging and Communications in Medicine (DICOM) standard and clinical data in various formats. It is currently seeking COVID-19–related CT scans, radiographs, MRI studies, and US images along with similar control cases. This study focuses on radiographs. The race, sex, and age data in MIDRC are self-reported. According to the MIDRC Data Contributor Reference Document, the outcome in MIDRC was confirmed through COVID-19 test results (polymerase chain reaction or rapid antigen test) within a time frame of 0 to 14 days before the imaging study. As of September 2022, 126 295 imaging studies with demographic information were included in the MIDRC data. We collected computed radiography and digital radiography studies with age, sex, and race information. Figure 1 provides an overview of the data selection process. A final total of 77 887 chest radiographs from 60 802 imaging studies in 27 796 patients were included in this study.

The second dataset used in this study was the publicly accessible National Institutes of Health ChestX-ray14 dataset, which comprised 112 120 frontal chest radiographs in 30 805 patients (3). In the ChestX-ray14 dataset, race and sex information were self-reported, while age was recorded at the time of the patient's first admission. In the ChestX-ray14 dataset, a thoracic abnormality refers to any abnormal finding in the chest area. This finding encompasses various conditions, such as lung masses.

Bias Definition

To assess the model's fairness, we used the difference between the maximum and minimum values of the marginal area under the receiver operating characteristic curve (ΔmAUC); mAUC represents the mean marginal pairwise equal opportunity criterion (24). The mAUC (24) is defined as:

$$A_{G_i} := P\left(f(x) > f(x') \mid y > y', (x,y) \in G_i^+, (x',y') \in G^-\right).$$

G is the dataset used, G_i is the subgroup in the dataset, $f(x)$ is the output of the AI model with input image x, and y is the ground truth label for x, indicating whether the input image shows disease. P stands for the mAUC, which measures the AUC for a specific subgroup. It is calculated by determining the probability that the model ranks a randomly selected positive sample from the subgroup over a randomly selected negative example from the entire data. For binary classification, the mAUC requires that positive labels have an equal chance to be predicted positively across subgroups (24). By subtracting the minimum value of mAUC from the maximum, ΔmAUC can be obtained. A higher ΔmAUC signifies significant disparities at the levels

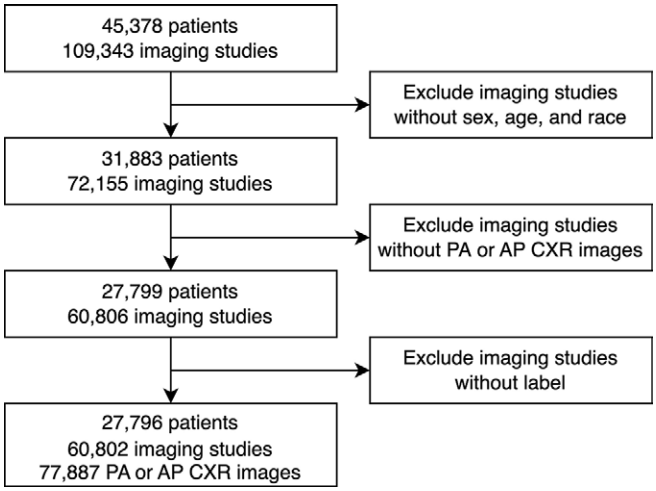

Figure 1: Flowchart of creation of Medical Imaging and Data Resource Center dataset. AP = anterior posterior, CXR = chest radiograph, PA = posterior anterior.

of individual subgroups and a lack of fairness in the model's predictions. For example, in the age subgroup, the mAUC for individuals younger than 75 years and their counterparts is 0.83 and 0.73, respectively, resulting in a ΔmAUC of 0.10. If the proposed method can reduce this value from 0.10 to a lower value, it successfully reduces bias.

Additionally, we use the difference between the maximum and minimum values of the subgroups in traditional evaluation metrics, specifically within-group AUC (wAUC), true-positive rate (TPR), false-positive rate (FPR), and Brier score (BS) to assess fairness. We refer to them as ΔwAUC, ΔTPR, ΔFPR, and ΔBS, respectively.

Overall Architecture

The overall architecture is presented in Figure 2. We first pretrained the model using contrastive learning, which learns the initial parameters for the model backbone. We then fine-tuned the model for the subsequent tasks. We used DenseNet-121 (25) as the backbone in this study.

Contrastive Learning Model

We used contrastive learning as a pretraining technique to minimize the bias among different subgroups, resulting in fair image embeddings. To implement contrastive learning, we replaced the final output layer of the prediction network with a single-layer perceptron, which served as the contrastive head. In contrastive learning, *anchor image* refers to an image that serves as a reference point within the contrastive loss function. For anchor images in a minibatch, we used images with the same label but originating from different subgroups as positive samples and images with different labels from the same subgroup as negative samples. In this scenario, a male individual with COVID-19 served as the anchor image, while the image of a female individual with COVID-19 that followed served as a positive sample. On the other hand, the image of a male individual without COVID-19 was considered a negative sample. In this context, positive sampling encouraged image embeddings from different subgroups to be similar to one another while still considering the label information. Image embeddings were the feature embeddings obtained by the convolutional part of the model. Conversely, negative sampling pushed image embeddings with distinct labels further apart, without emphasizing the group information. The contrastive loss can be expressed as follows (22):

$$L = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} log \frac{\exp\left(z_i \cdot z_p / \tau\right)}{\sum_{n \in N(i)} \exp\left(z_i \cdot z_n / \tau\right)}.$$

i is the anchor in the minibatch I, and the upper limit of the summation for i is the total number of anchor images in the minibatch. I is the set of all indices in the minibatch, $P(i)$ represents all the positive samples of i in the minibatch, $N(i)$ are all the negative samples of i in the minibatch, and z_i, z_p, and z_n are the image embeddings of i, p, and n, respectively. The loss function allows all positive pairs to contribute to the numerator, encouraging the encoder to provide closely aligned representations for all entries from the same class. The form of the loss function can distinguish between positive and negative samples.

Downstream Prediction

After we pretrained the model using contrastive learning, we replaced the contrastive head with the origin output layer, which is the prediction head in Figure 2. We then fine-tuned the model to generate the output result. We used binary cross-entropy loss in the downstream prediction.

Experimental Settings

For the MIDRC dataset, we followed the same image processing method as described in the study by Johnson et al (26) for the original chest radiographs. We started by converting all the posterior-anterior or anterior-posterior chest radiographs from DICOM to JPG format. Specifically, pixel values in the DICOM format were normalized to a range of [0, 255]. If necessary, all pixels were inverted to ensure that the air in the image appeared white and the area outside the patient's body appeared black. After that, we performed histogram equalization to enhance the image contrast. Finally, the processed image was saved in JPG format with a quality factor of 95.

All images were subsequently resized to 256 × 256 × 3 using PyTorch's default bilinear interpolation and center cropped to 224 × 224 × 3. Stochastic image augmentation was randomly applied to transform a chest radiograph into an augmented view. We sequentially applied two simple augmentation operations: *(a)* random rotation between 0° and 10° and *(b)* random flipping.

The proposed method is not exclusive to specific deep learning models, and DenseNet-121 (25) showed good performance in classification on ChestX-ray14 in the previous study (4). Therefore, we used a DenseNet-121 architecture pretrained on CheXpert (27) in this study. The last layer of the model is first substituted by a single-layer perceptron with an output dimension of 128 (backbone + contrastive head). We then fine-tuned the entire network for the subsequential tasks. Adam optimizer (28) with a learning rate of 0.0001 was used for contrastive learning. We set the temperature to 0.05 and trained the model for 10 epochs. After that, we replaced the output layer with the classification output layer (backbone + prediction head) and fine-tuned the model for 1 epoch. The experiments were conducted on an Intel Core i9–9960X 16-core processor and an NVIDIA Quadro RTX

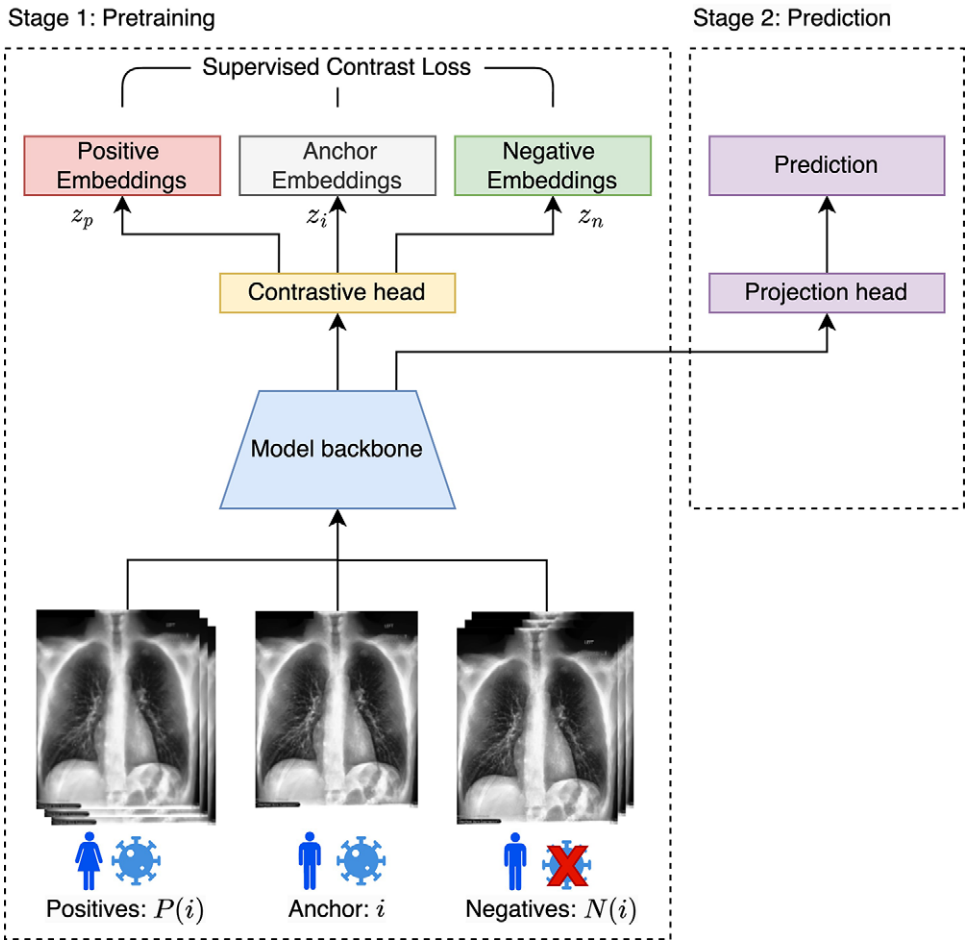

Figure 2: Graphic shows the overview of the proposed workflow using the contrastive learning model for fairness. A male individual with COVID-19 serves as the anchor image (middle), while the image of a female individual with COVID-19 serves as a positive sample (left), and the image of a male individual without COVID-19 is considered a negative sample (right).

6000 GPU. The models were implemented using PyTorch. The code is available at *https://github.com/bionlplab/CXRFairness*.

For the MIDRC dataset, we randomly split the entire dataset at the patient level. We designated one group (20% of the patients) as the held-out test set and used the remaining portion as the training and validation sets. For the ChestX-ray14 dataset, we used the official training, validation, and testing split.

We evaluated our methods on all subgroups across age, sex, and race. We compared our results with four baselines: empirical risk minimization (29), balanced empirical risk minimization (20), adversarial learning (30), and SCL (22). These four baselines were based on DenseNet-121 pretrained on CheXpert (27), which we refer to as DenseNet-121, Balance DenseNet-121, ADV, and SCL, respectively. DenseNet-121 is an original algorithm that does not consider the bias problem, and our proposed algorithm used DenseNet-121 as its backbone. Data resampling is a commonly used data preprocessing technique for reducing bias in subgroups, so we employed Balance DenseNet-121 as one of the baselines. In this study, we resampled the subgroups with fewer samples to ensure that the number of samples in all subgroups was the same. ADV is a widely used in-processing method derived from the domain adaptation field, which treats the sensitive attribute as a domain-specific label and attempts to use only domain-irrelevant features for the target task. SCL is a general contrastive learning approach without label definitions related to demographic information, which we use to demonstrate the effectiveness of the proposed method. To further evaluate the proposed method, we used both DenseNet-121 and the proposed model trained on the ChestX-ray14 dataset, testing them on the MIMIC-CXR test set.

Statistical Analysis

We used 200 bootstrap samples to obtain a distribution of the ΔmAUC and reported 95% CIs. For each bootstrap iteration, we sampled n images with replacements from the test set of n images. To compare the difference in ΔmAUC between the proposed model and baseline across all subgroups, we conducted a paired t test. Statistical analysis was conducted using SciPy 1.7.1 (Python Software Foundation), with statistical significance defined as a P value less than .05. We also attempted to perform bootstrapping at the patient level.

To analyze the bias within each dataset, we first employed logistic regression to analyze the association between demographic information (age, sex, and race) and the prevalence of COVID-19 on the MIDRC dataset. Age, sex, and race were used as predictors and compared with the reference group (eg, individuals younger than 75 years vs those 75 years and older, male vs female, Black vs White, and other race vs White). Other races included American Indian or Alaska Native, Asian, Native Hawaiian or other

Table 1: Patient Characteristics for Both Datasets

Characteristics per Dataset	Training Set	Test Set	Complete Set
MIDRC dataset	22 237	5559	27 796
Age (y)	59 (44–71)	59 (25–75)	59 (43–71)
Sex			
Male	11 257 (51)	2815 (51)	14 072 (51)
Female	10 980 (49)	2744 (49)	13 724 (49)
Race			
Black	7444 (33)	1912 (34)	9356 (34)
White	12 002 (54)	2998 (54)	15 000 (54)
Other	2791 (13)	649 (12)	3440 (12)
ChestX-ray14 dataset	28 008	2797	30 805
Age (y)	48 (34–59)	49 (34–59)	48 (34–59)
Sex			
Male	15 073 (54)	1557 (56)	16 630 (54)
Female	12 935 (46)	1240 (44)	14 175 (46)

Note.—Data are reported as medians, with IQRs in parentheses, for continuous variables and numbers of patients, with percentages in parentheses, for categorical variables. The racial category "other" includes American Indian or Alaska Native, Asian, Native Hawaiian or other Pacific Islander, and other race. MIDRC = Medical Imaging and Data Resource Center.

Table 2: Subgroup Information of Both Datasets at Image Level

Characteristics per Dataset	Training set		Test set	
	Positive	Total	Positive	Total
MIDRC dataset	31 434	62 178	7935	15 709
Age				
<75 years	26 868 (67)	52 427 (84)	6970 (88)	13 115 (83)
≥75 years	4566 (15)	9751 (16)	965 (12)	2594 (17)
Sex				
Male	17 991 (57)	35 081 (56)	4404 (56)	8799 (56)
Female	13 443 (43)	27 097 (44)	3531 (44)	6910 (44)
Race				
Black	16 836 (54)	24 104 (39)	4456 (56)	6135 (39)
White	11 616 (37)	30 667 (49)	2739 (35)	7790 (50)
Other	2982 (9)	7407 (12)	740 (9)	1784 (11)
ChestX-ray14 dataset	43 021	86 524	9671	25 596
Age				
<60 years	30 933 (72)	66 048 (76)	7579 (78)	19 634 (77)
≥60 years	12 088 (28)	20 476 (24)	2092 (22)	5962 (23)
Sex				
Male	24 409 (57)	48 458 (56)	5595 (58)	14 882 (58)
Female	18 612 (43)	38 066 (44)	4076 (42)	10 714 (42)

Note.—Data are presented as numbers of images, with percentages in parentheses. The racial category "other" includes American Indian or Alaska Native, Asian, Native Hawaiian or other Pacific Islander, and other race. MIDRC = Medical Imaging and Data Resource Center.

Pacific Islander, and other races. Odds ratios (ORs) larger than 1 indicated that the comparison groups had higher antecedent rates than the reference group. When comparing rates, 95% CIs were calculated. We also examined the association between demographic factors (age and sex) and thorax abnormality on the ChestX-ray14 dataset.

Results

Study Patients

Table 1 lists the patient characteristics for both datasets. For the MIDRC dataset, the training set included 22 237 patients (median age, 59 years [IQR: 44–71 years]; 11 257 [51%] male, 10 980 [49%] female), and the test set included 5559 patients (median age, 59 years [IQR: 25–75 years]; 2815 [51%] male, 2744 [49%] female). For the ChestX-ray14 dataset, the training set included 28 008 patients (median age, 48 years [IQR: 34–59 years]; 15 073 [54%] male, 12 935 [46%] female), and the test set included 2797 patients (median age, 49 years [IQR: 34–59 years]; 1557 [56%] male, 1240 [44%] female).

Table 2 presents the subgroup information of the datasets at the image level. Our study focused on training image-based classifiers for disease detection and evaluating the model's performance on subgroups based on sex, age, and race for the MIDRC dataset and sex and age for the ChestX-ray14 dataset. Due to the different age characteristics between these two datasets (Table 1), we set the age groups for each dataset differently.

Analysis of Bias within Each Dataset

As shown in Figure 3, age younger than 75 years (OR = 1.59 [95% CI: 1.53, 1.66]), male sex (OR = 1.04 [95% CI: 1.02, 1.08]), Black race (OR = 4.00 [95% CI: 3.87, 4.13]), and other race (OR = 1.14 [95% CI: 1.09, 1.20]) were associated with higher odds of COVID-19 diagnosis.

In the ChestX-ray14 dataset, age of 60 years or older (OR = 1.34 [95% CI: 1.30, 1.37]) and male sex (OR = 1.03 [95% CI: 1.00, 1.05]) were associated with higher odds of thorax abnormality (Fig 3).

Model Fairness Comparisons in MIDRC Dataset

Figure 4 shows that our proposed method produced significantly smaller ΔmAUC across all demographics compared with the baselines.

Table 3 presents a detailed performance comparison of various methods for COVID-19 diagnosis based on sex, race, and age. Individuals in subgroups with lower AUC values are at a higher risk of being misdiagnosed than their counterparts.

Specifically, compared with DenseNet-121, the ΔmAUC

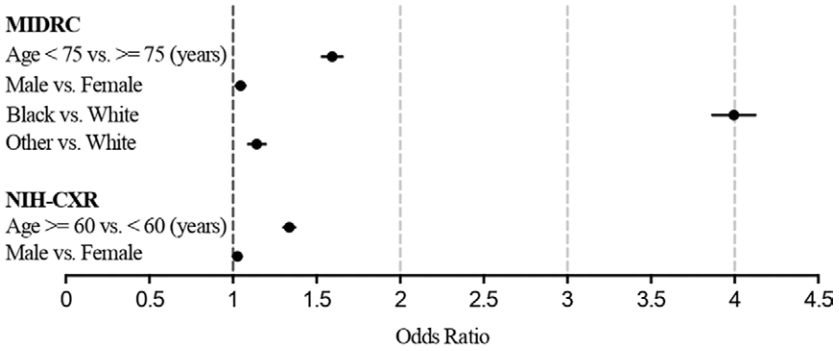

Figure 3: Forest plot of relative odds (95% CIs) of COVID-19 (Medical Imaging and Data Resource Center [MIDRC] dataset) and thorax abnormality (National Institutes of Health ChestX-ray14 [NIH-CXR] dataset) associated with age, sex, and race.

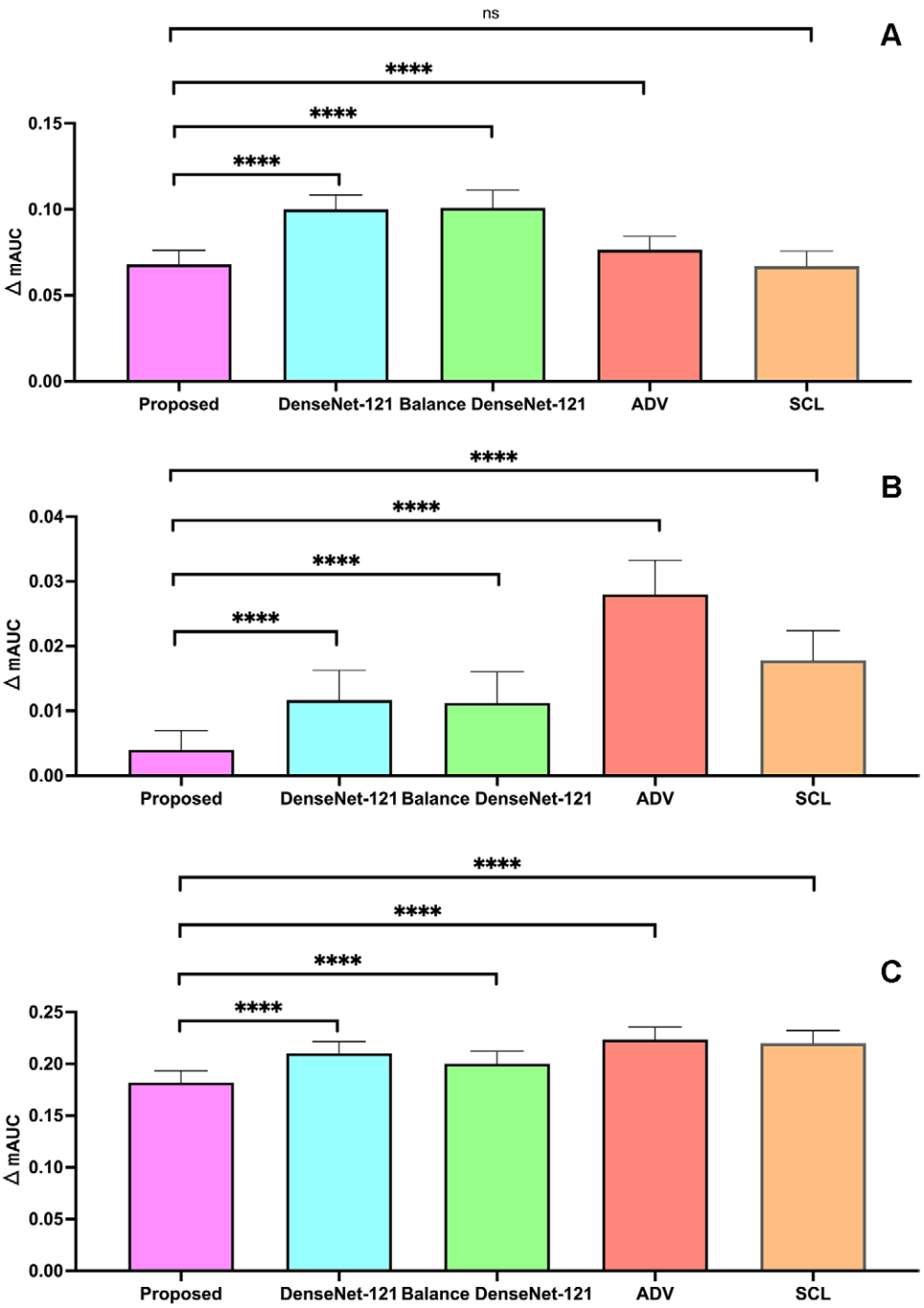

Figure 4: Bar graphs of ΔmAUC across subgroups of **(A)** sex, **(B)** age, and **(C)** race in COVID-19 detection on the Medical Imaging and Data Resource Center dataset. The results are averaged over 200 times in a bootstrap experiment. **** = $P \leq .001$, ADV = adversarial learning (30), Balance DenseNet-121 = DenseNet-121 with balanced empirical risk minimization (29), ΔmAUC = difference in marginal area under the receiver operating characteristic curve, ns = not significant, SCL = supervised contrastive learning (22).

Table 3: AUC, Marginal AUC, and Marginal AUC Difference of Baseline and Proposed Model for COVID-19 Diagnosis in MIDRC Dataset

Variable	DenseNet-121	Balance DenseNet-121	ADV	SCL	Proposed
Sex					
Overall AUC	0.82 (0.82, 0.82)	0.81 (0.81, 0.81)	0.81 (0.81, 0.81)	0.81 (0.81, 0.81)	0.81 (0.81, 0.81)
Male	0.82 (0.82, 0.82)	0.82 (0.82, 0.82)	0.82 (0.82, 0.82)	0.82 (0.82, 0.82)	0.81 (0.81, 0.81)
Female	0.81 (0.81, 0.81)	0.80 (0.80, 0.81)	0.79 (0.79, 0.79)	0.80 (0.80, 0.80)	0.81 (0.81, 0.81)
ΔmAUC	0.01 (0.01, 0.01)	0.01 (0.01, 0.01)	0.03 (0.03, 0.03)	0.02 (0.02, 0.02)	0.004 (0.003, 0.004)
Race					
Overall AUC	0.82 (0.82, 0.82)	0.81 (0.80, 0.81)	0.82 (0.82, 0.82)	0.81 (0.81, 0.81)	0.79 (0.79, 0.79)
White	0.76 (0.76, 0.76)	0.77 (0.77, 0.77)	0.76 (0.76, 0.76)	0.76 (0.75, 0.76)	0.75 (0.75, 0.75)
Black	0.88 (0.88, 0.88)	0.85 (0.85, 0.85)	0.88 (0.88, 0.88)	0.87 (0.87, 0.87)	0.84 (0.84, 0.84)
Other	0.67 (0.67, 0.67)	0.65 (0.65, 0.65)	0.66 (0.65, 0.66)	0.65 (0.65, 0.65)	0.66 (0.66, 0.66)
ΔmAUC	0.21 (0.21, 0.21)	0.20 (0.20, 0.20)	0.22 (0.22, 0.23)	0.22 (0.22, 0.22)	0.18 (0.18, 0.18)
Age					
Overall AUC	0.82 (0.82, 0.82)	0.80 (0.80, 0.80)	0.81 (0.81, 0.81)	0.81 (0.81, 0.81)	0.80 (0.80, 0.80)
<75 years	0.83 (0.83, 0.83)	0.81 (0.81, 0.81)	0.82 (0.82, 0.82)	0.82 (0.82, 0.82)	0.81 (0.81, 0.81)
≥75 years	0.73 (0.73, 0.73)	0.71 (0.71, 0.71)	0.75 (0.75, 0.75)	0.75 (0.75, 0.75)	0.74 (0.74, 0.74)
ΔmAUC	0.10 (0.10, 0.10)	0.10 (0.10, 0.10)	0.08 (0.08, 0.08)	0.07 (0.07, 0.07)	0.07 (0.07, 0.07)

Note.—Data in parentheses are 95% CIs. ADV = adversarial learning (30), AUC = area under the receiver operating characteristic curve, Balance DenseNet-121 = DenseNet-121 with balanced empirical risk minimization (29), ΔmAUC = difference between the maximum and minimum values of the marginal AUC, MIDRC = Medical Imaging and Data Resource Center, SCL = supervised contrastive learning (22).

obtained by the proposed method decreased from 0.01 (95% CI: 0.01, 0.01) to 0.004 (95% CI: 0.003, 0.004) for sex, with female individuals showing lower mAUC values than their male counterparts. For the race subgroup, the ΔmAUC obtained by the proposed method decreased from 0.21 (95% CI: 0.21, 0.21) to 0.18 (95% CI: 0.18, 0.18) compared with DenseNet-121. The other racial group showed lower mAUC values than the White and Black groups. Similarly, for the age subgroup, the ΔmAUC values decreased from 0.10 (95% CI: 0.10, 0.10) to 0.07 (95% CI: 0.07, 0.07) compared with DenseNet-121, with individuals younger than 75 years displaying lower mAUC values than their counterparts.

Tables S1–S5 present mAUC and ΔmAUC (bootstrap on patient level), wAUC and ΔwAUC, TPR and ΔTPR, FPR and ΔFPR, and BS and ΔBS of various methods for COVID-19 diagnosis based on sex, race, and age, respectively. The proposed method obtained comparable ΔmAUC in sex and lower ΔmAUC in age and race compared with DenseNet-121. The proposed method achieved comparable ΔwAUC in sex, higher ΔwAUC in race, and lower ΔwAUC in age compared with DenseNet-121. The proposed method demonstrated comparable TPR in sex and race subgroups to DenseNet-121, with lower ΔTPR in sex, race, and age subgroups. Moreover, the proposed method exhibited a lower FPR in the age subgroup and reduced ΔFPR in race and age subgroups compared with DenseNet-121. Additionally, the proposed method generated lower ΔBS in sex and age subgroups as well as lower BS in the age subgroup.

Model Fairness Comparisons in ChestX-ray14 Dataset

Figure 5 shows that our proposed method produced significantly smaller ΔmAUC across all demographics in diagnosing thorax abnormalities on the ChestX-ray14 dataset compared with the baselines.

Table 4 further presents a detailed analysis of the results. The proposed method achieved a lower ΔmAUC than the baselines for all demographic groups. In the sex subgroup analysis, the ΔmAUC obtained by the proposed method decreased from 0.01 (95% CI: 0.01, 0.01) to 0.005 (95% CI: 0.005, 0.01) compared with DenseNet-121. The proposed model performed similarly for male individuals as their counterparts in the sex subgroup analysis, while the baselines generated lower AUC for male individuals. In the age subgroup, the ΔmAUC obtained by the proposed method decreased from 0.05 (95% CI: 0.05, 0.05) to 0.04 (95% CI: 0.04, 0.04) compared with DenseNet-121. Individuals older than 60 years had lower AUCs than their counterparts.

Tables S6–S10 list mAUC and ΔmAUC (bootstrap on patient level), wAUC and ΔwAUC, TPR and ΔTPR, FPR and ΔFPR, and BS and ΔBS of various methods for diagnosing thorax abnormalities on the ChestX-ray14 dataset across sex and age, respectively. The proposed method obtained comparable ΔmAUC in the sex subgroup and lower ΔmAUC in the age subgroup compared with DenseNet-121. The proposed method achieved higher ΔwAUC in the sex subgroup and lower ΔwAUC in age than DenseNet. The proposed method achieved higher TPR in sex and age subgroups than DenseNet-121. Additionally, the proposed method exhibited lower ΔFPR in the sex subgroup than DenseNet-121. Furthermore, compared with DenseNet-121, the proposed method generated comparable BS and lower ΔBS in sex and age subgroups.

External Testing

The details of the results are presented in Table S11. For external testing, our proposed method achieved a higher AUC for both sex and age subgroups and a lower ΔmAUC for the sex subgroup.

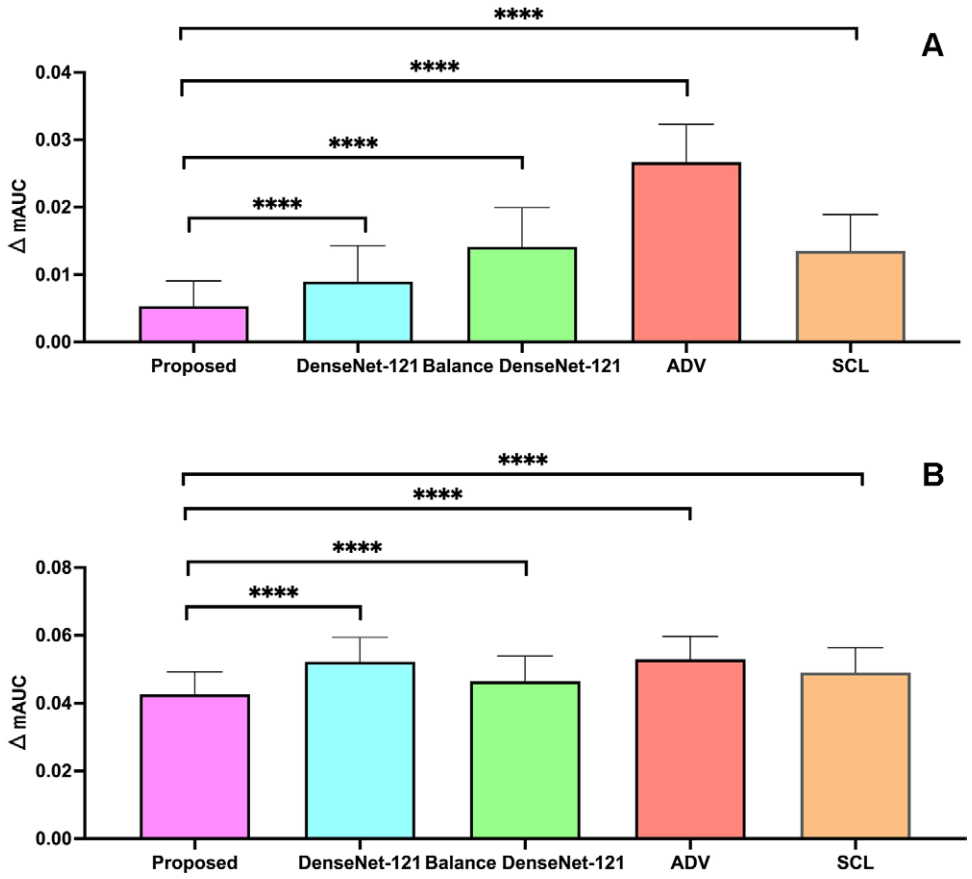

Figure 5: Bar graphs of ΔmAUC across subgroups of **(A)** sex and **(B)** age in thorax abnormality detection in the National Institutes of Health ChestX-ray14 dataset. The results are averaged over 200 times in a bootstrap experiment. **** = $P \leq .001$. ADV = adversarial learning (30), Balance DenseNet-121 = DenseNet-121 with balanced empirical risk minimization (29), ΔmAUC = difference in marginal area under the receiver operating characteristic curve, ns = not significant, SCL = supervised contrastive learning (22).

Table 4: AUC, Marginal AUC, and Marginal AUC Difference of Baseline and Proposed Model for Thorax Abnormalities Diagnosis in ChestX-ray14 Dataset

Variable	DenseNet-121	Balance DenseNet-121	ADV	SCL	Proposed
Sex					
Overall AUC	0.74 (0.73, 0.74)	0.73 (0.73, 0.73)	0.71 (0.71, 0.71)	0.73 (0.73, 0.73)	0.73 (0.73, 0.73)
Male	0.73 (0.73, 0.73)	0.73 (0.73, 0.73)	0.70 (0.70, 0.70)	0.72 (0.72, 0.73)	0.74 (0.73, 0.73)
Female	0.74 (0.74, 0.74)	0.74 (0.74, 0.74)	0.73 (0.73, 0.73)	0.74 (0.74, 0.74)	0.73 (0.73, 0.73)
ΔmAUC	0.01 (0.01, 0.01)	0.01 (0.01, 0.02)	0.03 (0.03, 0.03)	0.01 (0.01, 0.01)	0.005 (0.005, 0.005)
Age					
Overall AUC	0.74 (0.73, 0.74)	0.72 (0.72, 0.73)	0.71 (0.71, 0.71)	0.73 (0.73, 0.73)	0.73 (0.73, 0.73)
<60 years	0.75 (0.75, 0.75)	0.73 (0.73, 0.74)	0.72 (0.72, 0.72)	0.74 (0.74, 0.74)	0.74 (0.74, 0.74)
≥60 years	0.69 (0.69, 0.70)	0.69 (0.69, 0.69)	0.67 (0.66, 0.67)	0.69 (0.69, 0.69)	0.69 (0.69, 0.70)
ΔmAUC	0.05 (0.05, 0.05)	0.05 (0.05, 0.05)	0.05 (0.05, 0.05)	0.05 (0.05, 0.05)	0.04 (0.04, 0.04)

Note.—Data in parentheses are 95% CIs. ADV = adversarial learning (30), AUC = area under the receiver operating characteristic curve, Balance DenseNet-121 = DenseNet-121 with balanced empirical risk minimization (29), ΔmAUC = difference between the maximum and minimum values of the marginal AUC, SCL = supervised contrastive learning (22).

Relative Change

Table S12 lists the relative changes in AUC and mAUC, which are within 2%, while the relative change in ΔmAUC ranges from 13.5% to 68.10%.

Model Fairness Comparisons in Intersectional Groups

Table S13 provides a summary of the characteristics of the intersectional groups within the MIDRC dataset. It indicates that certain intersectional groups (eg, ≥75 years, other race, male

sex) contain a small number of positive cases. This small number not only poses challenges to model training but also affects the statistical significance of the results.

Table S14 presents the ΔmAUC between baseline and proposed models on the MIDRC across various intersectional subgroups. The proposed method demonstrates comparable or improved performance in terms of ΔmAUC across the different intersectional subgroups.

Discussion

In this study, we proposed a method leveraging SCL to reduce bias in AI models for chest radiograph image diagnosis across different groups. The proposed model was evaluated using two large-scale chest radiograph datasets. We observed systematic model biases in subgroups across all settings. This observation highlights the importance of addressing biases in AI models to ensure fair and accurate diagnoses across all demographic subgroups. Key observations are discussed below.

First, our proposed method effectively improves the fairness of chest radiograph diagnoses by using SCL to obtain fair image embeddings that retain label information. In contrast to state-of-the-art models, we modified the definitions of positive and negative samples to enable the network to capture more label and less group information. Our proposed method generated smaller ΔmAUC for both datasets across all demographics compared with the baselines. Additionally, the proposed method consistently maintains overall performance. We conducted quantitative analysis using the metric of relative change. The result showed that the relative change in AUC and mAUC remains consistently within 2%, while the relative change in ΔmAUC ranges from 13.5% to 68.10%. The results suggest that our proposed method can reduce bias (ΔmAUC) without significantly compromising AUC and mAUC.

Second, this study highlights the impact of data imbalance on the bias of AI models. The overrepresentation of prevalent patients in certain subgroups can lead to biased models, as evidenced by our findings. For example, in the MIDRC dataset, the prevalence of COVID-19 is significantly higher among Black individuals compared with their White counterparts (70.02% vs 37.33%). This overrepresentation can result in biased models trained on this dataset. Additionally, the sample size can still introduce bias even when subgroups have similar prevalence. For instance, in the MIDRC dataset, the number of White, Black, and other race individuals are 38 457, 30 239, and 9191, respectively. Although the COVID-19 prevalence is almost the same for the other race and White individuals (40.50% vs 37.33%), the former group, which had the smallest sample size among the racial subgroups, obtained the lowest AUC value. Similar phenomena were observed in the age subpopulations for thorax disease detection in the ChestX-ray14 dataset and COVID-19 detection in the MIDRC dataset.

Third, data resampling is a commonly used data preprocessing technique for reducing bias in subgroups, but our findings suggest that it may not always be effective. Specifically, our results show that the Balance DenseNet-121 model could not reduce bias for sex and age on the MIDRC dataset or bias for sex on the ChestX-ray14 dataset compared with the DenseNet-121 model. In this study, we employed only one resampling method to ensure an equal sample size across subgroups. However, future research could explore additional resampling methods to determine their effectiveness.

Fourth, ADV is widely used as an in-processing method to improve group fairness, but our results suggest that it may not always be effective. Specifically, our results demonstrate that the adversarial model could reduce bias only related to age in the MIDRC dataset when compared with the DenseNet-121 model.

Finally, SCL is a general contrastive learning approach without label definitions related to demographic information. The experiments conducted with SCL can be regarded as an ablation study to demonstrate that our proposed method considers demographic information to form positive and negative samples for learning image feature embeddings to improve group fairness. The results show that our proposed method effectively improved group fairness.

Our study had some limitations. Although this study assessed the fairness of binarized models, it did not examine the calibration of predicted probabilities. As a result, there was a possibility of overconfidence or underconfidence in certain cases. To address this limitation, future research should investigate the relationship between calibration and bias in disease detection and develop effective methods to reduce calibration bias. Moreover, expanding the proposed method to include continuous attributes and multiclass settings would increase its applicability.

Additionally, this study aimed to enhance the fairness of automated chest radiograph diagnosis through contrastive learning. We used two extensive chest radiograph datasets to showcase the effectiveness of the proposed method, both of which focused on thoracic diseases. In the future, we plan to extend the application of our method to other diseases and imaging modalities and test it on more models with other base architectures, not like this time when only using DenseNet-121.

Furthermore, given that MIDRC is a multi-institutional collaborative initiative and no exclusion criteria are specified in the dataset descriptions, we have taken measures to mitigate selection bias. However, it is important to acknowledge that MIDRC may not fully represent all patient populations. Finally, the data do not provide the comorbidity history of the patients.

Another limitation of our model was that it was designed to manage disparities linked to specific sensitive attributes individually, rather than addressing multiple variables like sex, race, and age simultaneously. Navigating multiple sensitive attributes necessitates a more complex model architecture. Additionally, it requires a more diverse dataset. However, in the datasets used in this study, certain intersectional groups (eg, ≥75 years old, other sexes) consisted of only a small number of positive cases, which not only created difficulties in model training but also influenced the statistical significance of outcomes. In the future, we plan to explore models and datasets that can effectively manage these challenges.

In summary, this study introduces an effective AI model that reduces bias toward racial, age, and sex subgroups in automated chest radiograph diagnosis. Notably, this represents the first attempt to address bias in deep learning for COVID-19 diagnosis. Our proposed approach uses SCL as a pretraining method to obtain fair image embeddings. Unlike previous supervised contrastive methods, our approach uses images with the same

label but from different protected groups as positive samples and images with different labels but from the same protected group as negative samples for each anchor image in a minibatch. This approach allows the network to capture more label information and less group information during pretraining. The backbone of the model is fine-tuned in the downstream task. We developed and evaluated the proposed method using two large multi-institutional datasets, demonstrating its effectiveness in reducing bias. Therefore, the proposed method may be suitable for clinical practice and can help alleviate concerns regarding disparities generated by AI models.

Author contributions: Guarantors of integrity of entire study, **M.L., Z.S., Y.P.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **M.L., T.L., Y.D., G.S., Y.P.**; clinical studies, **Y.P.**; experimental studies, **M.L., T.L., Z.S., Y.P.**; statistical analysis, **M.L., T.L., Z.S., F.W., Y.P.**; and manuscript editing, **M.L., T.L., G.H., Y.D., F.W., G.S., Y.P.**

Disclosures of conflicts of interest: **M.L.** No relevant relationships. **T.L.** No relevant relationships. **Z.S.** No relevant relationships. **G.H.** No relevant relationships. **Y.D.** No relevant relationships. **F.W.** No relevant relationships. **G.S.** No relevant relationships. **Y.P.** Support for the present manuscript from the National Library of Medicine under award number 4R00LM013001 and the National Science Foundation under award number 2145640.

References

1. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng 2018;2(10):719–731.
2. Lin M, Wynne JF, Zhou B, et al. Artificial intelligence in tumor subregion analysis based on medical imaging: A review. J Appl Clin Med Phys 2021;22(7):10–26.
3. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017; 2097–2106.
4. Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. ArXiv 1711.05225 [preprint] https://arxiv.org/abs/1711.05225. Posted November 14, 2017. Accessed July 21, 2024.
5. Lin M, Liu L, Gordon M, et al. Primary Open-Angle Glaucoma Diagnosis from Optic Disc Photographs Using a Siamese Network. Ophthalmol Sci 2022;2(4):100209.
6. Lin M, Hou B, Liu L, et al. Automated diagnosing primary open-angle glaucoma from fundus image by simulating human's grading with deep learning. Sci Rep 2022;12(1):14080.
7. Char DS, Shah NH, Magnus D. Implementing Machine Learning in Health Care - Addressing Ethical Challenges. N Engl J Med 2018;378(11):981–983.
8. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A Survey on Bias and Fairness in Machine Learning. ACM Comput Surv 2021;54(6):1–35.
9. Lin M, Xiao Y, Hou B, et al. Evaluate underdiagnosis and overdiagnosis bias of deep learning model on primary open-angle glaucoma diagnosis in under-served populations. AMIA Jt Summits Transl Sci Proc 2023;2023:370–377.
10. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med 2021;27(12):2176–2182.
11. Lin M, Li T, Yang Y, et al. Improving model fairness in image-based computer-aided diagnosis. Nat Commun 2023;14(1):6261.
12. Abbas A, Abdelsamea MM, Gaber MM. Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network. Appl Intell 2021;51(2):854–864.
13. Minaee S, Kafieh R, Sonka M, Yazdani S, Jamalipour Soufi G. Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning. Med Image Anal 2020;65:101794.
14. Liu Y, Xing W, Zhao M, Lin M. An end-to-end framework for diagnosing COVID-19 pneumonia via Parallel Recursive MLP module and Bi-LTSM correlation. Medical Imaging with Deep Learning. Proc Mach Learn Res 2024;227:416–425. https://proceedings.mlr.press/v227/liu24a.html.
15. Liu Y, Xing W, Zhao M, Lin M. A new classification method for diagnosing COVID-19 pneumonia based on joint CNN features of chest X-ray images and parallel pyramid MLP-mixer module. Neural Comput Appl 2023;35(23):1–13.
16. Joshi N, Burlina P. AI Fairness via Domain Adaptation. ArXiv 2104.01109 [preprint] https://arxiv.org/abs/2104.01109. Posted March 15, 2021. Accessed July 21, 2024.
17. Zhao Q, Adeli E, Pohl KM. Training confounder-free deep learning models for medical applications. Nat Commun 2020;11(1):6010.
18. Du S, Hers B, Bayasi N, Hamarneh G, Garbi R. FairDisCo: Fairer AI in Dermatology via Disentanglement Contrastive Learning. In: Karlinsky L, Michaeli T, Nishino K, eds. Computer Vision – ECCV 2022 Workshops. ECCV 2022. Lecture Notes in Computer Science, vol 13804. Springer, 2022; 185–202.
19. Wu Y, Zeng D, Xu X, Shi Y, Hu J. FairPrune: Achieving Fairness Through Pruning for Dermatological Disease Diagnosis. In: Wang L, Dou Q, Fletcher PT, Speidel S, Li S, eds. Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. MICCAI 2022. Lecture Notes in Computer Science, vol 13431. Springer, 2022; 743–753.
20. Zhang H, Dullerud N, Roth K, Oakden-Rayner L, Pfohl S, Ghassemi M. Improving the Fairness of Chest X-ray Classifiers. Proceedings of the Conference on Health, Inference, and Learning. Proc Mach Learn Res 2022;174:204–233. https://proceedings.mlr.press/v174/zhang22a.
21. Pfohl SR, Foryciarz A, Shah NH. An empirical characterization of fair machine learning for clinical risk prediction. J Biomed Inform 2021;113:103621.
22. Khosla P, Teterwak P, Wang C, et al. Supervised contrastive learning. Adv Neural Inf Process Syst 2020;33:18661–18673.
23. Lakhani P, Mongan J, Singhal C, et al. The 2021 SIIM-FISABIO-RSNA Machine Learning COVID-19 Challenge: Annotation and standard exam classification of COVID-19 chest radiographs. J Digit Imaging 2023;36(1):365–372.
24. Narasimhan H, Cotter A, Gupta M, Wang S. Pairwise Fairness for Ranking and Regression. In: Proceedings of the AAAI Conference on Artificial Intelligence. AAAI, 2020; 5248–5255.
25. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely Connected Convolutional Networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017; 4700–4708.
26. Johnson AEW, Pollard TJ, Berkowitz SJ, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data 2019;6(1):317.
27. Irvin J, Rajpurkar P, Ko M, et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence. AAAI, 2019; 590–597.
28. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. ArXiv 1412.6980 [preprint] https://arxiv.org/abs/1412.6980. Posted December 22, 2014. Accessed July 21, 2024.
29. Vapnik V. Principles of risk minimization for learning theory. Adv Neural Inf Process Syst 1991;4:831–838.
30. Wadsworth C, Vera F, Piech C. Achieving Fairness through Adversarial Learning: an Application to Recidivism Prediction. ArXiv 1807.00199 [preprint] https://arxiv.org/abs/1807.00199. Posted June 30, 2018. Accessed July 21, 2024.