Fairness Of AI Models in vector embedded Chest X-ray representations

Gebreyowhans H. Bahre[1,2,4], **Hassan Hamidi**[1,4],**Andrew Sellergren**[5],
Leo Anthony Celi[3],**Francesco Calimeri** [2],**Laleh Seyyed-Kalantari** [1,4]
[1]York University, [2]University of Calabria, [3]Massachusetts Institute of Technology, [4]Vector Institute,
[5]Google
bahre@yorku.ca

Abstract

As deep learning models and datasets expand, the demand for computational resources and memory storage intensifies; at the same time, data privacy concerns hinder data and model sharing. Hence, accessibility of model training is significantly challenged. Vector embeddings, as compact representations of medical images, offer a solution to the challenges of computational resource demands and data privacy concerns in AI-based medical imaging. In this study we investigate the suitability of vector embeddings as substitutes for original medical images in disease prediction tasks, focusing on performance and fairness. Using datasets like MIMIC-CXR and CheXpert, we find that vector embedding-based models generally improve disease detection performance and mitigate unfairness in diagnosis rates. The reduced demographic signals in these embeddings may contribute to fairer outcomes without compromising performance. Our findings suggest that vector embeddings can enable more accessible and equitable medical computer vision, particularly in low-resource settings.

1 Introduction

Artificial Intelligence (AI) can reduce healthcare costs, burnouts of staff, and geographical and social disparities in care access. AI application in radiology has been showing promising results [Irvin et al., 2019, Wang et al., 2017a, Ahluwalia et al., 2023, Rajpurkar et al., 2018].

However, building effective AI models is challenging, due to the need for extensive data [Akbarian et al., 2023], high-performance computing, human expertise, and the risk of biases and unfairness [Seyyed-Kalantari et al., 2021b,a, Nalla et al., 2024, Banerjee et al., 2023]. Here by unfairness we mean consistent disparate outcomes of an AI model for a predictive task against some, typically vulnerable, subpopulations.

Google recently released a CXR Foundation model[6] that transforms chest radiograph images into information-rich numerical vectors referred to as “vector embeddings”in an inference mode. So far, Google has released the vector embedding representation of the MIMIC-CXR and CheXpert datasets [2,4]. CXR Foundation models have been trained on a vast amount of natural and X-ray images. Notably, using vector embeddings instead of original images reduces or even eliminates the need for complex deep learning algorithm development, huge computation resources, and data storage, thus paving the way to AI access equity. Such practice seems inevitable as models and training datasets grow larger; however, whether vector embedding representations can effectively substitute for raw medical images from both model performance and fairness perspectives is still an open question.

[6] https://github.com/Google-Health/imaging-research/tree/master/cxr-foundation

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

In this work, we evaluate fairness and performance of AI models trained on vector embedding vs chest X-ray images in disease classification tasks. While there are concerns around AI race detection from medical images [Gichoya et. al., 2022] and its impact on AI model fairness, we further explore race and sex detection of AI models from vector embeddings vs medical images. The goal is to verify whether vector embedding representations carry less demographic data (e.g., race or sex) than medical images and explore its impact on model fairness. We compare models' fairness in correct disease diagnosis [Seyyed-Kalantari et al., 2021a] and underdiagnosis (unhealthy patient flagged as healthy) [Seyyed-Kalantari et al., 2021b] in models that are trained on vector embedding and medical images. We perform the analyses on large, publicly accessible vector embeddings of MIMIC-CXR (MIMIC) and CheXpert (CXP) chest X-ray datasets, and a multi-source aggregation of both datasets, referred to as *ALL*. Due to data availability, we use race, sex, and age as sensitive attributes for all datasets, and insurance type as a proxy for socioeconomic status [Seyyed-Kalantari et al., 2021a] in the MIMIC-CXR dataset. The main contribution of our work can be summarized as follows:

- Disease classification of AI model trained on vector embedding across 14 labels.
- Fairness analysis of the vector embedding-based disease detection model.
- Evaluating AI model race and sex detection from vector embedding vs medical images.
- Performing the aforementioned analyses on CheXpert, MIMIMIC-CXR, and their aggregation (ALL) datasets.

To the best of our knowledge, this work is the first benchmark of the above tasks to date. So far, only disease classification of vector embedding-based model on five labels in 234 test samples of the CheXpert has been reported [Sellergren et al., 2022].

2 Related Work

2.1 Fairness and debiasing in medical imaging

Recent studies showcased unfairness of AI models in disease diagnosis across various sensitive attributes and underdiagnosis in chest X-ray disease classification for historically underserved populations [Seyyed-Kalantari et al., 2021a,b, Ahluwalia et al., 2023, Gichoya et al., 2023, Zhang et al., 2022]. Underdiagnosis measured by False Positive Rate (FPR) on the No Finding label demonstrates that the patient has a disease, but the classifiers detect the patient as healthy, potentially leading to receiving no treatment. In the medical imaging domain, Larrazabal et al. [2020] evaluated unfairness under gender imbalance training datasets. Limited efforts have been spent to address unfairness in medical imaging, centred around benchmarking previous debiasing methods [Zhang et al., 2022] and combining fine-tuning and pruning techniques [Marcinkevics et al., 2022]. MEDFAIR framework Zong et al. [2022] assessed machine learning model fairness in medical imaging, highlighting the prevalent bias in Empirical Risk Minimization (ERM) models across various modalities. Also Zhang et al. [2021] evaluate the domain generalization techniques fairness and realize no method outperforms ERM. Unfair AI can lead to escalating unfairness [Bohdal et al., 2023]. Fairness and bias analysis in medical imaging needs domain-specific consideration of sensitive attributes [Heming et al., 2023]. These techniques often reduce performance for privileged groups (e.g. White) rather than improving it for non-privileged (e.g. Black) [Zhang et al., 2022, Marcinkevics et al., 2022].

2.2 Short-cut learning from medical images

AI models can predict human biological age [Lu et al., 2023], sex [Yang et al., 2021, Cao et al., 2021], and race [Gichoya et. al., 2022], and even body mass index Abbasi Bavil et al. [2024] from medical images. This is an undesired outcome as the AI model may use this data to further discriminate against historically underserved populations. We hope that using vector embedding will degrade AI demographic feature detection from medical images and mitigate unfairness, which needs further investigation.

2.3 Vector embedding representation

Foundation models [Bommasani and et al., 2021, Yang et al., 2023], being large-scale deep AI models trained on extensive datasets, can be applied across diverse tasks with minimal fine-tuning.

Google trained a CXR Foundation model and released vector embedding, the vector representation of X-ray images in embedding space [Sellergren et al., 2022]. Vector embeddings condense intricate information into concise vectors with 1376 floating-point representations for each chest X-ray image. The model was initially trained on a large dataset of natural images, JFT-300M dataset [Sun et al., 2017]. Subsequently, it was trained with supervised contrastive learning on noisy labels of normal/abnormal over a dataset of $821,544$ chest radiographs, collected from India and the US [Sellergren et al., 2022]. These datasets include five different hospitals in India, the ChestX-ray14 dataset (from the National Institutes of Health(NIH)), and the US1 dataset (from a hospital system in Illinois, United States). Note the datasets and disease labels in our study were not used to train CXR Foundation models, and the images of our dataset are gathered from different geographical regions.

The disease prediction performance of vector embeddings has been presented for five labels [Sellergren et al., 2022] of the CheXpert dataset on a limited 234 samples. Glocker et al. [2022] conducted a statistical bias analysis on the chest X-ray foundation model developed by Sellergren et al. [2022] on the CheXpert dataset. Their findings revealed that the model embeds characteristics such as biological sex and racial identity. Their disease detection performance shows around 5% degradation from Sellergren et al. [2022], which might be due to different problem setups. Also, their fairness investigation was based on fixed threshold selection leading to a demonstration of unfairness detection in CheXpert vector embedding. Threshold selection significantly impacts fairness analysis [Seyyed-Kalantari et al., 2022], and different values may be chosen based on needs. In the lack of specific preference for the cost of false negative or positive prediction, a common approach focuses on threshold selection based on maximizing the F1 score across all labels [Irvin et al., 2019, Seyyed-Kalantari et al., 2021a, Rajpurkar et al., 2018] which was not the selection criteria for Glocker et al. [2022].

2.4 Transfer learning

While using vector embeddings might resemble transfer learning where a model is pre-trained and its classification head is fine-tuned our approach goes beyond simple transfer learning. In the age of foundation models, we explore the potential of generating enriched vector embeddings that can substitute for original images, removing the need to continuously load, fine-tune, and deploy pre-trained models. This novel approach of utilizing the embedding dataset significantly improves AI accessibility in environments with limited resources, such as instrumentation and expertise, clearly differentiating our method from traditional transfer learning.

3 Methods

3.1 Data

There are two publicly available Chest X-ray vector embedding datasets corresponding to the MIMIC-CXR and Chexpert image datasets. We have done our analysis on these datasets and their aggregation called ALL dataset. MIMIC-CXR[5] dataset, collected from the Beth Israel Deaconess Medical Center in Boston, MA, between 2011 and 2016 [Johnson et al., 2019] and its corresponding vector embedding representation has been released by Google [Sellergren et al., 2023][6]. The Chexpert[7] dataset, which has gathered at the Stanford University Medical Center between October 2002 and July 2017 [Irvin et al., 2019], and its vector embedding representation has been released by Google [8]. Both vector embedding datasets were derived from Google's CXR-foundation model [Sellergren et al., 2022]. Detailed information regarding the datasets, including distribution across patient subgroups and diagnostic labels, can be found in Table A1 in supplementary materials and Table 1. We also aggregated these two datasets to further explore the impact of using multi-source datasets.

We should note while the CXR foundation model could encode new datasets like chest-Xray14 [Wang et al., 2017b], a data sharing agreement prevents us from sharing sensitive health data such as this dataset with a third party (Google) to get the Vector Embedding representation. Also, since chest-

[5] https://physionet.org/content/mimic-cxr-jpg/2.0.0/

[6] https://physionet.org/content/image-embeddings-mimic-cxr/1.0/

[7] https://stanfordaimi.azurewebsites.net/datasets/8cbd9ed4-2eb9-4565-affc-111cf4f7ebe2

[8] https://docs.google.com/forms/d/e/1FAIpQLSek0P-JSwSfonIiZJlz7gOTbL0lugsDug0FUnMhS1zVzpEKlg/viewform

Xray14 has been used for training the Google foundation model with noisy labels of normal/abnormal, we should not conduct our analysis in this dataset to avoid data leakage. By doing our analysis on MIMIC-CXR and CheXpert, we have ensured none of our datasets has been used in training the Google X-ray foundation model.

3.2 Benchmarks

As baselines, we benchmark the following image-based models in MIMIC, CXP and ALL:

- Disease classification model trained on raw chest x-ray images from Seyyed-Kalantari et al. [2021a] and our in-house image-based model trained on ALL dataset.
- Fairness evaluation in performance (area under the ROC Curve (AUC)) correct disease diagnosis and underdiagnosis [Seyyed-Kalantari et al., 2021a,b] and our in-house image-based model trained on ALL.
- race detection from medical images [Gichoya et. al., 2022] for MIMIC-CXR and CheXpert and our in-house sex detection model from medical images across all datasets.

For vector embedding datasets of MIMIC, CXP and ALL, we benchmark the performance of our trained models on:

- The disease classification from chest x-ray vector embedding.
- Fairness evaluation in correct disease diagnosis and underdiagnosis.
- Race and sex detection of AI models from vector embedding.

3.3 Fairness evaluation

In this study, S denotes sensitive attributes, a criterion for eligibility for protection. In particular, $S = \{S_{sex}, S_{age}, S_{race}\}$ for all datasets; and for MIMIC-CXR dataset also $S_{Insurance} \in S$. For every sensitive attribute, we consider a set of protected groups. Here, the protected groups are; $S_{sex} = \{male, female\}$, $S_{race} = \{White, Black, \ Hispanic, Other, Asian, AmericanIndian/Alaskanative\}$, $S_{age} = \{0-20, \ 20-40, 40-60, \ 60-80, 80-\}$, and $S_{insurance} = \{Medicare, Other, Medicaid\}$. Medicaid is a US governmental insurance for low-income families. Thus, we use insurance as a proxy for social economic status.

We evaluate the separation statistical fairness criteria, which, given the true label Y require orthogonality of predicted label $\hat{Y}$ and S_i, $\hat{Y} \perp\!\!\!\perp S_i \mid Y$. Here, Y , $\hat{Y} \in \mathbb{R}^N$ and their elements y_j, $\hat{y}_j \in$ {0, 1}. Here, N is the number of disease labels. In MIMIC-CXR and CheXpert N=14.

Equality of odds [Hardt et al., 2016] notion of fairness satisfies separation criteria by equalizing the True Positive Rate (TPR) and FPR. We evaluate TPR disparities across disease labels [Seyyed-Kalantari et al., 2021a] and FPR differences across “No Finding ”label [Seyyed-Kalantari et al., 2021b]. Similar to [Seyyed-Kalantari et al., 2021a], for binary S_i (e.g sex) the TPR disparity for the *l*th subpopulation within S_i, is given by

$$TPRDisp_{S_i;l} = TPR_{S_i;l} - TPR_{\neg S_i;l}. \tag{1}$$

Also, for the non-binary classification, similar to [Seyyed-Kalantari et al., 2021a], the TPR disparity for the *l*th subpopulation within S_i is given by:

$$TPRDisp_{S_i;l} = TPR_{S_i;l} - \text{Median}\left(\{TPR_{S_i;k}\}_{k=1}^{l}\right). \tag{2}$$

We calculate $TPRDisp_{S_i;l}$ per disease label y_j. For a given y_j, and S_i, the subgroup with maximum $TPRDisp_{S_i;l}$ is the most favorable as it has the largest disparity in favor. The most unfavorable groups revive the highest negative gap and $Gap_{i,j}$ are given by:

$$Gap_{i,j} = \max\left(\{TPRDisp_{S_i;k}\}_{k=1}^{l}\right) - \min\left(\{TPRDisp_{S_i;k}\}_{k=1}^{l}\right) \tag{3}$$

where, $Gap_{i,j}$ denotes the TPR disparity gap per disease label across subpopulations for a given S_i. We then calculate $\mathbb{E}[Gap_{i,j}]$, per S_i, across $\forall y_j$ and report it as the average $Gap_{i,j}$ for a given sensitive attribute. Additionally, we zoom in "No Finding "(no disease diagnosed) label and evaluate the FPRs of this label as it measures the underdiagnosis rate similar to Seyyed-Kalantari et al. [2021b]. A false positive of "No Finding"means the patient has a disease, but the classifier marks the patient as healthy.

3.4 Experiments

We conducted the following three major experiments.

A) Disease classification with vector embedding-based model: We evaluated three separate classifiers trained on vector embeddings of the MIMIC, CXP, and ALL datasets for disease classification and compared their outcomes to classifiers trained on chest X-ray images.

B) Fairness evaluation of vector embedding-base model: We assessed the fairness of the vector embedding-based model in correct disease diagnosis (TPR disparity) and flagging unhealthy patients healthy (underdiagnosis rate) in disease classification task.

C) Race and sex detection using vector embedding: We examine the ability of models trained on vector embeddings to detect race and sex.

3.5 Models

Label (Abbr.)	MIMIC(Img)	MIMIC(Emb)	CXP(Img)	CXP(Emb)	ALL(Img)	ALL(Emb)
Atelectasis (A)	0.837±0.001	0.809±0.001	0.717±0.001	**0.908±0.000**	0.891±0.004	0.887±0.001
Cardiomegaly (Cd)	0.828±0.002	0.805±0.001	0.855±0.003	**0.902±0.000**	0.887±0.004	0.884±0.000
Consolidation (Co)	0.844±0.001	0.826±0.002	0.734±0.004	**0.906±0.000**	0.938±0.003	0.936±0.000
Edema (Ed)	0.904±0.002	0.892±0.000	0.849±0.001	**0.904±0.000**	0.913±0.003	**0.914±0.001**
Enlarged Card (EC)	0.757±0.003	0.728±0.004	0.668±0.005	**0.921±0.000**	0.956±0.002	0.953±0.000
Fracture (Fr)	0.718±0.007	**0.798±0.002**	0.790±0.006	**0.878±0.001**	0.912±0.006	**0.917±0.001**
Lung Lesion (LL)	0.772±0.006	**0.809±0.003**	0.780±0.005	**0.872±0.001**	0.876±0.010	**0.878±0.000**
Lung Opacity (LO)	0.782±0.002	0.769±0.001	0.747±0.001	**0.934±0.000**	**0.898±0.004**	**0.898±0.000**
No Finding (NF)	0.868±0.001	0.867±0.000	0.885±0.001	**0.955±0.000**	0.911±0.005	**0.912±0.001**
Effusion (Ef)	0.933±0.001	0.909±0.000	0.885±0.001	**0.904±0.000**	0.916±0.004	0.911±0.000
Pleural Other (PO)	0.848±0.003	**0.877±0.005**	0.795±0.004	**0.894±0.001**	0.920±0.009	**0.922±0.001**
Pneumonia (Pa)	0.748±0.005	0.745±0.002	0.777±0.003	**0.864±0.000**	0.850±0.007	0.847±0.001
Pneumothorax (Px)	0.903±0.002	0.884±0.001	0.893±0.002	**0.905±0.000**	0.891±0.012	**0.898±0.001**
Sup. Dev. (SD)	0.927±0.001	**0.928±0.000**	0.898±0.001	**0.942±0.001**	0.929±0.006	**0.941±0.000**
Average (Avg)	0.834±0.001	0.832±0.000	0.805±0.001	**0.906±0.000**	0.906±0.006	**0.907±0.000**

Table 1: AUC (mean over 5 runs ± 95% CI) for disease classification, trained on raw chest X-ray image-based model (Img) vs. our models trained on vector embeddings (Emb). The datasets are MIMIC-CXR (MIMIC), CheXpert (CXP), and their aggregation (ALL). The Img baseline of MIMIC and CXP are from Seyyed-Kalantari et al. [2021a]. Here, Sup. Dev. stands for support device.

All disease detection models (i.e., MIMIC-CXR, CXP, ALL(Emb), the classification head of ALL(Img)) and race and sex classification models have two hidden layers. Detailed configurations of all models are provided in Appendix B of supplementary materials. For ALL dataset image-based models, we utilized the DenseNet121, similar to other literatures [Irvin et al., 2019, Pooch et al., 2020, Rajpurkar et al., 2017, Seyyed-Kalantari et al., 2021b, Zhang et al., 2022]. The dataset was partitioned into training, validation, and testing sets according to a $80 - 10 - 10$ split, ensuring no patient overlap. We report AUC and use TPR and FPR for fairness analysis.

4 Results

4.1 Disease classification performance using vector embedding

We present AUC for disease classification over 14 disease labels in MIMIC, CXP, and ALL datasets for both vector embedding-based model (Emb) and image-based model (Img). We used the results presented in Seyyed-Kalantari et al. [2021a] as the baseline for MIMIC and CXP, which itself compared its outcomes with other models [Tanno et al., 2019, Wang et al., 2020, Cohen et al., 2020, Allaouzi and Ben Ahmed, 2019] and achieved SOTA results. For ALL datasets, we trained an in-house image-based model. Notably, ALL datasets in Seyyed-Kalantari et al. [2021a] also include the Chest X-ray 14 dataset, which has been used in training of Google CXR Foundation model [Sellergren et al., 2022]. Therefore, we trained both image-based and vector-embedding models for ALL datasets, including only CXP and MIMIC datasets.

Table 1 shows the AUCs across labels. Our vector embedding-based models perform better on average across all labels in disease classification tasks for CXP and ALL datasets, particularly showing a notable 0.1 AUC boost for CXP. In MIMIC, the image-based model's AUC is negligibly 0.002 higher. The Google CXR Foundation model paper [Sellergren et al., 2022] provides vector embedding-based results for five CXP labels, only for 234 hand-labeled test images, which are not publicly available. However, our test set covered 14 labels on o large test set of $19,471$ images for CheXpert, $21,591$ for MIMIC-CXR, and $41,062$ for ALL datasets. Overall, our AUCs are better or similar for all those five labels [Sellergren et al., 2022], except for Effusion, where ours is 0.03 lower. We report the mean and 95% confidence interval achieved from different random seed. Training was conducted using 20 CPU cores, 32GB RAM, and an NVIDIA RTX 6000 GPU, completing in 7,5, and 12 minutes for MIMIC, CXP, and ALL vector embedding datasets, respectively. In contrast, training the ALL image-based model typically takes about 10 hours. In summary, vector embeddings allow to accomplish the task faster with much lower computational power, and lead to better performance compared to medical images based models [Seyyed-Kalantari et al., 2021b, Tanno et al., 2019, Cohen et al., 2020, Wang et al., 2020, Allaouzi and Ben Ahmed, 2019].

4.2 Fairness Results

4.2.1 TPR Disparities

We have evaluated TPR disparities using Eq. 1 for sex and Eq. 2 for the remaining sensitive attributes. Here, positive and negative disparities reflect biases favouring or unfavouring particular subgroups. Here, the most favorable groups have the largest frequency of positive gaps across 13 disease labels, and the most unfavorable has the largest frequency of negative gaps. Figure 1 shows the distribution of race TPR disparities with 95% CI, sorted by Gap_j for a model trained on the ALL dataset. Here, $\mathbb{E}[Gap_{race,j}], \forall j$, is 0.214, "Support Devices(SD) "has the least gap 0.037 and "Pneumonia(Pn) "has the most 0.376. "Black "patients constantly receive negative TPR disparities in $13/13$ disease labels. We refer to them as the most unfavorable group, while patients with "Other "races reviving the most frequently positive TPR disparities $13/13$ are referred to as the most favorable groups. We plot TPR disparities for remaining sensitive attributes and datasets in Figures $C1$ to $C9$ of supplementary materials. We summarized all TPR disparity average gaps across all labels, the disease with the lowest and highest gap, and the most favorable and unfavorable subpopulation in Table 2. Ideally, we would have negligible TPR disparities across all subgroups, within each label ("No Finding"label has been excluded to focus in disease diagnosis.).

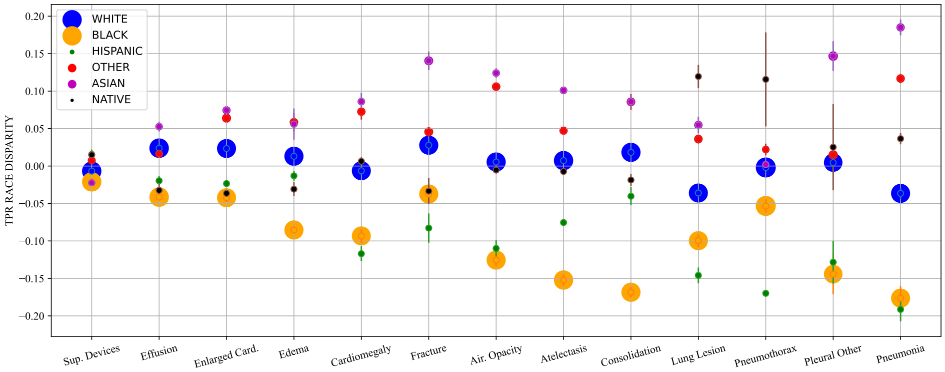

Figure 1: TPR race disparities (mean over 5 run ± 95% CI indicated by arrows) of ALL dataset (y-axis) across disease labels (x-axis). The scatter plot size corresponds to the subgroup sizes per label. Here, positive TPR disparities are favourable, while negative disparities are unfavourable. Notably, Black patients are the unfavourable group for all 13 disease labels, and patients of other racial groups are the most favourable subgroup. For a particular disease, the lower the distance, the fairer the model. We summarized these outcome in Table 2.

Attribute	Dataset	Average	Cross-Label Gap		Unfavorable	Favorable
		Gap	Lowest	Highest		
Sex	ALL(Emb)	**0.042**	Fr:0.007	LL:0.114	Female(10/13)	Male(10/13)
	ALL(Img)	0.069	PE:0.024	Ed:0.139	Female(12/13)	Male(12/13)
	MIMIC(Emb)	**0.071**	PE:0.008	LL:0.217	Female(11/13)	Male(11/13)
	MIMIC(Img)	0.072	Ed:0.011	EC:0.151	Female(10/13)	Male(10/13)
	CXP(Emb)	**0.024**	Pn:0.000	Ed:0.049	Female(9/13)	Male(9/13)
	CXP(Img)	0.062	ED:0.000	Co:0.139	Female(7/13)	Male(7/13)
Age	ALL(Emb)	**0.103**	PE:0.029	Px:0.266	20-40(11/13)	60-80(12/13)
	ALL(Img)	0.122	FR:0.054	EC:0.194	20-40(10/13)	60-80(13/13)
	MIMIC(Emb)	**0.190**	SD:0.059	PE:0.405	80-(9/13)	60-80(9/13)
	MIMIC(Img)	0.245	SD:0.091	Cd:0.440	0-20, 20-40(7/13)	60-80(10/13)
	CXP(Emb)	**0.114**	Co:0.037	Px:0.251	0-20,20-40(10/13)	60-80(13/13)
	CXP(Img)	0.270	SD:0.084	NF:0.604	0-20, 20-40, 80-(7/13)	40-60(8/13)
Race	ALL(Emb)	0.214	SD:0.037	Pn:0.376	Black(13/13)	Other(13/13)
	ALL(Img)	0.183	EC:0.113	PX:0.316	Black(13/13)	Asian(13/13)
	MIMIC(Emb)	0.280	Cd:0.109	Px:0.663	Black,Asian(9/13)	White(10/13)
	MIMIC(Img)	0.226	NF:0.119	Pa:0.440	Hispanic(9/13)	White(9/13)
	CXP(Emb)	**0.100**	LL:0.035	Fr:0.186	Black,Native(12/13)	White,Asian(10/13)
	CXP(Img)	0.119	Fr: 0.055	At:0.215	Native(9/13)	Other(7/13)
Insurance	MIMIC(Emb)	**0.008**	At:0.0005	Co:0.029	Medicare(8/13)	Other(9/13)
	MIMIC(Img)	0.100	SD:0.021	PO:0.190	Medicaid(10/13)	Other(10/13)

Table 2: Summary of TPR disparities across sensitive attributes for image-based (Img) [Seyyed-Kalantari et al., 2021a] versus vector embedding-based (Emb) models. We calculate the $\mathbb{E}[Gap_{i,j}], \forall i, \forall j$, as listed in the Average Gap column. A smaller average gap indicates a fairer model in disease diagnosis. The lowest and highest gaps per attribute/dataset, along with their values, are shown (full disease names in Table 1). The most Unfavorable/favorable subgroups have also been shown. Only MIMIC has insurance data.

In cases of minimal average gap, our model shows improved fairness regarding TPR disparity. As before, compare fairness between models trained on vector embeddings (Emb) and images (Img), with baseline results from Seyyed-Kalantari et al. [2021a], except for ALL. Vector embedding models consistently show a lower average gap for sex, age, and insurance attributes across MIMIC, CXP, and ALL datasets, indicating fairer outcomes compared to image models. However, for race in ALL and MIMIC, vector embeddings have a higher gap. The most and least favored subgroups generally remain unchanged between vector embedding and image models.

4.2.2 Underdiagnosis

For CXP, Fig. 2 shows the underdiagnosis rate using vector embeddings vs medical images across subgroups of sex, age, and race and the patients' intersection with two/three underserved groups. We report the baseline results from Seyyed-Kalantari et al. [2021b], shown in gray color in Fig. 2. We exclude groups with fewer than 10 patients with FPR from the plot to avoid conclusions based on small subsets. No three-group intersections meet this criterion, so we do not provide such plots.

Vector embedding reduces the underdiagnosis rate and narrows the fairness gap between the maximum and minimum rates per sensitive attribute, improving fairness in the max-min gap of underdiagnosis. We also evaluate the underdiagnosis rate for the MIMIC-CXR and ALL datasets. Table $D1$ in the supplementary materials summarizes underdiagnosis rate fairness, with detailed findings in Figures $D1$ and $D2$.

For the MIMIC-CXR dataset, vector embedding models reduced underdiagnosis rates and max-min gaps across all subgroups compared to image-based models. In ALL data, both models show similar underdiagnosis rates and max-min gaps (Fig $D1$). The image model has a slightly lower FPR for three age subgroups, but the difference is minimal, with the max-min gap only 0.002 higher for vector embedding.

4.3 Sex and race detection using vector embedding

We aim to determine if models can learn sensitive features like race and sex when using vector embeddings. Lower detection of these features is preferred, as using demographic data may lead to unfairness. Table $C1$ in the supplementary materials shows the AUC for sex and race detection in

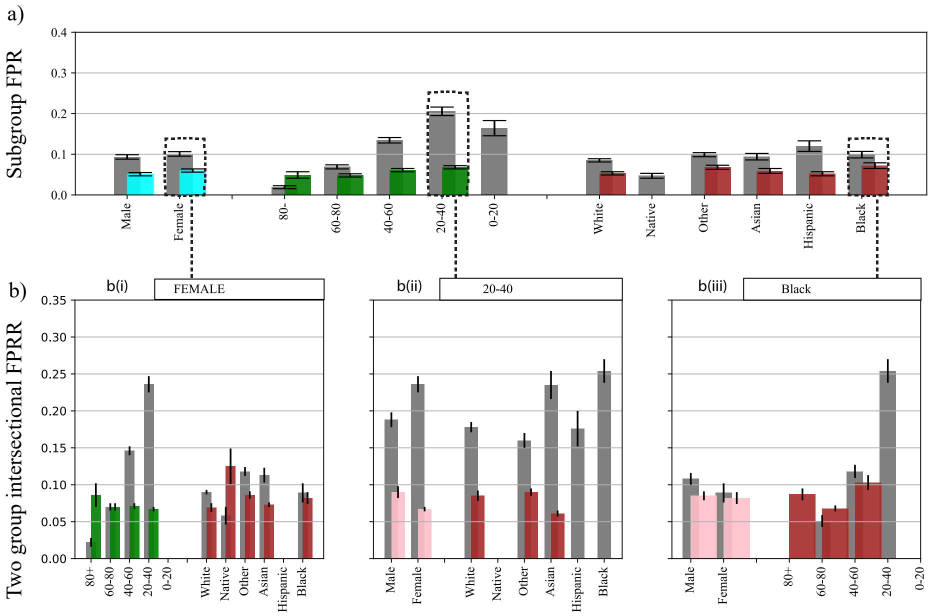

Figure 2: Exploration of underdiagnosis rates. (a) rates across sex, age, and race subgroups in CheXpert. (b), Two group intersection underdiagnosis rates for (b(i)) female, (b(ii)), 20-40, and (b(iii)) Black patients amidst all other subgroups. Subgroups with fewer than ten FPR occurrences are excluded. The gray bar represents the image-based model from Seyyed-Kalantari et al. [2021b]. Here, using vector embeddings reduced the max-min FPR gap and overall underdiagnosis rate,leading to more fairness. Most underdiagnosid groups and max-min gap are presented in Table $D1$ of supplementary materials.

different settings. While vector embeddings still carry these signals, detecting race and sex is easier in images, as shown by the lower AUC in vector embeddings.

5 Discussion

5.1 Vector embeddings: reliable substitute for X-ray images

Disease classification performance: On average, vector embedding-based disease classifiers outperform image-based models across all labels in CheXpert and multi-source ALL datasets (see Table 1). In MIMIC-CXR, the image-based model only slightly outperforms by 0.002, which is negligible compared to the computational savings. Thus, vector embeddings are a reliable substitute for raw images for AI model training.

Fairness: For fairness, we compared TPR disparity, underdiagnosis rate, and max-min gap in underdiagnosis. Vector embeddings generally improve TPR disparity across all labels in most of the dataset-sensitive attribute setup pairs, reducing the gap in 8 of 10 sensitive attribute setups (see Table 2). Similarly, vector embeddings often reduce both the underdiagnosis rate and the max-min gap compared to image-based models (Table $D1$), doing so in 7 out of 10 dataset-attribute pairs. For cases where the gap isn't smaller, the difference is minimal, ranging from 0.002 to 0.007. This outcome indicates greater fairness in the vector embedding model compared to the image-based model. We also examined multi-source data, where both model types perform similarly in disease detection, showing minimal max-min gaps in underdiagnosis and often small average diagnosis gaps. This suggests that large multi-source datasets can reduce disparities, aligning image-based models more closely with the representations learned by foundation models.

Voulnerable groups: The vector embedding-based model does not alter vulnerable subgroups, with female, younger, and Black patients still being the most underdiagnosed (see Tables 2 and Table $D1$ in supplementary materials). Additionally, TPR disparity shifts from Medicaid to Medicare when using vector embeddings. This group represents retired patients, typically of lower socioeconomic status, with Medicaid remaining the most underdiagnosed. Groups with multiple vulnerable traits, such as Black females, face higher underdiagnosis rates than white females, indicating amplified bias. These findings align with previously identified vulnerable groups in healthcare [Abdelmalek et al., 2023] and medical imaging [Seyyed-Kalantari et al., 2021a,b], reflecting existing societal biases.

Diversity and the size of data: The image-based and vector embedding-based models demonstrate similar performance in disease detection and underdiagnosis rates across various datasets and attributes. The multi-source dataset is notably larger and more diverse than individual datasets. These features may help achieve performance closer to the vector embedding dataset, originally derived from a foundation model trained on large, diverse data. Similarly, vector embedding yields greater performance improvements in the CheXpert dataset compared to the MIMIC-CXR dataset, as CheXpert was initially smaller. These findings suggest that vector embedding may offer greater benefits in fairness and performance with smaller, less diverse original datasets. As data size increases, the advantages of using vector embedding or image-based models for improved performance and fairness diminish. Nonetheless, vector embedding still provides the benefit of faster training with lower computational resources.

Generalizability: Across datasets, vector embedding-based models consistently improved model fairness compared to image-based models. However, the vulnerable subgroups remained unchanged with vector embedding. It's important to note that fairness analysis outcomes on binary predictions can vary significantly with different thresholds. In this work, similar to prior studies [Seyyed-Kalantari et al., 2021a,b, Rajpurkar et al., 2017], we use the threshold that maximizes the F1 score across all labels, treating precision and recall equally. However, one can set the threshold to achieve a fixed FPR for disease classification [Glocker et al., 2022]. The choice of threshold depends on the specific problem and the associated costs of precision and recall in the downstream task.

5.2 The fairer, the blinder to demographic features

Our findings suggest that demographic features such race and sex persists in vector embedding but the race and sex detection performance is less than image-based model. Concurrently, vector embeddings reduce unfairness in disease diagnosis and underdiagnosis rates compared to image-based models.Digging into numbers among three datasets, CXP has more fairness (less average gap) in correct disease diagnosis (See Table 2) and less max-min gap in underdiagnosis rate analysis (See Table $D1$ of supplementary materials. This co-occurs with often less sex and race detection performance (See Table $C1$ of supplementary materials). In particular, for the CheXpert dataset, we observe the race signal dropped more in vector embedding, co-occurring with higher performance in disease detection (See Table 1), and less unfairness (See Table 2). Such observation amplifies the importance of learning representation with less sensitive signals to mitigate unfairness.

5.3 Vector embedding: AI equity and lower environmental damage

Our work shows that vector embeddings enhance AI efficiency and fairness while reducing memory and GPU usage, leading to lower carbon emissions and environmental impact. This approach makes AI more accessible to those with limited computational resources or expertise. Releasing and using vector embedding datasets as image substitutes can promote global AI equity. As AI models grow and become constrained to high-tech companies, vector embeddings offer a viable alternative for those lacking advanced computing infrastructure.

6 Limitations and Future Work

Considering the potential benefits showcased by the vector embedding dataset, we propose the expansion of producing vector embedding versions of diverse datasets. This expansion will broaden our fairness analysis to include a wider range of vector embedding datasets, diverse demographic profiles, and various analytical techniques. Our work relies on two only available vector embedding datasets, MIMC-CXR and CheXpert, along with their aggregations. In addition, the backbone CXR foundation model [Sellergren et al., 2022] that generated the vector embeddings is trained on data collected from limited resources in the USA and India, raising concerns about data shift and drift. Using a larger and diversified dataset for these foundation models potentially leads to a more generalizable representation of learning. We plan to develop a fair vector embedding representation for future work that leads to fairer outcomes. Considering recent progress in large language models (LLMs), We also plan to consider multi-modality in analyzing the vector embedding or learning fair vector embeddings. In doing so, the fairness of applied LLMs needs to be considered so as not to enforce extra biases [Tian et al., 2023]. Following the hints from this research, locating demographic signals [Salvado et al., 2024] and disentangling or mitigating demographic signals from vector embedding representation seems to be a plausible path to reach our goal. We will also generate vector embedding representations for diverse public medical image datasets and release them for the public community's use.

7 Conclusion

We examined the fairness and performance of the disease classification AI model using vector embedding datasets and image-based datasets. Overall, the vector embedding-based model outperforms or has a negligible drop in disease classification performance and improved fairness compared to the image-based model, suggesting vector embeddings are a proper substitute for medical images in AI model training. We observed large and multi-source datasets demonstrate less difference in fairness and performance between models based on vector embedding and image. Additionally, there are fewer demographic features such as race and sex information in vector embedding vs images, which may guide researchers to look for ways to learn representation with fewer demographic features to reach better fairness. We should also note training a model for the classification of vector embedding datasets requires less computational power and specialized knowledge while promoting privacy and equity in AI access and reducing negative computational environmental impact.

8 Acknowledgments

This work has received partial support from the PNRR project FAIR - Future AI Research (PE00000013), Spoke 9 - Green-aware AI, under the NRRP MUR program funded by NextGenerationEU(to Gebreyowhans H. Bahre). We sincerely thank the Natural Sciences and Engineering Research Council of Canada (NSERC) for the Discovery Grant (to Dr. Laleh Seyyed-Kalantari) and the Connected Minds CFREF (to Dr. Seyyed-Kalantari). The authors also express their gratitude to the Vector Institute for providing high-performance computing platforms.

References

E. Abbasi Bavil, M. Ahluwalia, L. Seyyed-Kalantari, B. Fine, and M. Abdalla. Body mass index prediction from chest radiographs and associated performance gaps in chest radiograph abnormality prediction. In *CAR 2024 Annual Scientific Meeting (ASM)*, Montréal, Canada, 2024.

Fred M. Abdelmalek, Federico Angriman, Julie Moore, Kuan Liu, Lisa Burry, Laleh Seyyed-Kalantari, Sangeeta Mehta, Judy Gichoya, Leo Anthony Celi, George Tomlinson, Michael Fralick, and Christopher J. Yarnell. Association between patient race and ethnicity and use of invasive ventilation in the united states. *Annals of the American Thoracic Society*, 21(2):Specific Page Range, Nov 2023. doi: SpecificDOI. Impact Factor: 8.3.

M. Ahluwalia, M. Abdalla, J. Sanayei, L. S. Kalantari, M. Hussain, A. Ali, and B. Fine. The subgroup imperative: Chest x-ray classifier generalization gaps in patient, setting, and pathology subgroups. *Radiology: Artificial Intelligence*, 5(2):e230020, 2023. doi: 10.1148/ryai.230020. URL `https://doi.org/10.1148/ryai.230020`.

S. Akbarian, L. Seyyed-Kalantari, F. Khalvati, and E. Dolatabadi. Evaluating knowledge transfer in the neural network for medical images. *IEEE Access*, 11:85812–85821, 2023. doi: 10.1109/ACCESS.2023.3283216.

Imen Allaouzi and Mohamed Ben Ahmed. A novel approach for multi-label chest x-ray classification of common thorax diseases. *IEEE Access*, 7:64279–64288, 2019. doi: 10.1109/ACCESS.2019.2916849.

I. Banerjee, K. Bhattacharjee, J. L. Burns, H. Trivedi, S. Purkayastha, L. S. Kalantari, B. N. Patel, R. Shiradkar, and J. Gichoya. "shortcuts" causing bias in radiology artificial intelligence: Causes, evaluation, and mitigation. *Journal of the American College of Radiology*, 20(9), September 2023.

Ondrej Bohdal, Timothy Hospedales, Philip HS Torr, and Fazl Barez. Fairness in ai and its long-term implications on society. *arXiv preprint*, arXiv:2304.09826, 2023. URL https://arxiv.org/abs/2304.09826.

Rishi Bommasani and et al. On the opportunities and risks of foundation models. *arXiv preprint*, 2021. URL https://arxiv.org/abs/2108.07258.

Y. Cao, Y. Ma, D.N. Vieira, et al. A potential method for sex estimation of human skeletons using deep learning and three-dimensional surface scanning. *International Journal of Legal Medicine*, 135:2409–2421, 2021. doi: 10.1007/s00414-021-02675-z. URL https://doi.org/10.1007/s00414-021-02675-z.

Joseph Paul Cohen, Muhammad Hashir, Rupert Brooks, and Hannah Bertrand. On the limits of cross-domain generalization in automated x-ray prediction. *arXiv preprint arXiv:2002.02497*, 2020.

Judy Wawira Gichoya, Kaesha Thomas, Leo Anthony Celi, Nabile Safdar, Imon Banerjee, John D Banja, Laleh Seyyed-Kalantari, Hari Trivedi, and Saptarshi Purkayastha. Ai pitfalls and what not to do: mitigating bias in ai. *British Journal of Radiology*, 96(1150), October 2023. doi: 10.1259/bjr.20230023. URL https://doi.org/10.1259/bjr.20230023.

Judy Wawira Gichoya et. al. Ai recognition of patient race in medical imaging: a modelling study. *The Lancet. Digital health*, 4:e406–e414, 2022. doi: 10.1016/S2589-7500(22)00076-2.

Ben Glocker, Charles Jones, Melanie Bernhardt, and Stefan Winzeck. Risk of bias in chest x-ray foundation models. *arXiv preprint*, 2022. URL https://arxiv.org/abs/2209.02965.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413*, 2016. https://arxiv.org/abs/1610.02413.

Carolina A. M. Heming, Mohamed Abdalla, Monish Ahluwalia, Linglin Zhang, Hari Trivedi, MinJae Woo, Benjamin Fine, Judy Wawira Gichoya, Leo Anthony Celi, and Laleh Seyyed-Kalantari. Benchmarking bias: Expanding clinical ai model card to incorporate bias reporting of social and non-social factors, 2023.

Jeremy A. Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David Andrew Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):590–597, July 2019. doi: 10.1609/aaai.v33i01.3301590. URL https://doi.org/10.1609/aaai.v33i01.3301590.

Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *Computer Vision and Pattern Recognition*, 2019.

Agostina J. Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H. Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, 2020.

T. Lu, Yr. Diao, Xe. Tang, et al. Deep learning enables automatic adult age estimation based on ct reconstruction images of the costal cartilage. *Eur Radiol*, 33:7519–7529, 2023. doi: 10.1007/s00330-023-09761-3. URL https://doi.org/10.1007/s00330-023-09761-3.

Ricards Marcinkevics, Ece Ozkan, and Julia E. Vogt. Debiasing deep chest x-ray classifiers using intra-and post-processing methods. In *Machine Learning for Healthcare Conference*, pages 504–536. PMLR, 2022.

Vineela Nalla, Seyedamin Pouriyeh, Reza M. Parizi, Hari Trivedi, Quan Z. Sheng, Inchan Hwang, Laleh Seyyed-Kalantari, and MinJae Woo. Deep learning for computer-aided abnormalities classification in digital mammogram: A data-centric perspective. *Current Problems in Diagnostic Radiology*, 53(1):1–10, January 2024. doi: 10.1067/j.cpradiol.2023.10.001. URL `https://doi.org/10.1067/j.cpradiol.2023.10.001`.

Eduardo H P Pooch, Pedro Ballester, and Rodrigo C Barros. Can we trust deep learning based diagnosis? the impact of domain shift in chest radiograph classification. In *Thoracic Image Analysis. TIA 2020*, volume 12502, Cham, 2020. Springer. doi: 10.1007/978-3-030-62469-9_7. URL `https://doi.org/10.1007/978-3-030-62469-9_7`.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P Lungren, and Andrew Y Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

Pranav Rajpurkar, Jeremy Irvin, Robyn L Ball, Kaylie Zhu, Brandon Yang, Harsh Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P Langlotz, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLoS Med*, 15(11): e1002686, Nov 2018. doi: 10.1371/journal.pmed.1002686.

Olivier Salvado, Salamata Konate, Rodrigo Santa Cruz, Andrew Bradley, Judy Wawira Gichoya, Laleh Seyyed-Kalantari, Brandon Price, Clinton Fookes, and Leo Lebrat. Localisation of racial information in chest x-ray for deep learning diagnosis. In *Proceedings of the International Symposium on Biomedical Imaging (ISBI)*, Athens, Greece, 2024.

A. Sellergren, A. Kiraly, T. Pollard, W. Weng, Y. Liu, A. Uddin, and C. Chen. Generalized image embeddings for the mimic chest x-ray dataset (version 1.0). PhysioNet, 2023. `https://doi.org/10.13026/pxc2-vx69`.

A. B. Sellergren, C. Chen, Z. Nabulsi, Y. Li, A. Maschinot, A. Sarna, J. Huang, C. Lau, S. R. Kalidindi, M. Etemadi, F. Garcia-Vicente, D. Melnick, Y. Liu, K. Eswaran, D. Tse, N. Beladia, D. Krishnan, and S. Shetty. Simplified transfer learning for chest radiography models using less data. *Radiology*, 305(2):454–465, 2022. doi: 10.1148/radiol.212482.

Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew B. A. McDermott, and Marzyeh Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. *Pacific Symposium on Biocomputing*, 26:232–243, 2021a.

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew BA McDermott, Irene Y. Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27(12):2176–2182, 2021b.

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B.A. McDermott, Irene Y. Chen, and Marzyeh Ghassemi. Reply to: 'potential sources of dataset bias complicate investigation of underdiagnosis by machine learning algorithms' and 'confounding factors need to be accounted for in assessing bias by machine learning algorithms'. *Nature Medicine*, 28:1161–1162, 2022. doi: 10.1038/s41591-022-01854-8. URL `https://www.nature.com/articles/s41591-022-01854-8`.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era, 2017.

Ryosuke Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C Alexander, and Nathan Silberman. Learning from noisy labels by regularized estimation of annotator confusion. *arXiv preprint arXiv:1902.03680*, 2019.

Jacob-Junqi Tian, D. Emerson, Deval Pandya, Laleh Seyyed-Kalantari, and Faiza Khattak. Efficient evaluation of bias in large language models through prompt tuning. In *Proceedings of the Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=v1WL01lgp8`.

Xiaogang Wang, Zhen Xu, Dongkuan Yang, Lai Tam, Holger Roth, and Dong Xu. Learning image labels on-the-fly for training robust classification models. *arXiv preprint arXiv:2009.10325*, 2020.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1(2):2097–2106, July 2017a. doi: 10.1109/CVPR.2017.845. URL `https://doi.org/10.1109/CVPR.2017.845`.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In *Computer Vision and Pattern Recognition (CVPR) 2017*, pages 2097–2106. IEEE, 2017b. URL `http://openaccess.thecvf.com/content_cvpr_2017/html/Wang_ChestX-ray8_Hospital-Scale_Chest_CVPR_2017_paper.html`.

Chung-Yi Yang, Yi-Ju Pan, Yen Chou, Chia-Jung Yang, Ching-Chung Kao, Kuan-Chieh Huang, Jing-Shan Chang, Hung-Chieh Chen, and Kuei-Hong Kuo. Using deep neural networks for predicting age and sex in healthy adult chest radiographs. *Journal of Clinical Medicine*, 10(19):4431, 2021. doi: 10.3390/jcm10194431. URL `https://doi.org/10.3390/jcm10194431`.

Sherry Yang, Ofir Nachum, Yilun Du, Jason Wei, Pieter Abbeel, and Dale Schuurmans. Foundation models for decision making: Problems, methods, and opportunities. *arXiv preprint*, 2023.

H. Zhang, N. Dullerud, L. S. Kalantari, Q. Morris, Shalmali Joshi, and M. Ghassemi. An empirical framework for domain generalization in clinical settings. In *ACM Conference on Health, Inference, and Learning (CHIL)*, Virtual, 2021.

Haoran Zhang, Natalie Dullerud, Karsten Roth, Lauren Oakden-Rayner, Stephen Pfohl, and Marzyeh Ghassemi. Improving the fairness of chest x-ray classifiers. In *Conference on Health, Inference, and Learning*, pages 204–233. PMLR, 2022.

Yongshuo Zong, Yongxin Yang, and Timothy Hospedales. Medfair: Benchmarking fairness for medical imaging. *arXiv preprint*, arXiv:2210.01725, 2022. URL `https://arxiv.org/abs/2210.01725`.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer:[Yes]

   Justification: The main claims in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses potential limitations of the work performed.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The paper includes all necessary assumptions and complete, correct proofs for each theoretical result, with proper references and cross-referencing.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [No]

Justification: The paper does not fully disclose all information needed for reproducing the experimental results, as the code will not be released until the paper is accepted for publication at another conference. However, the dataset is openly available for use in experimentation.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
    (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
    (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
    (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
    (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The paper does not provide open access to the code, as the code will not be released until the paper is accepted for publication at another conference. However, the dataset is openly available for experimentation.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper does not fully disclose all information needed for reproducing the experimental results, as the code will not be released until the paper is accepted for publication at another conference. However, the dataset is openly available for use in experimentation.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The models for each experiment are trained five times with different seed numbers, and results are reported with a 95% confidence interval for robustness, with error bars appropriately included.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer:[Yes]

   Justification: The compute resources used and run times are clearly explained in the results section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics as it ensures informed consent, protects participant privacy, adheres to data integrity standards, and considers the societal impact of the findings.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

Justification: This research does not discuss any negative societal impacts. It is focused on the performance and fairness of AI models utilizing vector-embedded chest x-ray datasets, which are primarily intended for improving healthcare outcomes.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer:[Yes]

Justification: The datasets used in this research are deidentified, which significantly reduces the risk of exposing personal information. Deidentification involves removing or altering identifiable information so that individuals cannot be readily identified.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper properly credits the creators of all assets used, explicitly mentions the licenses and terms of use, and includes the relevant citations, asset versions, and URLs where applicable. Each asset's license type is clearly stated, ensuring compliance with copyright and terms of service.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer:[Yes]

    Justification: The new assets introduced in the paper are well documented, with comprehensive details provided alongside the assets, including information on training, limitations, and consent processes where applicable.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer:[NA]

    Justification: This work does not involve with crowd sourcing and Research with Human Subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [No]

Justification: This work does not require Institutional Review Board (IRB) approvals or equivalent review for research involving human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.