Embedding vs Image-Based AI: A Comparative Fairness Study in Chest X-ray Analysis

Gebreyowhans H. Bahre[1,2,5], Hassan Hamidi [1,5], Andrew B. Sellergren[4], Leo Anthony Celi[3], Francesco Calimeri[2], Laleh Seyyed-Kalantari[1,5]

[1]York University, Toronto, Canada
[2]University of Calabria, Rende, Italy
[3]Laboratory for Computational Physiology, Massachusetts Institute of Technology, USA
[4]Google Health, Google, USA
[5]Vector Institute, Toronto, Canada
bahre@yorku.ca, hhamidi@yorku.ca, asellerg@google.com, lceli@mit.edu, francesco.calimeri@unical.it, lsk@yorku.ca

Abstract

AI has shown remarkable potential in healthcare, but faces accessibility challenges due to high computational and expertise demands, especially in medical image analysis. Vector embeddings, compact representations of medical images achieved from foundation models in zero-shot inference, offer a potential solution. Recently, an equivalent vector embeddings dataset of existing large publicly available medical images has been released, for which training an AI model requires significantly lower computing infrastructure and storage needs. Such data sets provide greater accessibility to AI in medical imaging for those who do not have access to large computing resources. The burning question remains: What is the gain or loss in using vector embedding to replace medical images, particularly from a fairness and utility point of view? In this work, we compare AI models trained in vector embeddings (Emb) with raw chest radiograph images for disease diagnosis, focusing on both performance and fairness. Our results show that Emb-based models match or exceed image-based models in diagnostic performance while improving fairness. Crucially, Emb achieve this with far less computational cost. These findings position Emb as a powerful, scalable alternative to image-based AI, especially valuable for low-resource settings where access to GPUs and expert infrastructure is limited.

Introduction

Artificial Intelligence (AI) has demonstrated highly promising outcomes across various biomedical and healthcare fields, including radiology (Jeremy et al. 2019; Xiaosong et al. 2017; Akbarian et al. 2023). However, these applications face several critical challenges, such as high computational demands, substantial data storage needs, a shortage of domain experts, and concerns regarding fairness in decision-making (Seyyed-Kalantari et al. 2021b,a; Konate et al. 2025; Gichoya et al. 2023; Ahluwalia et al. 2023). Fairness issues arise when AI models systematically generate unequal results for underrepresented or vulnerable subpopulations (Seyyed-Kalantari et al. 2021a; Yang et al. 2024; Larrazabal et al. 2020; Zhang et al. 2022; Marcinkevics, Ozkan, and Vogt 2022; Salvado et al. 2024).

Google introduced a CXR foundation model (Sellergren et al. 2022) (FM) that converts chest X-ray images into numerical representations called "vector embeddings (Emb)". This foundation model was trained on $JFT-300M$ natural images and further refined through supervised contrastive learning using noisy normal/abnormal labels from $821,544$ chest X-rays collected in India and the US (ChestX-ray14 (Xiaosong et al. 2017) and US1 from Illinois) (Sellergren et al. 2022). Embeddings for the MIMIC-CXR (MIMIC) (Johnson et al. 2019a) and CheXpert (CXP) (Jeremy et al. 2019) datasets generated from this CXR Foundation model have also been made publicly available. These embeddings represent the static zero-shot inference of the foundation model on entirely new datasets, MIMIC and CheXpert, where the model was neither trained on nor exposed to during training, including for detecting their specific disease labels on any dataset.

Training models using Emb offer several advantages, including reduced algorithmic complexity, lower computational costs, and decreased storage requirements. As the size of AI models and datasets continues to grow, making AI less accessible, the use of Emb representations instead of original image datasets appears to be an increasingly necessary approach, especially for low-resource communities.

In this context, a comprehensive comparison between models trained on Emb, obtained from a foundation model, and those trained on traditional image data is essential. Particlularly, medical images has shown to contain detactable demographic features of the patients such as sex, race (Konate et al. 2025; Salvado et al. 2024; Gichoya et al. 2022), while AI models trained on them have lower performance for some races or sexes. However, Emb are much lower representation and may contain less demographic information. Such a comparison can reveal the distinct patterns each approach captures through its training process. The findings can also help determine whether Emb can serve as a viable alternative to the original images, offering comparable or improved fairness and performance while benefiting from the advantages previously outlined. In this study, we investigate whether Emb datasets can effectively serve as a substitute for chest X-ray image datasets. To achieve this, we assess both the performance and fairness of Emb-based models in comparison to

traditional medical image-based AI models in disease classification tasks. In our context, "unfairness" refers to situations where AI models produce unequal outcomes across different sub-populations, typically favoring already privileged groups. We examine fairness by evaluating the consistency of correct disease diagnoses across various demographic groups (e.g., sex, race, age, and socioeconomic status) for each disease label. Although using Embeddings might seem similar, using Emb goes beyond this paradigm. We use embeddings as static representations obtained through *zero-shot inference* from the foundation model, which, first, has never encountered our data and second, has never been trained for our specific task. This method eliminates the need to load, fine-tune, and deploy large pre-trained models. In contrast, transfer learning requires repeatedly loading and tuning pre-trained models for each new task. Additionally, we investigate the use of enriched vector embeddings as a replacement for original images, an objective not addressed by transfer learning. This approach addresses scalability and accessibility challenges that traditional transfer learning is not designed to solve.

Our main contribution can be summarized as follows.

- Assessment of disease classification performance of the Emb-based model across 14 labels.
- Fairness analysis of the correct disease diagnoses across disease labels and demographic features
- Comparison of fairness and performance against "standard" image-based trained models, with the aim of validating Emb as a proper replacement for images.
- Analyses are performed on existing large, publicly available Chest-Xray Emb datasets: MIMIC (Johnson et al. 2019a) and CXP (Jeremy et al. 2019) chest X-ray datasets, and their multi-source aggregation (ALL).
- We found that the Emb-based model vs image-based model reduces unfairness and does not often change the vulnerable groups.
- Emb-based model is significantly faster, approximately 50 times faster than the image-based model, leading to more accessibility of AI in medical imaging in low computational resource settings.

Our codes are available at https://github.com/Gebreyowhans/Emb_fairness.

Dataset and Methods

Datasets

We utilized vector embeddings from publicly available MIMIC (Johnson et al. 2019b) and CXP (Jeremy et al. 2019) datasets, both generated using Google's CXR foundation model (Sellergren et al. 2022). To further examine the effects of combining data sources, we also aggregated these two datasets. Unlike the original datasets, the embeddings used in this study exclude lateral view images. Comprehensive details about the datasets, including their distribution across patient subgroups, are provided in Table 2.

Subgroup	Attribute	MIMIC	CXP	ALL
	# Images	227,641	219,946	447,587
	# Patients	52,874	63,467	116,341
Sex	Male	54.1 %	59.3 %	56.7 %
	Female	45.9 %	40.7 %	43.3 %
Age	0-20	0.4 %	0.8 %	0.6 %
	20-40	10.6 %	13.1 %	11.8 %
	40-60	30.6 %	31 %	30.8 %
	60-80	42.2 %	39.1 %	40.7 %
	80-	16.3 %	16.1 %	16.2 %
Race	White	66.1 %	63.3 %	64.8 %
	Black	16.1 %	6.1 %	11.4 %
	Asian	3.2 %	11.8 %	7.2 %
	Hispanic	5.5 %	2.4 %	4 %
	Native	0.3 %	1.9 %	1.2 %
	Other	4.7 %	14.6 %	9.3 %
Insurance	Medicare	44.7 %		
	Other	47.1 %		
	Medicaid	8.2 %		

Table 1: The chest X-ray vector embedding datasets used in this study include MIMIC, CXP, and their combined aggregation (ALL).

Dataset	Training	Validation	Test
MIMIC	182,896	22,855	22,855
CXP	178,402	22,085	22,073
ALL	361,298	41,063	41,063

Table 2: Training, validation, and test splits using the 80-10-10 method across all datasets (i.e., MIMIC, CXP, and ALL).

Benchmarks

For the MIMIC, CXP, and their combined dataset (referred to as ALL in our study), we benchmark disease classification performance using an image-based model from Seyyed-Kalantari et al. (Seyyed-Kalantari et al. 2021a), along with our inhouse developed image-based model trained on the ALL dataset (baseline), and compare these with the Emb-based model. We then evaluate fairness in terms of correct disease diagnosis (Seyyed-Kalantari et al. 2021a), measured using the true positive rate (TPR), by comparing the baseline image-based model to the Emb-based model. The image-based models used for MIMIC and CXP are the same as those from (Seyyed-Kalantari et al. 2021a), while for the ALL dataset, we developed a custom in-house model trained on the combined MIMIC and CXP data.

Models and Training

All disease classification models, MIMIC, CXP, ALL (Emb), and the classification head of ALL (Img), shared the same architecture, consisting of two hidden layers with 768 and 256 units, respectively. Each layer used ReLU activation, batch normalization, and a dropout rate of 0.3. Models were trained using a batch size of 48, a weight decay of $1e-5$, and an initial learning rate of $1e-4$, optimized with the

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±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±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±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±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 3: Disease classification AUC (mean over 5 runs ± 95% CI) for image-based (Img) versus our vector embedding-based (Emb) models across MIMIC, CXP, and their combined dataset (ALL). The Img baselines for MIMIC and CXP are sourced from (Seyyed-Kalantari et al. 2021a). Bold values indicate cases where the Emb-based model outperforms the Img-based model.

Adam optimizer and a cosine decay learning rate schedule (Loshchilov and Hutter 2016). The output layer used sigmoid activation and was trained using the BCEWithLogitsLoss function. Each model included a fully connected classification layer with 14 output neurons to generate probability scores for each disease label. The final binary predictions were made using optimized per-label thresholds for the $F1$ score.

To establish a baseline for the ALL(Img) dataset, we trained a DenseNet-121 model, an architecture widely adopted in chest X-ray analysis (Jeremy et al. 2019; Pooch, Ballester, and Barros 2020; Rajpurkar et al. 2017; Seyyed-Kalantari et al. 2021b; Zhang et al. 2022). Since no prior benchmark existed for this combined dataset, we aggregated MIMIC-CXR and CheXpert to train a new image-based model. The ALL dataset introduced in (Seyyed-Kalantari et al. 2021a) comprises samples from MIMIC-CXR, CheXpert, and ChestX-ray14 (Xiaosong et al. 2017). However, ChestX-ray14 was part of the training data for the Google foundation model used to generate vector embeddings. Using this data for evaluation would introduce data leakage between the foundation model's training set and our test set. Therefore, we constructed a new ALL dataset using only MIMIC-CXR and CheXpert, and the model from (Seyyed-Kalantari et al. 2021a) was not suitable for our analysis.

Traning is caried out by spliting the dataset into into 80% training, 10% validation, and 10% testing, with early stopping applied if no improvement was observed for five consecutive epochs. The models are trained using the same hyperparameters and optimization strategy as previously described. To ensure reproducibility and robustness, we report the mean performance and 95% confidence interval across five independent runs with different random seeds. Model performance is measured using AUC, while fairness is assessed using TPR metrics.

Computational cost For the image-based pipeline (DenseNet-121, 224×224 RGB input), the forward pass costs approximately 2.83 GFLOPs. Including backward propagation ($\sim 3\times$ forward), a training sample costs ~ 8.49 GFLOPs. In contrast, the embedding-based model uses a lightweight MLP head ($1376 \to 768 \to 256 \to 14$), totaling ~ 1.68 MFLOPs per forward pass or ~ 5.04 MFLOPs per training sample. This leads to a $\sim 1{,}700\times$ difference in FLOPs per sample.

Fairness Evaluation

Fairness analysis requires a sensitive attribute S, such as race, gender, or age, that may cause biased outcomes. Here, we use sex, age, and race for CXP and ALL datasets, and additional attribute insurance type (similar to previous works (Seyyed-Kalantari et al. 2021a) as a proxy for socioeconomic status) in MIMIC. The protected groups are male and female; White, Black, Hispanic, Other, Asian, and Native for race; ages 0-20, 20-40, 40-60, 60-80, and 80+; and Medicare, Medicaid, and Other for insurance, with Medicaid often indicating low-income status. We use Equality of odds fairness notion (Hardt, Price, and Srebro 2016) which requires the orthogonality of the predicted label $\hat{Y}$ and the sensitive attribute S, $\hat{Y} \perp\!\!\!\perp S \mid Y$, given true label Y. Here, $Y, \hat{Y} \in \mathbb{R}^N$, N=14 is the number of disease labels, and their elements y_j, $\hat{y}_j \in \{0, 1\}$. We evaluate True Positive Rate (TPR) disparities across disease labels similar to image based model in (Seyyed-Kalantari et al. 2021a). For binary S (e.g sex) the TPR disparity for the *l*th is defined as (Seyyed-Kalantari et al. 2021a) (here, $S_{\neg l}$ means not belonging to S_l.):

$$TPRDisp_{S_l} = TPR_{S_l} - TPR_{S_{\neg l}} \quad (1)$$

For the non-binary S, the TPR disparity for the *l*th subpopulation (e.g Black in race)) we have (Seyyed-Kalantari et al. 2021a):

$$TPRDisp_{S_l} = TPR_{S_l} - \text{Median}\left(\{TPR_{S_k}\}_{k=1}^{l}\right) \quad (2)$$

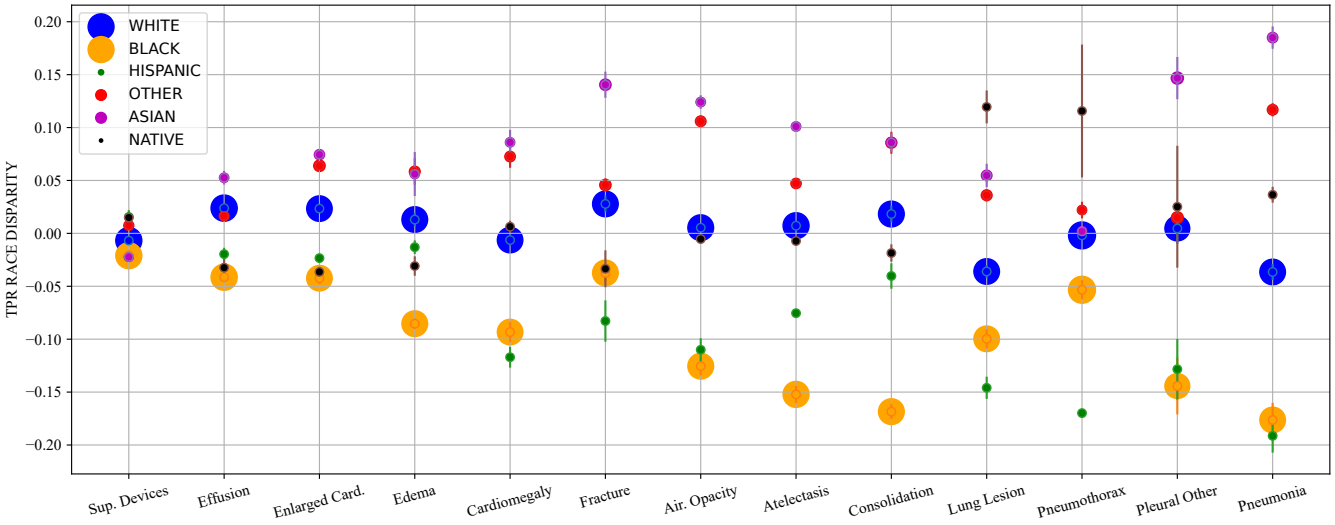

Figure 1: TPR race disparities of Emb-based model (mean across 5 runs ± 95% confidence interval, indicated by arrows) on the ALL dataset are shown along the y-axis, across disease labels on the x-axis. In the scatter plot, marker size represents the size of each subgroup for disease j. Positive $TPRDisp$ values indicate favorable outcomes, while negative values indicate unfavorable outcomes for a group. A lower Gap_j signifies a fairer model.

We compute $TPRDisp_{S_l}$ per disease label y_j. For a given y_j, S, the subgroup with maximum $TPRDisp_{S_l}$ is considered the most favorable as it demonstrates the largest disparity in its favor. The most unfavorable groups exhibit the highest negative gap. The TPR disparity gap per disease label j across subpopulations for a given S, is defined as :

$$Gap_j = \max\left(\{TPRDisp_{S_k}\}_{k=1}^{l}\right) - \min\left(\{TPRDisp_{S_k}\}_{k=1}^{l}\right). \quad (3)$$

We then calculate $\mathbb{E}[Gap_{i,j}]$, per S_i, across $\forall y_j$ and report it as the average Gap for a given sensitive attribute.

Results

Disease Classification Performance

Table 3 presents the disease classification AUC scores across 14 labels for the MIMIC, CXP, and ALL datasets, comparing both Emb-based and image-based models. For MIMIC and CXP, we use the results from Seyyed-Kalantari et al. (Seyyed-Kalantari et al. 2021a) as baselines, as that study has benchmarked its models against state-of-the-art methods. For the ALL dataset, we trained both image-based and Emb-based models in-house, using only CXP and MIMIC data. Unlike the ALL dataset in (Seyyed-Kalantari et al. 2021a), which includes Chest X-ray14 (Xiaosong et al. 2017), a dataset used to train the Google CXR Foundation Model (Sellergren et al. 2022), we exclude it from our ALL dataset to avoid overlap with the foundation model's training data.

The Google CXR Foundation model study (Sellergren et al. 2022) reports Emb-based results for five CheXpert (CXP) labels, evaluated on a small, non-public, hand-labeled test set of 234 images. In contrast, our evaluation spans all 14 disease labels using substantially larger test sets: 19,471 images for CXP, 21,591 for MIMIC, and 41,062 for the combined ALL dataset. For the five labels reported in (Sellergren et al. 2022), our AUC scores are generally comparable or higher, with the exception of Effusion, where our AUC is 0.03 lower. We present the mean and 95% confidence interval based on five runs using different random seeds.

Additionally, (Glocker et al. 2022) also investigated fairness in Emb-based models for four CXP labels using the statistic J, employing thresholds that enforce a fixed false-positive rate. Their results revealed unfairness in the Emb-based approach. It is important to recognize that the choice of threshold significantly influences fairness outcomes; a common strategy is to select thresholds that maximize the $F1$ score across all labels to balance precision and recall (Jeremy et al. 2019; Seyyed-Kalantari et al. 2021a), which we also applied in our analysis. To the best of our knowledge, no previous work has conducted a comprehensive comparison of Emb-based and Img-based models across all 14 labels, incorporating multiple Emb datasets, large and diverse test sets (to ensure generalizability), and using both threshold-independent (e.g., AUC) and threshold-dependent (e.g., correct disease diagnosis) evaluation metrics.

Fairness Analysis in Correct Disease Diagnosis

We present TPR disparity ($TPRDisp$) measurements across different subgroups and disease categories (excluding "No Finding"), where positive values reflect bias that benefits a particular subgroup and negative values reflect bias against it. The subgroup receiving the most favorable treatment demonstrates consistently positive $TPRDisp$ values across 13 diseases labels, whereas the most unfavoured group consistently shows negative $TPRDisp$ values. Lower $TPRDisp$ reflects

Attribute	Dataset	Average Gap	Cross-Label Gap Lowest	Highest	Unfavorable	Favorable
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 4: The $\mathbb{E}[Gap_j], \forall j$, values for image-based (Img) (Seyyed-Kalantari et al. 2021a) versus embedding-based (Emb) models are presented, with improved fairness in the Emb-based model highlighted in bold.

better fairness in disease diagnosis.

Figure 1 presents the race-based $TPRDisp$ with 95% confidence intervals, sorted by Gap_j for a model trained on the ALL dataset (as defined in (3)). Diseases are sorted from left to right by increasing disparity, those with the smallest gaps between racial subpopulations appear on the left (e.g., "Support Devices"), while those with the largest gaps (e.g., "Pneumonia") appear on the right. The average gap across all disease labels, $\mathbb{E}[Gap_{race,j}], \forall j$, is 0.214, with "Support Devices (SD)" showing the smallest gap of 0.037 and "Pneumonia (Pn)" the largest at 0.376. Patients identified as "Black" consistently experience negative TPR disparities in all 13 disease labels, indicating they are the most disadvantaged group, whereas patients categorized as "Other" consistently receive positive TPR disparities across all labels, making them the most favored group.

Table 4 provides a summary of fairness outcomes across Img-based and Emb-based models, evaluated over various demographic attributes. A lower average gap indicates improved fairness, meaning the performance difference between the best- and worst-performing groups is smaller across all disease labels. The results show that embedding-based models consistently exhibit lower average gaps (highlighted in bold), suggesting they achieve fairer outcomes compared to their image-based counterparts.

Embedding-based models more frequently demonstrate a lower average gap, indicating better fairness, compared to image-based models. However, the most favourable and unfavourable subgroups generally remain consistent across both model types. In other words, using vector embeddings does not change which groups are most vulnerable. Female patients, younger individuals, and Black patients are consistently among the most underdiagnosed or disadvantaged subpopulations (see Table 4). These findings align with previously identified vulnerable groups in healthcare (Abdelmalek et al. 2023; Walker et al. 2024; Parkhimchyk et al. 2024; Bahre et al. 2024) and medical imaging (Seyyed-Kalantari et al. 2021a,b), reflecting broader societal biases that need to be taken into account in machine learning in healthcare operations (Khattak et al. 2024) and properly communicated to the end users (Heming et al. 2023).

Generalizability to External Datasets

We used the embedding that came from Google foundation model(FM) in inference mode. Importantly, this model has never been exposed to MIMIC-CXR or CheXpert images during training. Also, Google FM has been trained on noisy high-level healthy vs. unhealthy labels and not the fine-grained disease-specific labels analyzed in our study. As such, the FM had no prior knowledge of the downstream tasks we evaluated.

Due to data sharing agreement, we are not allowed to upload X-rays in their party model (Google FM) to collect vector embedding and do this analysis on other dataset. Instead we had to focus on existing public Vector embedding data.

Conclusion

We conducted a comparative analysis of embedding-based and image-based models in terms of fairness and performance for disease diagnosis. The embeddings were obtained through *zero-shot inference* using Google's foundation model,

which had neither seen our target datasets nor been trained to detect our specific disease labels. Our results indicate that embedding-based models performed on par with or better than image-based models, while also demonstrating improved fairness. These findings suggest that embeddings can serve as effective alternatives to chest X-ray images, helping broaden access to AI tools for communities lacking high computational resources or expert clinicians to process complex medical images.

Acknowledgments

This work was supported in part by the Italian National Recovery and Resilience Plan (PNRR/NRRP) project FAIR–Future Artificial Intelligence Research (PE00000013), Spoke 9 "Green-aware Artificial Intelligence (AI)," funded by the Italian Ministry of University and Research (MUR) under the EU NextGenerationEU program, and from the projects Fa.Per.M.E. (CUP H53C22000640006) and NutriDieMMe (CUP H53C22000940001), both funded by the Italian Ministry of Health. The work of Leo Anthony Celi was supported in part by the National Institute of Health through DS-I Africa under Grant U54 TW012043-01 and Bridge2AI under Grant OT2OD032701, in part by the National Science Foundation through ITEST under Grant number 2148451, and in part by the Korea Health Technology Research and Development Project through the Korea Health Industry Development Institute (KHIDI) funded by the Ministry of Health and Welfare, Republic of Korea, under Grant RS-2024-00403047. The work of Laleh Seyyed-Kalantari was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant, and in part by the Connected Minds Canada First Research Excellence Fund (CFREF). The authors also express their gratitude to the Vector Institute for providing high-performance computing platforms. Francesco Calimeri is member of the Gruppo Nazionale Calcolo Scientifico-Istituto Nazionale di Alta Matematica (GNCS-INdAM).

References

Abdelmalek, F. M.; others. "Association between patient race; ethnicity; and use of invasive ventilation in the United States." Annals of the American Thoracic Society 21.2 (2024): 287-295. 2023. 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.

Ahluwalia, M.; Abdalla, M.; Sanayei, J.; Kalantari, L. S.; Hussain, M.; Ali, A.; and Fine, B. 2023. The Subgroup Imperative: Chest X-Ray Classifier Generalization Gaps in Patient, Setting, and Pathology Subgroups. *Radiology: Artificial Intelligence*.

Akbarian, S.; Seyyed-Kalantari, L.; Khalvati, F.; and Dolatabadi, E. 2023. Evaluating Knowledge Transfer in the Neural Network for Medical Images. *IEEE Access*, 11: 85812–85821.

Bahre, G.; et al. 2024. Fairness of AI Models in Vector Embedded Chest X-ray Representations. In *Advancements In Medical Foundation Model: Explainability, Robustness, Security, and Beyond, NeurIPS Workshop*.

Gichoya, J. W.; Banerjee, I.; Bhimireddy, A. R.; Burns, J. L.; Celi, L. A.; Chen, L.-C.; Correa, R.; Dullerud, N.; Ghassemi, M.; Huang, S.-C.; et al. 2022. AI recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, 4(6).

Gichoya, J. W.; Thomas, K.; Celi, L. A.; Safdar, N.; Banerjee, I.; Banja, J. D.; Seyyed-Kalantari, L.; Trivedi, H.; and Purkayastha, S. 2023. AI pitfalls and what not to do: mitigating bias in AI. *British Journal of Radiology*, 96(1150).

Glocker, B.; Jones, C.; Bernhardt, M.; and Winzeck, S. 2022. Risk of bias in chest x-ray foundation models. *arXiv preprint arXiv:2209.02965*.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. *NeurIPs*.

Heming, C. A. M.; et al. 2023. Benchmarking bias: Expanding clinical AI model card to incorporate bias reporting of social and non-social factors. arXiv:2311.12560.

Jeremy, J. A.; et al. 2019. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *AAAI*.

Johnson, A.; et al. 2019a. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data*.

Johnson, A. E. W.; Pollard, T. J.; Greenbaum, N. R.; Lungren, M. P.; ying Deng, C.; Peng, Y.; Lu, Z.; Mark, R. G.; Berkowitz, S. J.; and Horng, S. 2019b. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data*.

Khattak, F. K.; Subasri, V.; Krishnan, A.; Pou-Prom, C.; Akinli-Kocak, S.; Dolatabadi, E.; Pandya, D.; Seyyed-Kalantari, L.; and Rudzicz, F. 2024. MLHOps: Machine Learning Health Operations. *IEEE Access*, 4: 1–47.

Konate, S.; Lebrat, L.; Cruz, R. S.; Gichoya, J. W.; Price, B.; Seyyed-Kalantari, L.; Fookes, C.; Bradley, A.; and Salvado, O. 2025. Interpretability of AI race detection model in medical imaging with saliency methods. *Computational and Structural Biotechnology Journal (CSBJ)*.

Larrazabal, A. J.; Nieto, N.; Peterson, V.; Milone, D. H.; and Ferrante, E. 2020. Gender Imbalance in Medical Imaging Datasets Produces Biased Classifiers for Computer-Aided Diagnosis. *Proceedings of the National Academy of Sciences*, 117(23): 12592–12594.

Loshchilov, I.; and Hutter, F. 2016. SGDR: Stochastic Gradient Descent with Warm Restarts. *arXiv preprint arXiv:1608.03983*.

Marcinkevics, R.; Ozkan, E.; and Vogt, J. E. 2022. Debiasing Deep Chest X-ray Classifiers Using Intra-and Post-Processing Methods. In *Machine Learning for Healthcare Conference*, 504–536. PMLR.

Parkhimchyk, A.; et al. 2024. Exploring Visual Prompt Tuning for Demographic Adaptation in Foundation Models for Medical Imaging. In *Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning, NeurIPS workshop*.

Pooch, E. H. P.; Ballester, P.; and Barros, R. C. 2020. 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: Springer.

Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.; Shpanskaya, K.; Lungren, M. P.; and Ng, A. Y. 2017. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv preprint arXiv:1711.05225*.

Salvado, O.; Konate, S.; Cruz, R. S.; Bdadley, A.; Gichoya, J. W.; Seyyed-Kalantari, L.; Price, B.; Fookes, C.; and Lebrat, L. 2024. Localisation of Racial Information in Chest X-Ray for Deep Learning Diagnosis. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, 1–4.

Sellergren, A.; et al. 2022. Simplified Transfer Learning for Chest Radiography Models Using Less Data. *Radiology*, 305(2): 454–465.

Seyyed-Kalantari, L.; Liu, G.; McDermott, M. B. A.; and Ghassemi, M. 2021a. CheXclusion: Fairness gaps in deep chest X-ray classifiers. *Pacific Symposium on Biocomputing*, 26: 232–243.

Seyyed-Kalantari, L.; Zhang, H.; McDermott, M. B.; Chen, I. Y.; and Ghassemi, M. 2021b. Underdiagnosis Bias of Artificial Intelligence Algorithms Applied to Chest Radiographs in Under-served Patient Populations. *Nature Medicine*, 27(12): 2176–2182.

Walker, S. L.; et al. 2024. Association Between Sex and Race and Ethnicity and IV Sedation Use in Patients Receiving Invasive Ventilation. *CHEST Critical Care*, 2(4): 100100.

Xiaosong, W.; et al. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *IEEE CVPR*, 2097–2106.

Yang, Y.; Liu, Y.; Liu, X.; Gulhane, A.; Mastrodicasa, D.; Wu, W.; Wang, E. J.; Sahani, D. W.; and Patel, S. 2024. Demographic Bias of Expert-Level Vision-Language Foundation Models in Medical Imaging. *arXiv preprint arXiv:2402.14815*.

Zhang, H.; Dullerud, N.; Roth, K.; Oakden-Rayner, L.; Pfohl, S.; and Ghassemi, M. 2022. Improving the Fairness of Chest X-ray Classifiers. In *Conference on Health, Inference, and Learning*, 204–233. PMLR.