FairPneum: Improving Fairness in Pneumonia Diagnosis and Lesion Segmentation

Yuetan Chu[1⋆], Xinran Jin[1⋆], Xinghua Ma[1], Gongning Luo[1], and Xin Gao[1](✉)

King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia
xin.gao@kaust.edu.sa

Abstract. Machine learning-based computer-aided diagnostic systems often exhibit compromised fairness due to biases inherent in training datasets, leading to suboptimal and biased performance across diverse populations. These issues are particularly evident in pneumonia diagnosis and lesion segmentation, where disease prevalence and imaging characteristics vary significantly across demographic groups and pneumonia subtypes. To address these challenges, we propose FairPneum, a generalizable framework designed to enhance fairness in learning from biased datasets. FairPneum first introduces a hierarchical embedding disentanglement method that progressively isolates sensitive attributes using adversarial training, while an anti-bias training strategy is employed to leverage counterfactual embeddings to improve robustness against sensitive biases. We also introduce two novel metrics to effectively evaluate the global fairness in continuous performance metrics across different attributes while addressing the dependency between different sensitive attributes. Implementing FairPneum with both convolutional- and transformer-encoders, we demonstrate significant improvements in both performance and fairness on the public COVID-19 chest X-ray dataset and an in-house pneumonia CT dataset. Compared to SOTA fairness learning approaches, FairPneum effectively reduces performance disparities across demographic groups while maintaining or enhancing overall model accuracy. This study highlights a promising approach for developing unbiased and clinically generalizable solutions for pneumonia diagnosis and segmentation. The source code and dataset are available at https://github.com/JXXrrrrR/FairPneum.

Keywords: Fairness study · Pneumonia · Segmentation · Classification

1 Introduction

Over the past decade, machine learning approaches for computer-aided diagnosis (CAD) have been increasingly integrated into clinical practice, offering significant benefits for both patients and healthcare professionals [1]. However, these approaches are often compromised by inherent biases within medical datasets,

⋆ These authors contributed equally to this work.

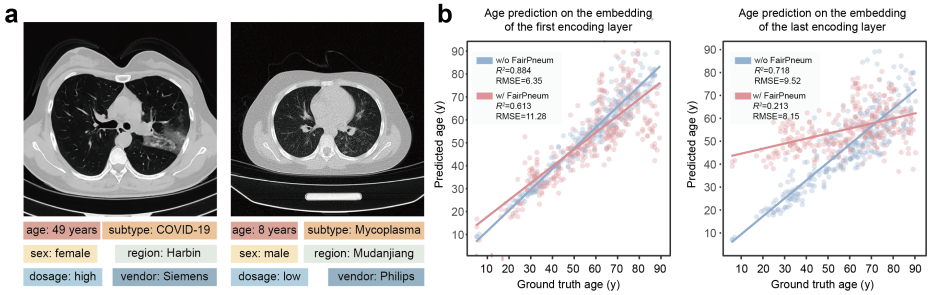

Fig. 1. The motivation and embedding disentanglement evaluation of FairPneum. a Representative samples of pneumonia containing sensitive attributes. **b** The age prediction results on the embeddings of the first and last encoding layer. Blue and red represent results with and without FairPneum. Without FairPneum, embeddings will encode age information. However, when integrating FairPneum, embeddings contain significantly less age information.

which arise from demographic imbalance and acquisition heterogeneity [2]. Such biases can cause systematic performance disparities across subpopulations, leading to unfair predictions and limited generalizability in clinical deployment. Fairness learning has therefore become an essential component in medical AI research, aiming to ensure equitable model performance across diverse demographic and clinical subgroups.

Motivated by these challenges, this work investigates fairness learning for pneumonia diagnosis and lesion segmentation. Accurate subtype identification enables timely intervention, and precise segmentation supports quantitative assessment and longitudinal monitoring [9]. However, radiological manifestations vary substantially across demographic groups [10]. For instance, *Mycoplasma pneumoniae* primarily affects pediatric patients, whereas COVID-19 pneumonia is more prevalent in older adults with distinct imaging patterns [11]. Models trained on dominant subtypes may therefore overfit to characteristic appearances and underperform on underrepresented groups. Such biological and demographic coupling makes pneumonia segmentation highly sensitive to dataset imbalance, potentially biasing severity estimation and clinical decisions. Ensuring fairness is therefore both a methodological necessity and a prerequisite for equitable clinical deployment.

Traditional fairness learning in natural image analysis typically focuses on relatively disentangled attributes, such as sex, illumination, or object category [12], which can be explicitly separated by neural networks [13]. In contrast, radiological images encode tightly coupled demographic and physiological factors: attributes such as age, region, and disease subtype influence both imaging appearance and disease manifestation. Such entanglement makes sensitive information difficult to remove through single-layer adversarial training, as biased signals may persist across hierarchical encoding stages [14] (Fig. 1.b). Consequently, existing disentanglement or adversarial approaches [12,15,28] are often insuffi-

cient for medical imaging. Moreover, although fairness learning is well studied in classification, its extension to segmentation remains limited [16], and current fairness metrics are largely designed for discrete outcomes rather than continuous measures such as Dice similarity. These limitations motivate fairness frameworks specifically tailored to medical imaging and pneumonia-related tasks.

To address the aforementioned challenges, we propose **FairPneum**, a comprehensive fairness learning framework for unbiased pneumonia diagnosis and lesion segmentation. The contributions are summarized as follows:

- **Hierarchical embedding disentanglement.** We design a multi-level adversarial module that progressively separates sensitive attributes (e.g., age, sex, subtype) from image representations, mitigating bias embedded across feature scales.
- **Anti-bias re-entanglement.** We propose a counterfactual re-entanglement strategy that combines debiased and sensitive embeddings, encouraging attribute-invariant yet pathology-relevant representations.
- **Fairness metrics for continuous outcomes.** We introduce *Local Performance Equality (LPE)* and *Global Performance Equality (GPE)* to evaluate fairness in continuous metrics (e.g., DSC), extending fairness assessment to segmentation tasks.
- **Establishment of a multi-subtype pneumonia dataset.** We establish an in-house chest CT dataset with 4,817 slices from 315 studies covering COVID-19, bacterial, and *Mycoplasma* pneumonia, providing a benchmark for segmentation fairness evaluation across demographics and subtypes.

Comprehensive experiments across diverse datasets demonstrate that FairPneum consistently improves fairness across demographic and pathological subgroups while maintaining competitive or superior overall performance compared with state-of-the-art (SOTA) methods.

2 Method

2.1 Algorithm Overview

As illustrated in Fig. 2, FairPneum augments a hierarchical encoder with stage-wise fairness-aware modules that progressively disentangle and re-entangle feature representations after each encoding stage. After each encoding layer, two plug-and-play components are inserted: a *hierarchical embedding disentanglement module* and a *counterfactual re-entanglement module*. The disentanglement module consists of an embedding decoupling block $\mathcal{B}$ and a sensitive attribute discriminator $\mathcal{D}$, where $\mathcal{B}$ separates intermediate features into attribute-sensitive and task-relevant subspaces, and $\mathcal{D}$ adversarially suppresses sensitive information. The re-entanglement module employs a reconstruction block $\mathcal{R}$ (Sec. 2.3) to recombine debiased embeddings and generate counterfactual variants during training, encouraging attribute-invariant yet pathology-preserving representations. The framework integrates seamlessly into existing encoders without altering the backbone or task-specific loss.

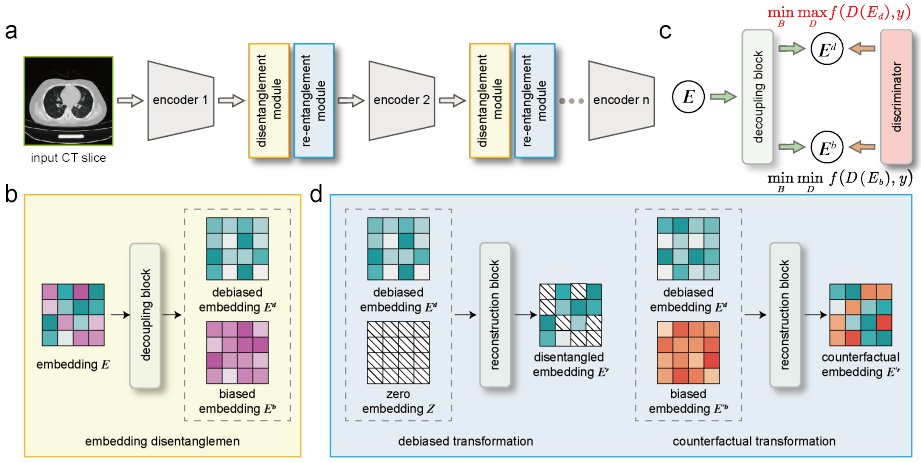

Fig. 2. Overview of FairPneum approach. **a** FairPneum can be integrated with any hierarchical encoder by inserting a paired embedding disentanglement and re-entanglement module after each encoding layer. **b** The embedding decoupling block separates the embedding into biased embedding E^b and debiased embedding E^d, with and without encoding biased attribute information, respectively. **c** The schematic of adversarial training of the decoupling block and discriminator. The red and green arrows represent the positive and adversarial training, respectively. **d** We employ two anti-bias training strategies, debiased and counterfactual transformation for the disentangled embedding. E'^b is the biased embedding derived from another case.

2.2 Hierarchical Embedding Disentanglement via Adversarial Training

To separate sensitive attributes from task-relevant representations, we introduce a hierarchical embedding disentanglement mechanism. Given an input image x, the hierarchical encoder progressively extracts multi-scale features for classification and segmentation tasks [18]. For an intermediate embedding $E \in R^{C\times H\times W}$ produced by one encoding layer, the representation may contain both desired task information and sensitive attributes.

To decompose E, we employ an embedding decoupling block $\mathcal{B}$ that splits it into two sub-embeddings, E^d and E^b, each sharing the same dimensionality as E (Fig. 2(b)):

$$\mathcal{B}(E) \to (E^d, E^b). \tag{1}$$

Here, E^d represents the debiased embedding with minimal sensitive information, while E^b retains the attribute-related components. The decoupling block consists of two 3×3 convolutional layers followed by a GELU activation [19], where the first convolution preserves the channel size ($C \to C$) and the second expands it ($C \to 2C$). The first C channels of the output correspond to E^d, and the remaining C channels to E^b.

To guide this separation, we introduce a lightweight attribute discriminator $\mathcal{D}$ [20] trained to predict sensitive attributes y, which form a multi-label vec-

tor representing all studied factors, each discretized into categorical values. The adversarial training follows a min–max scheme, where $\mathcal{D}$ and $\mathcal{B}$ are optimized alternately without gradient sharing. The discriminator minimizes its classification loss:

$$L_{\mathcal{D}} = f(\mathcal{D}(E^d), y) + f(\mathcal{D}(E^b), y), \tag{2}$$

where $f(\cdot)$ denotes the focal loss (FL) [21]. In contrast, $\mathcal{B}$ is optimized to confuse $\mathcal{D}$ on E^d while allowing sensitive information in E^b:

$$L_{\mathcal{B}} = -f(\mathcal{D}(E^d), y) + f(\mathcal{D}(E^b), y). \tag{3}$$

Through this adversarial interplay, the decoupling block learns to generate attribute-invariant embeddings E^d that preserve task-relevant features while isolating the attribute-specific components E^b. The disentangled pair (E^d, E^b) is subsequently utilized in the re-entanglement stage to synthesize counterfactual representations (Sec. 2.3).

2.3 Debiased Embedding Re-entanglement via Anti-bias Training

After disentanglement, a re-entanglement stage restores representation completeness while suppressing bias propagation, encouraging the decoder to rely on unbiased features rather than sensitive attributes during training. In the initial anti-bias stage, we apply a debiased transformation by replacing the biased embedding E^b with a zero-tensor Z of identical shape, simulating the removal of sensitive components. A reconstruction block $\mathcal{R}$ then recombines the debiased and zero embeddings into an integrated representation:

$$E^r = \mathcal{R}(E^d \oplus Z), \tag{4}$$

where $\oplus$ denotes channel-wise concatenation (Fig. 2(d)). The reconstruction block consists of two 3×3 convolutional layers and a GELU activation [19], with the first convolution maintaining channels ($2C \to 2C$) and the second compressing them ($2C \to C$). Each encoder layer contains a corresponding lightweight $\mathcal{R}$ block with the same architecture, and these blocks are only activated during the anti-bias stage. During inference, all $\mathcal{R}$ modules are bypassed.

The reconstructed embedding E^r replaces the original embedding E and is passed to the next encoding layer. For networks employing skip connections [22], E^r rather than E is concatenated with the corresponding decoder features.

To further enhance decoder robustness, we introduce a counterfactual transformation strategy. For each training batch, 50% of samples are randomly selected for counterfactual transformation. Each selected sample is paired with at most one image x' from the same task class in the same mini-batch, where at least one sensitive attribute differs. If no eligible partner exists in the mini-batch, the sample is not swapped, and only the debiased transformation is applied. Given the disentangled embeddings (E^d, E^b) for x and (E'^d, E'^b) for x', we construct counterfactual representations by replacing the biased component:

$$E'^r = \mathcal{R}(E^d \oplus E'^b). \tag{5}$$

These counterfactual embeddings are used only during training to regularize the decoder, promoting attribute-invariant yet task-preserving representations. The reconstruction block $\mathcal{R}$ is jointly optimized with the overall segmentation/classification objective, without any additional reconstruction loss.

2.4 Fairness Evaluation Metrics

Conventional group fairness metrics measure the balance of outcomes between groups with different sensitive attributes [23]. For instance, demographic parity (DP) is defined as $\mathrm{DP}_{i,j} = \frac{p(Y=1 \mid a=a_i)}{p(Y=1 \mid a=a_j)}$, where $a_i, a_j \in A$ denote the sensitive attributes and Y represents the prediction label. However, such definitions assume discrete outcomes and are not applicable to continuous values, such as the Dice Similarity Coefficient (DSC) in segmentation tasks.

To address this limitation, we propose a *Local Performance Equality* (LPE) metric that extends DP to continuous outcomes by comparing the distributions of performance scores across different attribute groups. The distributions $p(Y \mid a)$ are estimated from per-sample DSC values using kernel density estimation (KDE), and the similarity between groups is quantified using the first-order Wasserstein distance (Earth Mover's Distance):

$$\mathrm{LPE}_{i,j}^{(A)} = \mathcal{W}(p(Y \mid a = a_i),\, p(Y \mid a = a_j))\,, \tag{6}$$

where $\mathcal{W}$ is computed symmetrically and normalized by the support range of Y. A smaller LPE indicates higher parity between the two sensitive groups.

While LPE captures pairwise fairness, it does not account for correlations among multiple sensitive attributes. For instance, children may show a higher incidence of mycoplasma pneumonia and also receive lower radiation doses, causing cross-attribute bias. To jointly evaluate fairness across correlated attributes, we introduce the *Global Performance Equality* (GPE) metric, which aggregates weighted LPE terms:

$$\mathrm{GPE} = \sum_{A \in \mathcal{K}} \frac{|\rho_{A,Y}|\,(1 - p_{A,Y}) + \varepsilon}{\sum_{A \in \mathcal{K}} \left(|\rho_{A,Y}|\,(1 - p_{A,Y})\right) + \varepsilon} \sum_{a_i, a_j \in A} \mathrm{LPE}_{i,j}^{(A)}, \tag{7}$$

where $\rho_{A,Y}$ and $p_{A,Y}$ denote the Pearson correlation coefficient and its significance level between attribute group A and outcome Y, $\mathcal{K}$ is the set of all attribute groups, and $\varepsilon = 0.001$ ensures numerical stability. The weighting term $|\rho_{A,Y}|\,(1-p_{A,Y})$ emphasizes fairness evaluation for attributes strongly correlated with the outcome. A lower GPE value indicates smaller cross-group disparity and thus a fairer model.

Specifically, GPE is intended for settings where sensitive attributes exhibit non-negligible dependency with the performance outcome or with each other. When all such dependencies are uniformly weak, LPE should be interpreted as the primary fairness measure, because the contribution of cross-attribute dependency becomes negligible.

a

MIDRC		
Age	< 75 y	65,542 (84.15%)
	⩾ 75 y	12,345 (15.85%)
Sex	male	43,880 (56.33%)
	female	34,007 (43.67%)

b

In-house dataset (training&validation)		
Age	< 18 y	102 (39.53%)
	⩾ 18 y	156 (60.47%)
Sex	male	132 (51.16%)
	female	126 (48.84%)
Region	Harbin	187 (72.48%)
	Mudanjiang	71 (27.52%)
Subtypes	COVID-19	96 (37.21%)
	bacterial	85 (32.95%)
	mycoplasma	77 (29.84%)

c

In-house Dataset (testing)		
Age	< 18 y	26 (45.61%)
	⩾ 18 y	31 (54.39%)
Sex	male	25 (43.86%)
	female	32 (56.14%)
Region	Harbin	35 (61.40%)
	Mudanjiang	22 (38.60%)
Subtypes	COVID-19	16 (28.07%)
	bacterial	19 (33.33%)
	mycoplasma	22 (38.60%)

Fig. 3. Distribution of imbalanced attributes. **a**, **b**, and **c** illustrate the imbalanced attributes within the publicly available MIDRC dataset, and our in-house training&validation and testing sets.

3 Experiments and Results

Methods	ACC (%) for <75 y ↑	ACC (%) for >=75 y ↑	DP for age ↑	ACC (%) for male ↑	ACC (%) for female ↑	DP for sex ↑	overall ACC (%) ↑
SwinT	78.34	72.23	0.922	77.02	77.91	0.987	77.41
SwinT + DFA	75.54	72.87	0.966	74.70	75.68	0.986	75.13
SwinT + PnD	77.54	73.67	0.951	76.68	77.30	0.991	76.95
SwinT + Lin et al.	76.35	72.97	0.957	75.46	76.33	0.990	75.84
SwinT + FairPneum-	76.87	73.14	0.954	76.52	76.02	0.993	76.30
SwinT + FairPneum	77.58	74.97	0.966	77.24	77.10	0.998	77.18
MedNext	84.70	76.23	0.900	83.02	83.91	0.989	83.41
MedNext + DFA	82.07	75.87	0.923	80.70	81.68	0.989	81.13
MedNext + PnD	83.72	78.67	0.938	82.78	83.17	0.994	82.95
MedNext + Lin et al.	82.53	77.97	0.948	81.36	82.46	0.985	81.84
MedNext + FairPneum-	83.05	78.14	0.942	82.52	82.02	0.994	82.30
MedNext + FairPneum	83.87	79.97	0.953	83.14	83.46	0.996	83.28

Table 1. Classification performance and fairness evaluation on the MIDRC dataset. The baseline performance (without fairness learning) is highlighted in blue. For each metric, the best result among fairness-learning methods is highlighted in red.

This study used one publicly available MIDRC dataset[25] and one in-house dataset to evaluate the proposed FairPneum framework. The MIDRC dataset includes 77,887 chest radiographs from 62,219 studies[2], with self-reported demographic information. Age and sex were selected as sensitive attributes (Fig. 3.a), with age dichotomized at 75 years to reflect the common clinical definition of elderly patients and the higher-risk subgroup in COVID-19 cohorts. The in-house dataset comprises 315 CT studies with 4,817 axial slices for lesion segmentation, split into a training–validation set (n = 258) and an independent testing set (n = 57) (Fig. 3.b, c). Four sensitive attributes were examined: age, sex, geographic region, and pneumonia subtype. Age was dichotomized at 18 years, corresponding to the pediatric–adult clinical boundary and the pronounced subtype-age coupling in the training–validation set, where Mycoplasma pneumonia cases predominantly occurred in patients under 18; the testing set showed a more balanced age distribution. Lesion annotations for pneumonia-positive slices were performed by a radiologist with over 10 years of clinical experience.

Fairness learning was evaluated using two representative architectures, Swin-Transformer (SwinT) and MedNeXt[26]. GCE loss[27] and Dice loss were used for classification and segmentation, respectively, with training without fairness intervention serving as the baseline. FairPneum was benchmarked against several SOTA fairness methods: DFA[28], PnD[29], and Lin et al.[2] for classification, and ADV[30] and FEBS[3] for segmentation. To avoid confounding effects from backbone capacity, all fairness methods were implemented on the same task back-

bone and trained with their original fairness objectives or representation-learning strategies. Hierarchical disentanglement and counterfactual re-entanglement were used only in FairPneum, enabling comparison under a shared-backbone setting that isolates the effect of the proposed fairness strategy.

Performance was evaluated using the proposed fairness metrics together with accuracy (ACC), DSC, and DP to assess both task performance and fairness. An ablated variant, FairPneum^{-}, which removes counterfactual transformation during training, was included to analyze its contribution. Segmentation performance significance was tested using the Wilcoxon signed-rank test, with $p < 0.05$ deemed significant. All experiments were conducted on a Linux workstation with an NVIDIA A6000 GPU. Models were optimized using AdamW ($\beta_1 = 0.9, \beta_2 = 0.999$). Input images were normalized and resized to 512×512, with a training batch size of 16.

3.1 Performance and Fairness Evaluation

Methods	DSC (%) for <18 y ↑	DSC (%) for >=18 y ↑	LPE for age ↓	DSC (%) for COVID-19 ↑	DSC (%) for Bacterial ↑	DSC (%) for Mycoplasma ↑	LPE for subtype ↓	overall DSC (%) ↑	GPE ↓
SwinT	70.48	77.01	6.54	78.71	76.70	68.32	7.19	74.03	6.84
SwinT + ADV	72.52	77.25*	4.71	79.11*	77.12	70.41	6.12	75.05	5.36
SwinT + FEBS	72.20	76.05	3.84	77.06	76.35	70.50	4.81	74.29	3.95
SwinT + FairPneum-	72.07	76.74	4.66	76.68	75.83	72.05	3.61	74.61	3.89
SwinT + FairPneum	73.73*	76.76	3.07	76.91	76.95	72.91	3.46	75.38*	3.09
MedNext	75.62	81.74	6.13	83.54	81.34	73.55	7.02	78.95	6.10
MedNext + ADV	77.88	81.39	3.67	84.62*	81.14	75.11	6.68	79.59	4.87
MedNext + FEBS	77.04	81.00	4.03	81.60	81.60	75.37	4.94	79.19	3.87
MedNext + FairPneum-	76.90	80.99	4.10	82.10	80.01	76.19	4.36	79.12	3.40
MedNext + FairPneum	78.58*	81.61	3.13	81.92	81.74	77.69	3.41	80.23*	3.11

Methods	DSC (%) for male ↑	DSC (%) for female ↑	LPE for sex ↓	DSC (%) for Harbin ↑	DSC (%) for Mudanjiang ↑	LPE for region ↓
SwinT	72.92	74.90	1.98	74.79	72.82	2.03
SwinT + ADV	73.87	76.04	2.21	75.89	73.81	2.69
SwinT + FEBS	72.98	75.32	2.63	74.37	74.17	1.44
SwinT + FairPneum-	74.23	74.91	1.00	74.53	74.74	1.75
SwinT + FairPneum	75.48*	75.30	0.84	75.21	75.64*	1.75
MedNext	77.28	80.25	2.95	79.24	78.48	1.48
MedNext + ADV	78.30	80.96	2.66	80.44	78.75	1.89
MedNext + FEBS	77.33	80.65	3.30	79.52	78.68	2.33
MedNext + FairPneum-	78.27	79.79	1.75	78.75	79.71	1.86
MedNext + FairPneum	80.34*	80.15	1.53	80.52	79.77	1.73

Table 2. Segmentation performance and fairness evaluation. The baseline performance (without fairness learning) is highlighted in blue. For each metric, the best result among fairness-learning methods is highlighted in red. The best performance with statistical significance is marked in *.

Tables 1 and 2 show the results for classification and segmentation. Without fairness learning, clear gaps appear across groups (ACC difference 6–8%; DSC difference 4–8%), reflecting the impact of age, sex, and subtype imbalance. After applying fairness methods, these gaps consistently decrease. FairPneum yields the largest gains across both backbones, improving the inferior group by 2–3% in ACC and 3–5% in DSC. It also achieves the highest overall ACC/DSC among fairness approaches and remains comparable to or better than the baseline. Regarding fairness-specific measures, FairPneum achieves the best DP for age/sex, the best LPE across age and subtype, and the lowest GPE, indicating the most balanced performance. These results indicate that FairPneum effectively reduces group disparity while preserving, and in several cases slightly improving, the overall predictive performance.

4 Conclusion

In this work, we investigated fairness learning for both classification and lesion segmentation in pneumonia-related tasks. We introduced FairPneum, a framework that integrates hierarchical embedding disentanglement with anti-bias training to mitigate attribute-related biases within feature representations. In addition, we proposed two new metrics enabling global assessment of fairness for continuous-value performance measures. Extensive experiments across datasets, modalities, and network backbones demonstrate that FairPneum consistently enhances fairness while preserving overall performance. These findings highlight a promising direction toward more equitable and clinically reliable diagnostic and segmentation models for pneumonia.

Acknowledgments. This work was supported by KAUST Office of ORA under Awards RGC/3/6655-01-01, RGC/3/6208-01-01, RGC/3/5679-01-01, REI/1/5289-01-01, REI/1/5992-01-01, URF/1/6713-01-01, FCC/1/5932-12-11, URF/1/6599-01-02, RGC/3/6295-01-01, URF/1/6993-01-01, KCSH Award 5932, and Center of Excellence on Generative AI Award 5940.

Disclosure of Interests. The authors declare no competing interests.

References

1. Mhasawade, V., Zhao, Y., Chunara, R.: Machine learning and algorithmic fairness in public and population health. Nature Machine Intelligence 3(8), 659–666 (2021)
2. Lin, M., Li, T., Yang, Y., et al.: Improving model fairness in image-based computer-aided diagnosis. Nature Communications 14(1), 6261 (2023)
3. Tian, Y., Shi, M., Luo, Y., et al.: FairSeg: A large-scale medical image segmentation dataset for fairness learning with fair error-bound scaling. In: The Twelfth International Conference on Learning Representations (2023)
4. Yang, F., Servadio, J.L., Le Thanh, N.T., et al.: A combination of annual and nonannual forces drive respiratory disease in the tropics. BMJ Global Health 8(11), e013054 (2023)
5. Tan, H.H.G., Westeneng, H.J., Nitert, A.D., et al.: MRI clustering reveals three ALS subtypes with unique neurodegeneration patterns. Annals of Neurology 92(6), 1030–1045 (2022)
6. Kalra, M.K., Rizzo, S., Maher, M.M., et al.: Chest CT performed with z-axis modulation: scanning protocol and radiation dose. Radiology 237(1), 303–308 (2005)
7. Brown, A., Tomasev, N., Freyberg, J., et al.: Detecting shortcut learning for fair medical AI using shortcut testing. Nature Communications 14(1), 4314 (2023)
8. Luo, L., Xu, D., Chen, H., et al.: Pseudo bias-balanced learning for debiased chest x-ray classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 621–631. Springer Nature Switzerland, Cham (2022)
9. Chu, Y., et al.: Point-annotation supervision for robust 3D pulmonary infection segmentation by CT-based cascading deep learning. Computers in Biology and Medicine 187, 109760 (2025)

10. Youn, Y.S., Lee, K.Y.: Mycoplasma pneumoniae pneumonia in children. Korean Journal of Pediatrics 55(2), 42 (2012)
11. Li, K., Wu, J., Wu, F., et al.: The clinical and chest CT features associated with severe and critical COVID-19 pneumonia. Investigative Radiology (2020)
12. Sarhan, M.H., Navab, N., Eslami, A., et al.: On the fairness of privacy-preserving representations in medical applications. In: MICCAI Workshop on Domain Adaptation and Representation Transfer, pp. 140–149. Springer International Publishing, Cham (2020)
13. Shrestha, R., Kafle, K., Kanan, C.: OccamNets: Mitigating dataset bias by favoring simpler hypotheses. In: European Conference on Computer Vision, pp. 702–721. Springer Nature Switzerland, Cham (2022)
14. Zeng, S., et al.: JanusVLN: Decoupling semantics and spatiality with dual implicit memory for vision-language navigation. arXiv preprint arXiv:2509.22548 (2025)
15. Sarhan, M.H., et al.: Fairness by learning orthogonal disentangled representations. In: European Conference on Computer Vision, pp. 746–761. Springer International Publishing, Cham (2020)
16. Schaefer, G., Rajab, M.I., Celebi, M.E., et al.: Colour and contrast enhancement for improved skin lesion segmentation. Computerized Medical Imaging and Graphics 35(2), 99–104 (2011)
17. Xiao, C., Ke, Z., Zhao, P.: Focus where it matters: LLM-guided attention priors for few-shot segmentation. SSRN preprint 6281833
18. Zhang, H., Dana, K., Shi, J., et al.: Context encoding for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7151–7160 (2018)
19. Hendrycks, D., Gimpel, K.: Gaussian error linear units. arXiv preprint arXiv:1606.08415 (2016)
20. Wang, F., Jiang, M., Qian, C., et al.: Residual attention network for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164 (2017)
21. Lin, T.Y., Goyal, P., Girshick, R., et al.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
22. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241. Springer International Publishing, Cham (2015)
23. Du, M., Mukherjee, S., Wang, G., et al.: Fairness via representation neutralization. Advances in Neural Information Processing Systems 34, 12091–12103 (2021)
24. Tranmer, M., Elliot, M.: Multiple linear regression. The Cathie Marsh Centre for Census and Survey Research 5(5), 1–5 (2008)
25. Lakhani, P., Mongan, J., Singhal, C., et al.: The 2021 SIIM-FISABIO-RSNA machine learning COVID-19 challenge: Annotation and standard exam classification of COVID-19 chest radiographs. Journal of Digital Imaging 36(1), 365–372 (2023)
26. Roy, S., Koehler, G., Ulrich, C., et al.: MedNeXt: Transformer-driven scaling of ConvNets for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 405–415. Springer Nature Switzerland, Cham (2023)
27. Zhang, Z., Sabuncu, M.: Generalized cross entropy loss for training deep neural networks with noisy labels. Advances in Neural Information Processing Systems 31 (2018)

28. Lee, J., Kim, E., Lee, J., et al.: Learning debiased representation via disentangled feature augmentation. Advances in Neural Information Processing Systems 34, 25123–25133 (2021)
29. Li, J., Vo, D.M., Nakayama, H.: Partition-and-debias: Agnostic biases mitigation via a mixture of biases-specific experts. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4924–4934 (2023)
30. Madras, D., Creager, E., Pitassi, T., et al.: Learning adversarially fair and transferable representations. In: International Conference on Machine Learning, pp. 3384–3393. PMLR (2018)