On the Interplay of Human-AI Alignment, Fairness, and Performance Trade-offs in Medical Imaging

Haozhe Luo[1,3], Ziyu Zhou[2], Shelley Zixin Shu[1], Aurélie Pahud de Mortanges[1], Robert Berke[3], and Mauricio Reyes[1,4]

[1] ARTORG Center for Biomedical Engineering Research, University of Bern
{haozhe.luo,mauricio.reyes3}@unibe.ch
[2] Shanghai Jiao Tong University
[3] Kaiko.AI, Zurich, Switzerland
[4] Department of Radiation Oncology, Inselspital, Bern University Hospital and University of Bern, Bern, Switzerland

Abstract. Deep neural networks excel in medical imaging but remain prone to biases, leading to fairness gaps across demographic groups. We provide the first systematic exploration of Human-AI alignment and fairness in this domain. Our results show that incorporating human insights consistently reduces fairness gaps and enhances out-of-domain generalization, though excessive alignment can introduce performance trade-offs, emphasizing the need for calibrated strategies. These findings highlight Human-AI alignment as a promising approach for developing fair, robust, and generalizable medical AI systems, striking a balance between expert guidance and automated efficiency. The code is available at https://github.com/Roypic/Aligner.

Keywords: Fairness · Human-AI Alignment · Vision Language Model.

1 Introduction

Deep neural networks have become indispensable for a wide range of medical image computing applications. Nevertheless, their data-driven nature renders them susceptible to learning spurious correlations and biases [8,3,9,27,34]. This susceptibility not only undermines robustness and generalization, especially under out-of-distribution (OOD) conditions, but also raises concerns about the fairness of these systems across diverse patient populations. Human-AI alignment has recently emerged as a promising avenue for mitigating such issues by directing the learned representations toward human-centric knowledge. Although existing work on Human-AI alignment - also referred to as Explanation-Guided Learning (EGL) [5,6,7,23,31,23,28,19,20] - has shown improved robustness and performance, its relationship with model fairness remains largely unexplored. In medical imaging, unfairness often manifests as systematic performance disparities across demographic subgroups (e.g., sex, race, age), stemming from biases in

training data and inconsistencies in annotation practices, among other factors. For instance, several studies have revealed fairness gaps in chest X-ray classifiers [25,26], racial disparities in brain image analysis [29,14], and gender imbalances yielding skewed diagnostic outcomes [16]. Other lines of research highlight unfairness resulting from socioeconomic biases [22] or presentation and annotation disparities [10,32]. Given the potential of Human-AI alignment to mitigate these issues, its impact on reducing fairness gaps warrants deeper investigation. In this paper, we investigate the interplay between Human-AI alignment and model fairness, a relationship that remains largely unexplored. Specifically, we ask: *"Does Human-AI alignment contribute to reducing disparities in trained models?"*. To this end, we design a study on disease classification from chest X-ray images, a commonly benchmarked task for fairness research for which associated fairness variables are available. We systematically analyze fairness with respect to two subgroups (gender and age), using multiple group fairness metrics. Our experiments are conducted on Vision Transformer (ViT) under various degrees of human-AI alignment (including deliberate misalignment). Our findings demonstrate that Human-AI alignment consistently reduces fairness gaps across diseases and demographic subgroups while also enhancing out-of-domain generalization. This supports recent studies [29,14,32] suggesting that mitigating spurious correlations can improve real-world performance, challenging the notion that fairness interventions necessarily degrade model accuracy. However, we also find that excessive or misguided alignment can introduce trade-offs, emphasizing the need for carefully calibrated strategies. To the best of our knowledge, this is the first systematic study of Human-AI alignment's impact on fairness in medical imaging, highlighting its potential to develop fair, robust, and generalizable AI models.

2 Methods

2.1 Experimental Design to Assess Impact of Human-AI and Fairness

Figure 1 summarizes our study design, describing the multi-center training and out-of-domain (OOD) data used in our experiments, comprising fairness attributes (sex and age), different levels of human-AI alignment (including a randomized alignment ablation), and evaluation metrics for fairness, and performance (including a subanalysis at different regimes of training data).

Training and OOD Data: We selected multiple publicly available chest X-ray datasets to train classification models for detecting (i) nodules and masses, (ii) pleural effusion, and (iii) edema. These conditions were chosen for their clinical relevance and their distinct semantic characteristics, e.g., spatial location is a key factor for nodules, whereas texture plays a crucial role in identifying pleural effusion. These training datasets come equipped with expert-based annotations reflecting human-based attention areas a radiologist uses for diagnosis. These areas were used to guide the learning process, as detailed in section 2.2. Table 1

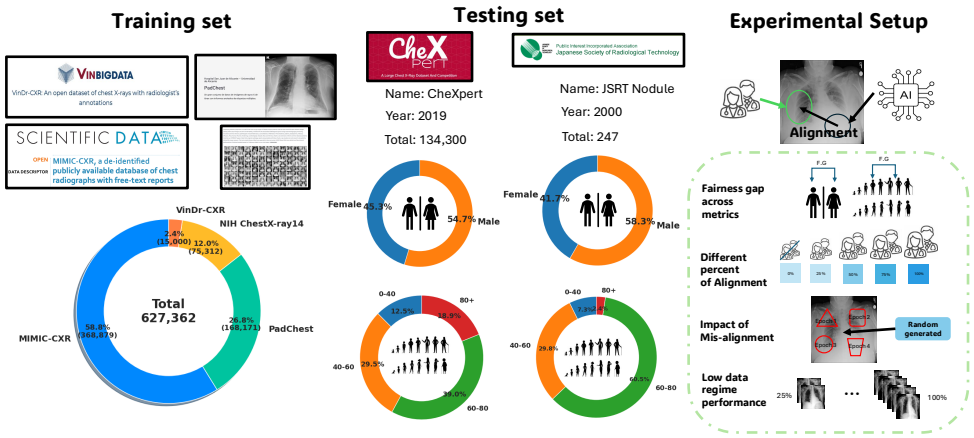

Fig. 1: Experimental setup to assess the impact of Human-AI alignment on fairness. Models trained on multicenter training datasets are trained without Human-AI alignment (i.e., baseline), and with various degrees of Human-AI alignment, and their fairness gap and classification performance metrics are assessed across two demographic groups on out-of-domain datasets across three different classification tasks. Additionally, the impact of Human-AI attention on low-data regimes and when alignment is randomized are further evaluated.

provides details on the datasets used for training and OOD evaluation, including NIH ChestX-Ray14 [30], MIMIC-CXR [13], VinDr-CXR [21], PadChest [2], CheXpert [12], and CheXlocalize [24]. These datasets ensure a diverse and comprehensive representation of different pathologies, with a total of 627, 362 cases used for training and 134, 547 for OOD testing.

Varying degrees of Human-AI alignment: We investigated the impact of Human-AI alignment on fairness by systematically varying the degree of Human guidance. Specifically, we considered five levels: *Level-1*: **0% (No Alignment):** A fully data-driven approach where **no Human-AI alignment** is conducted. *Level-2*: **25% (Weak Alignment):** Human-labeled data constitutes 25% of the total dataset used for standard training. *Level-3*: **50% (Moderate Alignment):** Human guidance is incorporated at a level equal to 50% of the dataset used for standard training. *Level-4*: **75% (Strong Alignment):** A predominantly Human-aligned setting where 75% of the dataset used for standard training consists of Human-labeled data. *Level-5*: **100% (Full Alignment):** A model trained on the same total dataset size as the standard cross-entropy baseline, but with complete Human guidance.

Varying training dataset size regimes: We also performed a comparative analysis of fairness between fully Human-AI aligned models and non-aligned models across different training data ratios (i.e., 25%, 50%, 75% and 100%),

Table 1: Datasets used for training and out-of-distribution (OOD) testing. The CheXlocalize dataset is utilized as an additional attention evaluation set of the CheXpert dataset.

Condition	Training				OOD Testing	
	NIH ChestX-ray14	PadChest	VinDr CXR	MIMIC CXR	JSRT	CheXpert (CheXlocalize)
Nodule & Mass	✓	✓	✓	–	✓	–
Pleural Effusion	✓	–	✓	✓	–	✓
Edema	✓	–	✓	✓	–	✓

where we measured group performance disparities across four key metrics: Accuracy, AUC, F1-score, and True Positive Rate (Sensitivity).

Ablation study - Effect of Random Alignment: We also performed an experiment where we randomized the attention areas the models are promoted to be aligned to. For each epoch, we randomly generate different shapes of attention maps at random locations.

Evaluation metrics: On the OOD datasets, we assessed fairness using the fairness gap metric proposed in [15], which quantifies the disparity in AUC performance between the best- and worst-performing demographic subgroups (e.g., male vs. female, and across different age subgroups(binary encoded based on the threshold); see Fig. 1 for details on subgroup definitions). Following [15], we considered AUC performance disparity as most relevant given that the positive and negative ratio of samples across all conditions is imbalanced. In addition, we evaluated the performance of models using the F1 score, accuracy, area under the ROC curve (AUC), and sensitivity across classes. Finally, to assess the degree of Human-AI alignment, we assessed the level of hit rate, as proposed in the XAI literature [24].

2.2 Human-AI Alignment for ViTs

Fig. 2 illustrates the architecture employed to perform Human-AI Alignment. It builds upon a recently proposed pre-trained medical Vision-Language Model (VLM) for chest X-ray diagnosis [18]. Below, we provide a summary of the approach for completeness reasons and derive the reader to [18] for details.

Overall Pipeline. Given an input image $\mathbf{I}$ and a textual prompt $\mathbf{T}$ (e.g., "Edema"), we first extract visual features $\mathbf{v} = \Phi_{\text{image}}(\mathbf{I})$ and language embeddings $\mathbf{t} = \Phi_{\text{text}}(\mathbf{T})$. These features are subsequently fused by a cross-attention module [17] to integrate visual features from chest X-rays with textual embeddings of clinical findings, producing cross-attention maps $\{\mathbf{M}_c\}$, where each $\mathbf{M}_c \in \mathbb{R}^{h \times w}$ corresponds to a particular class label c.

Attention Alignment. To address the discrepancy between clinicians' attention and the model's attention, the *Attention Aligner* module refines each cross-attention map. We excluded attention loss computation for negative sam-

ples because it has been shown that transformers can still focus on specific image regions[1] even when no relevant features are present. Consequently, the refined maps are supervised using two loss terms that are computed only on positive samples—that is, only on the pixels where the ground-truth annotation is non-zero. Let $\Omega^+ = \{i \in \Omega \mid Y_i \neq 0\}$ denote the set of positive pixels.

First, for attention alignment and following [18], we use a modified dice loss with false positive suppression defined as

$$\mathcal{L}_{\text{AL}} = 1 - \frac{2\sum_{i\in\Omega^+} Y_i P_i + \alpha + \varepsilon}{\sum_{i\in\Omega^+}(Y_i + P_i) + (w_{\text{FP}} - 1)\sum_{i\in\Omega^+} \text{FP}_i + \alpha + \varepsilon} \tag{1}$$

where Y_i and P_i denote the Human-annotated ground-truth and predicted attention values at pixel i, respectively. The terms α and ε are smoothing terms, and w_{FP} is a weighting factor for false positives.

The term $\mathcal{L}_{\text{AL}}$ enforces attention alignment between the provided Human-based attention and the model's attention map, indirectly enforcing the model to learn features yielding similar attention behavior as for the Human expert.

Classification Learning via Cross-Entropy Loss. This corresponds to the traditional Cross-Entropy loss term used to learn to solve the main task of classification. Let $\mathbf{y}_c \in \{0, 1\}$ denote the ground-truth label for finding c, and let $\mathbf{z}_c$ be the corresponding logit output from the classification head. Hence, the probability of image $\mathbf{I}$ to be classified as class c is $\mathbf{p}_c = \sigma(\mathbf{z}_c)$, where $\sigma(\cdot)$ denotes the sigmoid function. The cross-entropy loss is computed as $\mathcal{L}_{\text{CE}} = -\sum_{c\in\mathcal{N}} [\mathbf{y}_c \log(\mathbf{p}_c) + (1 - \mathbf{y}_c) \log(1 - \mathbf{p}_c)]$. This loss aligns the classification predictions with the ground-truth class labels, improving the model's diagnostic performance. The final loss is constructed as follows $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{AL}}$

Training Details: For training, we used ViT-B [4] as the visual backbone on an image size of 224 and Med-KEBERT [33] as the textual backbone. The model was optimized with AdamW using a learning rate of 5×10^{-5}. The training was conducted on a single H100 96G GPU with a total batch size of 32 for up to 1000 epochs, applying early stopping with patience of 30 epochs. The best-performing model was selected based on the highest validation AUC score. w_{FP} is set as 2.0. Each experiment was repeated five times, with all reported values averaged over the runs.

3 Results

Human-AI alignment improves fairness gap among different demographic groups: Table 2 shows the main results of our study assessing the impact of Human-AI alignment on fairness among demographic groups and performance across three datasets (CheXpert Edema, JSRT Nodule, and CheXpert Pleural Effusion). The results demonstrate that, compared to the baseline model without alignment (i.e., labeled as *w/o* in Table 2), Human-AI alignment improves the fairness gap for sex and age groups across different performance metrics and datasets, with improvements in 27 out of 30 comparisons (i.e., 5

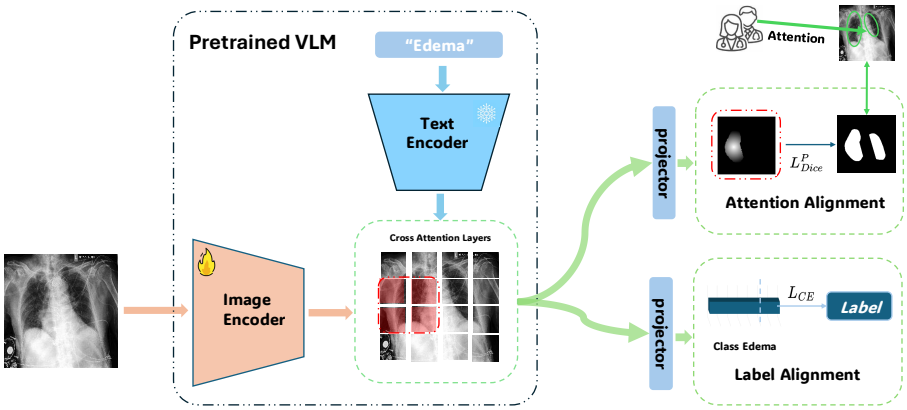

Fig. 2: Human-AI Alignment flow chart adapted from [18]. The approach is based on Visual-Language-Model (VLM) fusing image and language embeddings via cross-attention. The model is trained sequentially for each class per epoch, with the disease name as a prompt (e.g., "Edema"). Two projector heads are used to (i) optimize Human-AI alignment, and (ii) perform disease classification.

metrics × 3 datasets × 2 demographic groups, Table 2). Similar trends are observed in Fig. 3, showing that Human-AI alignment improves fairness metrics. However, exacerbating the alignment can also lead to diminished gains or even unintended trade-offs in fairness and performance. This finding aligns with the recent observations reported in [11].

Human-AI alignment improves performance in out-of-domain samples: Figure 6 shows the effect of Human-AI alignment on out-of-domain samples for nodule and mass detection, Edema, and Pleural Effusion (Table 1). Each

Table 2: Fairness gap comparison between the baseline model (w/o) and the Human-AI aligned model (Aligner) across OOD datasets and two demographic groups. Human-AI alignment reduced fairness gaps in 27 out of 30 comparisons with notable improvements across metrics (lower is better). The Hit Rate indicates the degree of Human-AI alignment (Higher is better). Best metrics in bold.

Method	Dataset	Demographic group	Accuracy gap (%) ↓	AUC gap↓	Sensitivity gap↓	F1 Score gap↓	Hit Rate↑
w/o [33]	CheXpert Edema	Age group	3.20 ± 0.19	2.91 ± 0.40	3.86 ± 2.11	3.26 ± 0.44	3.03 ± 1.50
Aligner	CheXpert Edema	Age group	**3.01 ± 0.57**	**2.01 ± 0.07**	**2.07 ± 0.42**	**2.82 ± 0.33**	**14.47 ± 8.71**
w/o [33]	CheXpert Edema	Gender group	1.69 ± 0.94	1.14 ± 1.25	2.69 ± 2.02	1.62 ± 1.23	3.03 ± 1.50
Aligner	CheXpert Edema	Gender group	**1.27 ± 0.45**	**0.17 ± 0.10**	**2.07 ± 0.84**	**1.01 ± 0.34**	**14.47 ± 8.7**
w/o [33]	JSRT Nodule	Age group	**16.11 ± 2.42**	**10.47 ± 2.16**	38.28 ± 9.69	33.29 ± 9.01	14.47 ± 8.71
Aligner	JSRT Nodule	Age group	20.09 ± 4.07	11.40 ± 3.31	**28.52 ± 5.11**	**25.98 ± 7.49**	**22.83 ± 7.96**
w/o [33]	JSRT Nodule	Gender group	**7.97 ± 3.30**	3.50 ± 2.31	12.91 ± 3.98	6.02 ± 3.52	14.47 ± 8.71
Aligner	JSRT Nodule	Gender group	8.25 ± 4.31	**1.07 ± 1.15**	**9.59 ± 9.20**	**7.68 ± 4.70**	**22.83 ± 7.96**
w/o [33]	CheXpert Pleural Effusion	Age group	9.50 ± 1.82	3.20 ± 0.48	17.42 ± 3.29	12.39 ± 3.52	10.82 ± 4.56
Aligner	CheXpert Pleural Effusion	Age group	**6.14 ± 0.38**	**3.84 ± 0.16**	**11.28 ± 1.02**	**7.51 ± 0.73**	**24.24 ± 13.33**
w/o [33]	CheXpert Pleural Effusion	Gender group	1.50 ± 0.99	**0.21 ± 0.18**	2.09 ± 1.24	1.30 ± 0.98	10.82 ± 4.56
Aligner	CheXpert Pleural Effusion	Gender group	**0.83 ± 0.93**	0.23 ± 0.20	**1.51 ± 1.19**	**0.64 ± 0.68**	**24.24 ± 13.33**

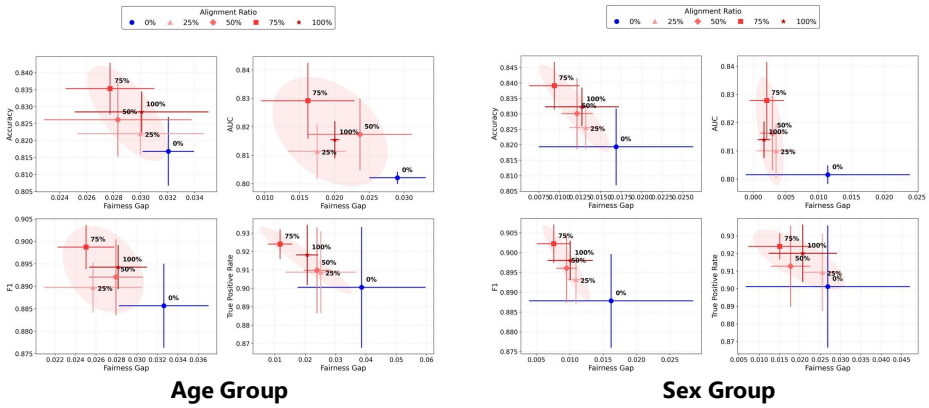

Fig. 3: Fairness-performance trade-off for age and sex groups across five levels of Human-AI alignment. Blue points represent non-aligned models (0%), while red-shaded points (25%–100%) indicate increasing alignment. Error bars show variability, and red-shaded ellipses highlight the trends. Fairness improves up to 75% alignment but degrades at 100%, suggesting an overconstraining effect.

radar chart shows the four performance metrics (higher the better), with and without human-AI alignment. Results show considerable performance improvements suggesting that Human-AI alignment promoted not only fairness improvements but also performance improvement on out-of-domain datasets, reflecting an important property for real-world clinical scenarios.

Human-AI alignment ensures stable fairness improvements in low-data scenarios: Figure 5 shows that Human-AI alignment improves fairness across all training data ratios, with the most significant impact in low-data scenarios (25%–50%), where it helps mitigate disparities more effectively.

Randomized Human-AI alignment reduces performance and fairness gap: Figure 4 presents the fairness-performance trade-off when Human-AI guidance is randomized (green points, the generation of random attention is illustrated at Fig1). As expected, randomization degrades performance but also reduces fairness gaps, suggesting a decorrelation effect on demographic attributes. This trade-off aligns with fairness-aware modeling literature, where reducing bias can sometimes come at the cost of lower performance.

4 Conclusion

Our study provides the first systematic exploration of the interplay between Human-AI alignment and fairness in medical image classification. Our results demonstrate that Human-AI alignment consistently reduces fairness gaps across sex and age groups, with improvements observed across datasets, tasks, and performance metrics, reinforcing the robustness of these findings. Beyond fairness

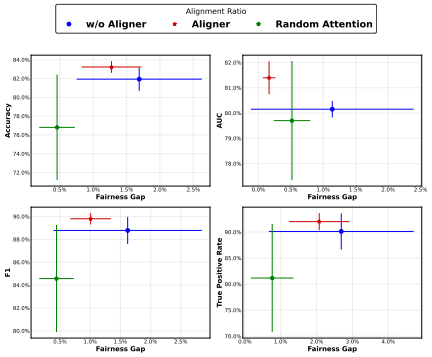

Fig. 4: Fairness-performance trade-off results under randomized Human-AI alignment (green) for sex group demographics, compared to a baseline model without alignment (blue), and with Human-AI alignment (red). Error bars represent variability.

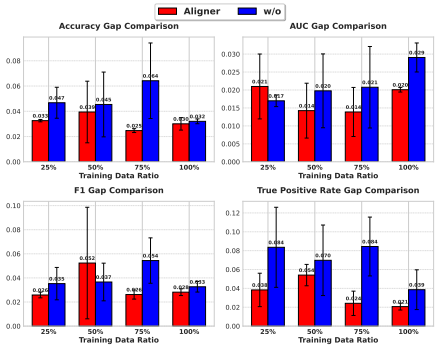

Fig. 5: Fairness comparison of Human-AI aligned (red) and non-aligned (blue) models across training data ratios and performance metrics. Lower values indicate better fairness. Alignment reduces fairness gaps, especially in low-data settings

benefits, we found that Human-AI alignment enhances out-of-domain performance, an essential property for real-world clinical deployment. These gains suggest that aligning model representations with human knowledge not only reduces bias but also strengthens performance when applied to unseen data, challenging the notion that fairness-improving interventions necessarily degrade accuracy. However, our findings also highlight the need for careful design and calibration of alignment strategies. While alignment generally improves both fairness and performance, excessive alignment can lead to diminished gains or even unintended trade-offs. Our randomized alignment ablation study further revealed that misguided alignment degrades performance while also reducing fairness gaps, suggesting a decorrelation effect between model predictions and demographic attributes. These results emphasize that the effectiveness of fairness interventions depends on how they are applied, underscoring the importance of balancing alignment for fairness and model utility. Overall, these findings highlight Human-AI alignment as a promising avenue for developing fair, robust, and generalizable AI models in medical imaging.

Acknowledgments. We thank the members of the MIA Group at ARTORG, University of Bern, and the Kaiko.AI team for their insightful discussions and support.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

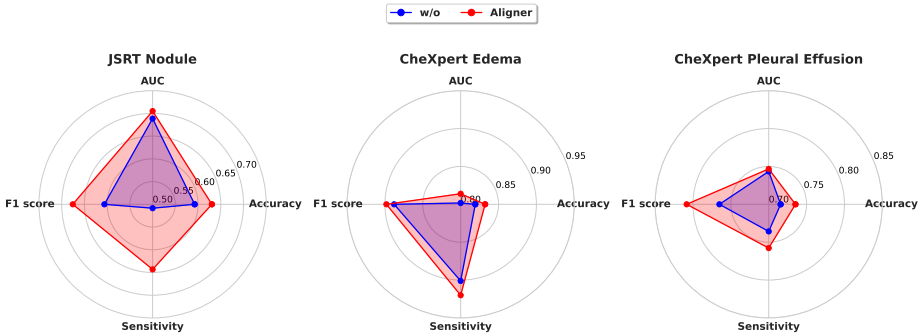

Fig. 6: Effect of Human-AI alignment on out-of-domain samples for nodule detection, Edema, and Pleural Effusion. Radar charts compare four performance metrics (outwards is better) with and without alignment. Results show consistent performance improvements across out-of-domain datasets.

References

1. Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. arxiv 2020. arXiv preprint arXiv:2005.00928 **10** (2022)
2. Bustos, A., Pertusa, A., Salinas, J.M., De La Iglesia-Vaya, M.: Padchest: A large chest x-ray image dataset with multi-label annotated reports. Medical image analysis **66**, 101797 (2020)
3. DeGrave, A.J., Janizek, J.D., Lee, S.I.: Ai for radiographic covid-19 detection selects shortcuts over signal. Nature Machine Intelligence **3**(7), 610–619 (2021)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
5. Gao, Y., Gu, S., Jiang, J., Hong, S.R., Yu, D., Zhao, L.: Going beyond xai: A systematic survey for explanation-guided learning. ACM Computing Surveys **56**(7), 1–39 (2024)
6. Gao, Y., Sun, T.S., Bai, G., Gu, S., Hong, S.R., Liang, Z.: Res: A robust framework for guiding visual explanation. In: proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining. pp. 432–442 (2022)
7. Gao, Y., Sun, T.S., Zhao, L., Hong, S.R.: Aligning eyes between humans and deep neural network through interactive attention alignment. Proceedings of the ACM on Human-Computer Interaction **6**(CSCW2), 1–28 (2022)
8. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020)
9. 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.: Ai recognition of patient race in medical imaging: a modelling study. The Lancet Digital Health **4**(6), e406–e414 (2022)
10. Glocker, B., Jones, C., Bernhardt, M., Winzeck, S.: Algorithmic encoding of protected characteristics in chest x-ray disease detection models. EBioMedicine **89** (2023)

11. Gorbatovski, A., Shaposhnikov, B., Malakhov, A., Surnachev, N., Aksenov, Y., Maksimov, I., Balagansky, N., Gavrilov, D.: Learn your reference model for real good alignment. arXiv preprint arXiv:2404.09656 (2024)
12. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 590–597 (2019)
13. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
14. Jones, C., Castro, D.C., De Sousa Ribeiro, F., Oktay, O., McCradden, M., Glocker, B.: A causal perspective on dataset bias in machine learning for medical imaging. Nature Machine Intelligence **6**(2), 138–146 (2024)
15. Ktena, I., Wiles, O., Albuquerque, I., Rebuffi, S.A., Tanno, R., Roy, A.G., Azizi, S., Belgrave, D., Kohli, P., Cemgil, T., et al.: Generative models improve fairness of medical classifiers under distribution shifts. Nature Medicine **30**(4), 1166–1173 (2024)
16. Larrazabal, A.J., Nieto, N., Peterson, V., Milone, D.H., Ferrante, E.: 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)
17. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems **32** (2019)
18. Luo, H., de Mortanges, A.P., Inel, O., Reyes, M.: Dwarf: Disease-weighted network for attention map refinement. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 59–68. Springer (2024)
19. Luo, H., Zhou, Z., Royer, C., Sekuboyina, A., Menze, B.: Devide: Faceted medical knowledge for improved medical vision-language pre-training. arXiv preprint arXiv:2404.03618 (2024)
20. Pahud de Mortanges, A., Luo, H., Shu, S.Z., Kamath, A., Suter, Y., Shelan, M., Pöllinger, A., Reyes, M.: Orchestrating explainable artificial intelligence for multimodal and longitudinal data in medical imaging. NPJ digital medicine **7**(1), 195 (2024)
21. Nguyen, H.Q., Lam, K., Le, L.T., Pham, H.H., Tran, D.Q., Nguyen, D.B., Le, D.D., Pham, C.M., Tong, H.T., Dinh, D.H., et al.: Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. Scientific Data **9**(1), 429 (2022)
22. Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**(6464), 447–453 (2019)
23. Rieger, L., Singh, C., Murdoch, W., Yu, B.: Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In: International conference on machine learning. pp. 8116–8126. PMLR (2020)
24. Saporta, A., Gui, X., Agrawal, A., Pareek, A., Truong, S.Q., Nguyen, C.D., Ngo, V.D., Seekins, J., Blankenberg, F.G., Ng, A.Y., et al.: Benchmarking saliency methods for chest x-ray interpretation. Nature Machine Intelligence **4**(10), 867–878 (2022)
25. Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I.Y., Ghassemi, M.: Chexclusion: Fairness gaps in deep chest x-ray classifiers. In: BIOCOMPUTING 2021: proceedings of the Pacific symposium. pp. 232–243. World Scientific (2020)

26. Seyyed-Kalantari, L., Zhang, H., McDermott, M.B., Chen, I.Y., Ghassemi, M.: Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nature medicine **27**(12), 2176–2182 (2021)
27. Song, C., Hui, C., Lin, Q., Zhang, W., Li, S., Zhang, S., Zhu, H., Li, Z., Liu, S., Jiang, F., Li, X.: Lvpnet: A latent-variable-based prediction-driven end-to-end framework for lossless compression of medical images (2025), https://arxiv.org/abs/2506.17983
28. Song, C., Hui, C., Zhang, W., Zhu, H., Liu, S., Huang, H., Jiang, F.: Bpclip: A bottom-up image quality assessment from distortion to semantics based on clip (2025), https://arxiv.org/abs/2506.17969
29. Stanley, E.A., Wilms, M., Mouches, P., Forkert, N.D.: Fairness-related performance and explainability effects in deep learning models for brain image analysis. Journal of Medical Imaging **9**(6), 061102–061102 (2022)
30. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE CVPR. pp. 2097–2106 (2017)
31. Wu, S., Zhang, X., Wang, B., Jin, Z., Li, H., Feng, J.: Gaze-directed vision gnn for mitigating shortcut learning in medical image. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 514–524. Springer (2024)
32. Zhang, H., Gerych, W., Ghassemi, M.: A data-centric perspective to fair machine learning for healthcare. Nature Reviews Methods Primers **4**(1), 86 (2024)
33. Zhang, X., Wu, C., Zhang, Y., Xie, W., Wang, Y.: Knowledge-enhanced visual-language pre-training on chest radiology images. Nature Communications **14**(1), 4542 (2023)
34. Zhou, Z., Luo, H., Taher, M.R.H., Pang, J., Ding, X., Gotway, M., Liang, J.: Ace: Anatomically consistent embeddings in composition and decomposition. arXiv preprint arXiv:2501.10131 (2025)