

REPRESENTATION DISENTANGLEMENT FOR FAIR CHEST X-RAY DIAGNOSIS

Yujie Sun¹, Ruizhe Li², Xiaowu Sun¹

¹ Academy of Artificial Intelligence and Advanced Technology, Xi'an Jiaotong-Liverpool University, China

² Nottingham Biomedical Research Centre, School of Medicine, University of Nottingham, UK

ABSTRACT

Deep learning has advanced chest X-ray (CXR) diagnosis, yet demographic biases in learned representations may contribute to performance disparities across intersectional groups. We propose a single-encoder framework combining dual-level decorrelation with prototype-guided cross-group contrastive learning to reduce demographic dependence while accounting for within-class variation. We further propose Demographic Representation Alignment Reduction (DRAR), a new metric that quantifies the reduction in demographic structure within disease representations. The framework is evaluated on four classification tasks using 34,809 CheXpert test images across eight intersectional groups, defined by age, sex and ethnicity. Compared with empirical risk minimization (ERM), our method reduces the mean equalized-odds gap from 15.41% to 10.86% and the AUC gap from 5.95% to 5.01%. Our method achieves a DRAR of 59.04% relative to ERM, with only a slight decrease in mean AUC. These results demonstrate that representation disentanglement can reduce demographic bias and improve intersectional fairness. Code is available at <https://github.com/06Yujie/Fair-Medical-Imaging>.

Index Terms— fair medical imaging, chest X-ray classification, representation disentanglement, contrastive learning

1. INTRODUCTION

Advances in chest X-ray (CXR) representation learning have improved disease recognition [1]. However, demographic bias remains a critical concern in CXR diagnosis [2, 3]. Models can exploit demographic and acquisition-related cues that correlate with disease labels. Reliance on these cues can encourage shortcut learning and impair generalization under distribution shifts [4, 5]. Such shortcuts can contribute to uneven diagnostic performance across demographic groups.

Several strategies have been developed to reduce demographic bias in disease diagnosis, including data-level augmentation [6] and bias-aware optimization [7]. More recently, representation-based methods have learned separate disease and demographic representations and disentangled them during training [8, 9, 10]. However, such methods often rely on separate encoders, complicating joint learning.

In addition, cross-group alignment based on disease labels may overlook within-class variation, as patients sharing the same label can exhibit heterogeneous imaging patterns [8]. Uniform alignment may therefore suppress diagnostically relevant variation and affect performance differently across demographic groups [9]. These limitations motivate a unified framework that jointly learns and disentangles disease and demographic representations while accounting for within-class patterns during cross-group alignment.

Moreover, evaluating fairness from model predictions provides limited insight into the learned representation space. Subgroup metrics, such as equalized-odds and area under the receiver operating characteristic curve (AUC) disparities, quantify performance gaps across demographic groups [11], while demographic probing measures the recoverability of sensitive attributes [5]. However, neither directly quantifies the correspondence between disease representation geometry and demographic structure. This motivates a representation-level metric that captures such correspondence, complementing performance-based fairness evaluation.

To address these challenges, we propose a unified framework that disentangles disease and demographic representations using a single encoder. We further account for within-class variation during cross-group alignment and introduce representation-level evaluation of demographic bias. We evaluate the framework across intersectional groups defined by age, sex, and ethnicity. Our main contributions are:

- We develop a single-encoder framework with feature- and relation-level decorrelation to disentangle disease and demographic representations.
- We propose prototype-guided cross-group contrastive learning to account for within-class variation during alignment.
- We introduce a representation-level fairness metric that quantifies demographic structure retained in disease representations.

2. METHODOLOGY

The framework has three main components: disease and demographic representation learning with an encoder and heads, feature- and relation-level decorrelation for representation disentanglement, and prototype-guided cross-group alignment that accounts for within-class variation.

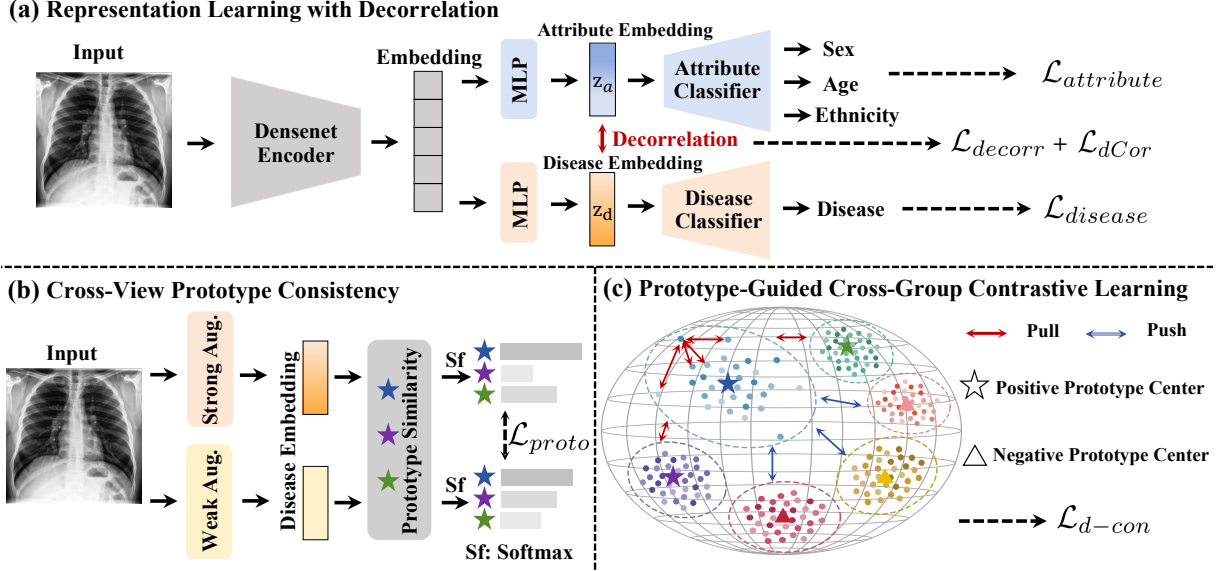


Fig. 1. Overview of the proposed framework. A shared encoder learns disease and attribute spaces through prototype-guided cross-group alignment and dual-level decorrelation.

2.1. Disease and Demographic Representation Learning

Given an image, a single DenseNet encoder [12] extracts a common representation, which is mapped by separate projection heads to ℓ_2 -normalized disease and demographic representations $\mathbf{z}_{d,i}$ and $\mathbf{z}_{a,i}$, respectively, as shown in Fig. 1.a.

The two representation spaces are learned through task-specific supervision. The disease representation is supervised by binary cross-entropy $\mathcal{L}_{\text{disease}}$, while the demographic representation is supervised by the averaged cross-entropy loss $\mathcal{L}_{\text{attribute}}$ for age, sex, and ethnicity. This supervision encourages the two heads to encode disease and demographic information in their respective representation spaces. During inference, only the encoder, disease projection head, and disease classifier are retained.

2.2. Feature- and Relation-Level Decorrelation

Separate projection heads do not ensure disentanglement of disease and demographic information. As shown in Figure 1.a, we therefore reduce their dependence at two complementary levels: feature-level decorrelation suppresses associations between feature dimensions, while relation-level decorrelation reduces dependence between pairwise distance patterns in the two spaces.

Feature-level decorrelation. To suppress feature-level dependence without penalizing between-class disease differences, we center both representations within each disease class and minimize their class-conditional centered cross-product:

$$\mathcal{L}_{\text{decorr}} = \frac{1}{|\mathcal{C}|h^2} \sum_{c \in \mathcal{C}} \left\| \tilde{\mathbf{Z}}_d^{(c)} \text{sg}(\tilde{\mathbf{Z}}_a^{(c)})^\top \right\|_F^2, \quad (1)$$

where $\mathcal{C} \subseteq \{0, 1\}$ is the set of disease classes represented by at least two samples in the mini-batch, and $\tilde{\mathbf{Z}}_d^{(c)}, \tilde{\mathbf{Z}}_a^{(c)} \in$

$\mathbb{R}^{h \times n_c}$ are the corresponding class-centered disease and demographic representation matrices, with samples arranged as columns. Here, h is the representation dimension, $\text{sg}(\cdot)$ denotes stop-gradient, and $\|\cdot\|_F$ is the Frobenius norm.

Relation-level decorrelation. Feature-level decorrelation does not necessarily remove demographic structure from the geometry of the disease representation. We therefore characterize sample relationships using pairwise Euclidean distance matrices $D_d, D_a \in \mathbb{R}^{n \times n}$, with entries $D_d^{ij} = \|\mathbf{z}_{d,i} - \mathbf{z}_{d,j}\|_2$ and $D_a^{ij} = \|\mathbf{z}_{a,i} - \mathbf{z}_{a,j}\|_2$ for a mini-batch of n samples, where $i, j \in \{1, \dots, n\}$. Each matrix is double-centered by subtracting its row and column means and adding its grand mean, yielding A_d and A_a . We minimize

$$\mathcal{L}_{\text{dCor}} = \frac{\max(0, \langle A_d, \text{sg}(A_a) \rangle_F)}{\sqrt{\langle A_d, A_d \rangle_F \langle \text{sg}(A_a), \text{sg}(A_a) \rangle_F} + \epsilon}, \quad (2)$$

where $\langle \cdot, \cdot \rangle_F$ is the Frobenius inner product, $\max(0, \cdot)$ avoids numerical negatives, and ϵ ensures stability. Minimizing $\mathcal{L}_{\text{dCor}}$ reduces dependence between pairwise distance patterns in the two spaces.

2.3. Prototype-Guided Variation-Aware Cross-Group Alignment

Beyond decorrelation, disease representations should be consistent across demographic groups while preserving meaningful within-class variation. Uniformly aligning same-label samples may suppress such variation. We therefore use class-specific prototypes [13] to characterize within-class patterns and guide cross-group alignment between samples with similar disease-related characteristics.

To model heterogeneous within-class patterns, we use $K = 3$ learnable unit-norm prototypes $\{\mathbf{u}_{c,k}\}_{k=1}^K$ for each class $c \in \{0, 1\}$ in the disease space. For sample i , a softmax

with temperature τ_p converts its cosine similarities to the prototypes of class y_i into an assignment $\mathbf{p}_i = (p_{i1}, \dots, p_{iK})^\top$, representing its affinity to different within-class patterns. As shown in Fig. 1.b, we enforce consistency of these assignments across augmented views and combine prototype affinity with demographic differences to weight pairs for contrastive alignment.

Cross-view prototype consistency. To obtain reliable within-class pattern assignments, we construct weak and strong views x_i^w and x_i^s using cropping, resizing, flipping, and affine transformations, with additional brightness and contrast jitter and Gaussian blur for the strong view. Both views are processed by the same encoder and disease projection head to obtain prototype assignments $\mathbf{p}_i^w$ and $\mathbf{p}_i^s$. Using the weak-view assignment as the target, we minimize

$$\mathcal{L}_{\text{proto}} = \frac{1}{n} \sum_{i=1}^n D_{\text{KL}}(\text{sg}(\mathbf{p}_i^w) \parallel \mathbf{p}_i^s), \quad (3)$$

where n is the mini-batch size and D_{KL} denotes the Kullback-Leibler divergence.

Prototype-guided pair weighting. As illustrated in Fig. 1.c, samples sharing the same disease label form positive pairs, while those with different labels form negative pairs. For pair (i, j) , we define prototype affinity r_{ij} and demographic difference δ_{ij} as

$$r_{ij} = 1 - \rho + \rho \text{sg}(\mathbf{p}_i^s)^\top \text{sg}(\mathbf{p}_j^s), \quad \delta_{ij} = \frac{1}{3} \sum_{t=1}^3 \mathbf{1}[a_{it} \neq a_{jt}], \quad (4)$$

where $t \in \{1, 2, 3\}$ indexes the three demographic attributes, and a_{it} denotes the t th demographic attribute of sample i , $\mathbf{1}[\cdot]$ is the indicator function, and ρ controls the contribution of prototype affinity. Their pair weights are

$$\begin{aligned} w_{ij}^+ &= \mathbf{1}[y_i = y_j](r_{ij} + \epsilon)^\gamma (1 + \alpha \delta_{ij}), \\ w_{ij}^- &= \mathbf{1}[y_i \neq y_j] \exp(-\delta_{ij}/\sigma), \end{aligned} \quad (5)$$

where α , γ , and σ control cross-group emphasis, prototype selectivity, and negative-pair decay, respectively, and ϵ ensures numerical stability. Thus, positive pairs with similar within-class patterns and greater demographic differences receive larger weights, whereas negative pairs with similar demographics are emphasized.

Prototype-guided contrastive learning. Using these pair weights, we perform supervised contrastive learning in the disease representation space. Let $q_{ij} = (\bar{\mathbf{z}}_{d,i}^s)^\top \bar{\mathbf{z}}_{d,j}^s / \tau$ denote the temperature-scaled cosine similarity, where $\bar{\mathbf{z}}_{d,i}^s = \mathbf{z}_{d,i}^s / \|\mathbf{z}_{d,i}^s\|_2$ and τ is the contrastive temperature. For each anchor i , let $\mathcal{P}(i) = \{j \neq i : y_j = y_i\}$ denote its positive set. We normalize the positive-pair weights as $\pi_{ij} = w_{ij}^+ / \sum_{\ell \in \mathcal{P}(i)} w_{i\ell}^+$ for $j \in \mathcal{P}(i)$, and define the weighted normalization term $b_i = \sum_{j' \neq i} (w_{ij'}^+ + w_{ij'}^-) \exp(q_{ij'})$. The contrastive loss is

$$\mathcal{L}_{d\text{-con}} = -\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{P}(i)} \pi_{ij} \log \frac{w_{ij}^+ \exp(q_{ij})}{b_i}, \quad (6)$$

where $\mathcal{I} = \{i : \mathcal{P}(i) \neq \emptyset\}$ contains anchors with at least one positive pair. This objective promotes disease-consistent representations across demographic groups while accounting for within-class variation.

The complete training objective combines disease and demographic supervision, dual-level decorrelation, prototype consistency, and contrastive alignment:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{\text{disease}} + \lambda_a \mathcal{L}_{\text{attribute}} + \mathcal{L}_{d\text{-con}} + \mathcal{L}_{\text{proto}} \\ &\quad + \lambda_{\text{decorr}} \mathcal{L}_{\text{decorr}} + \lambda_{\text{dCor}} \mathcal{L}_{\text{dCor}}, \end{aligned} \quad (7)$$

where λ_a , λ_{decorr} , and λ_{dCor} weight attribute supervision, feature-level decorrelation, and relation-level decorrelation.

2.4. Representation-Level Fairness Metric

Existing fairness evaluation does not directly characterize demographic structure in disease representations. We therefore propose **Demographic Representation Alignment Reduction (DRAR)**, which measures the structural alignment between disease representations and a demographic reference space and quantifies its reduction relative to ERM.

We train a reference model jointly supervised by age, sex, and ethnicity to obtain a demographic reference representation Z_a^{ref} . For each evaluated model, we measure the structural alignment between its disease representation Z_d and the demographic reference using centered kernel alignment (CKA) [14], denoted by $S = \text{CKA}(Z_d, Z_a^{\text{ref}})$. DRAR is then defined as

$$\text{DRAR} = \frac{S_{\text{ERM}} - S}{S_{\text{ERM}}}. \quad (8)$$

ERM has $\text{DRAR} = 0$, while higher values indicate a greater reduction in demographic alignment. DRAR complements outcome-level fairness metrics and demographic probing.

3. EXPERIMENTS AND RESULTS

3.1. Dataset and Demographic Groups

We use the CheXpert dataset [15], which comprises 224,316 chest radiographs from 65,240 patients. Four binary classification tasks are considered: *No Finding*, *Cardiomegaly*, *Pleural Effusion*, and *Pneumothorax*. The test set consists of 34,809 images and is divided into eight intersectional groups based on sex (Female/Male), ethnicity (White/Non-White), and age ($< 60 / \geq 60$ years).

3.2. Experimental Setup

We compare our method with ERM, ReSample [16], GroupDRO [17], DANN [18], FairAdaBN [19], and FairREAD [9]. For our method, we evaluated $K \in \{1, 3, 5\}$ and selected $K = 3$ based on the validation set AUC, we further set $h = 128$, $\tau = 0.1$, $\tau_p = 0.2$, $\rho = \alpha = \gamma = 1$, $\sigma = 0.5$, $\epsilon = 10^{-12}$, $\lambda_a = 0.25$, $\lambda_{\text{decorr}} = 500$, and $\lambda_{\text{dCor}} = 1$.

3.3. Evaluation Metrics

We evaluate fairness and diagnostic utility at both prediction and representation levels:

Prediction level: Diagnostic utility is evaluated using balanced accuracy (BACC), area under the receiver operating characteristic curve (AUC), and macro F1 score. Intersectional fairness across the eight demographic groups is measured by AUC disparity (Δ_{AUC}) and equalized-odds disparity (Δ_{EO}). Their trade-off is further assessed using $FATE_{AUC}$ and $FATE_{EO}$ [19], with BACC as the utility measure.

Representation level: Besides our proposed DRAR metric, age, sex, and ethnicity are predicted from frozen disease embeddings using auxiliary classifiers, with lower accuracy indicating less demographic information.

3.4. Results

We compare the methods at prediction and representation levels. We assess diagnostic utility and intersectional fairness, demographic information retained in disease representations, and cross-group representation patterns. Ablation studies further examine the contributions of the proposed components.

Diagnostic Utility and Intersectional Fairness. Table 1 reports the four-task mean results on the test cohort. Our method achieves the lowest disparities, reducing Δ_{EO} from 15.41% to 10.86% and Δ_{AUC} from 5.95% to 5.01% relative to ERM. Meanwhile, AUC decreases only from 83.80% to 83.36% and BACC from 76.44% to 75.47%. These results show that demographic disparities are substantially reduced while diagnostic utility is largely preserved, consistent with our objective of improving fairness rather than aggregate classification performance. Accordingly, our method achieves the highest $FATE_{EO}$ and $FATE_{AUC}$, indicating the best utility–fairness trade-off among the compared methods.

Table 1. Four-task mean diagnostic and intersectional fairness results (%). F denotes the corresponding FATE score.

Method	BACC $\uparrow$	AUC $\uparrow$	F1 $\uparrow$	Δ_{EO} $\downarrow$	Δ_{AUC} $\downarrow$	F_{EO} $\uparrow$	F_{AUC} $\uparrow$
ERM	76.44	83.80	66.22	15.41	5.95	0.00	0.00
GroupDRO	74.99	82.07	65.22	12.43	5.42	18.85	10.01
ReSample	75.98	83.09	65.14	14.30	5.78	5.21	0.39
DANN	74.11	81.83	66.11	15.73	5.56	-1.60	3.50
FairAdaBN	74.70	82.02	65.13	14.10	5.95	1.33	-1.67
FairREAD	74.07	82.08	66.56	19.44	6.46	-34.01	-10.30
Ours	75.47	83.36	65.34	10.86	5.01	28.26	14.53

Representation-Level Fairness. Table 2 evaluates demographic information retained in disease representations. We predict age, sex, and ethnicity from frozen disease embeddings by training multilayer perceptrons (MLPs). Our method achieves the lowest accuracy for the three attributes (69.43%, 82.50%, and 59.35%) and the highest DRAR (59.04%), indicating reduced demographic predictability and structural alignment with the demographic reference. Figure 2 analyzes the learned representations for Pleural Effusion. Compared with ERM, our method shows greater overlap across demographic groups while preserving disease-label separation. The pairwise distances in the feature space show a consistent pattern, supporting improved cross-group alignment without compromising disease discrimination.

Ablation Study. Table 3 evaluates the effects of dual-level decorrelation and prototype-guided cross-group align-

Table 2. Representation-level fairness results (%). Demographic prediction accuracy is reported as mean $\pm$ standard deviation across four tasks. “–” denotes methods using demographic attributes at inference.

Method	Demographic Prediction Accuracy $\downarrow$				DRAR $\uparrow$
	Age	Sex	Ethnicity		
ERM	72.32 $\pm$ 1.66	86.22 $\pm$ 2.12	62.00 $\pm$ 1.34		0.00
GroupDRO	71.98 $\pm$ 1.73	85.63 $\pm$ 2.01	61.51 $\pm$ 1.48		22.22
ReSample	72.20 $\pm$ 1.90	85.80 $\pm$ 2.44	61.99 $\pm$ 1.40		2.81
DANN	70.94 $\pm$ 2.27	85.95 $\pm$ 1.83	61.94 $\pm$ 1.13		20.92
FairAdaBN	–	–	–		-0.60
FairREAD	–	–	–		13.07
Ours	69.43$\pm$4.09	82.50$\pm$3.70	59.35$\pm$1.61		59.04

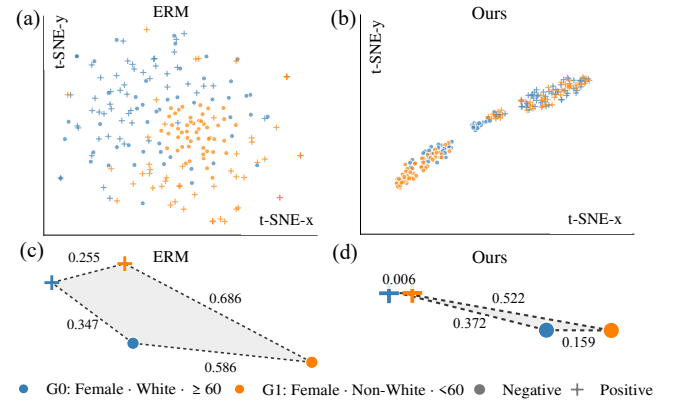


Fig. 2. Pleural Effusion feature analysis. (a,b) t-SNE plots for ERM and our method. (c,d) Mean Euclidean distances between group–label representations in original feature space.

ment on disease and demographic representation learning. Removing prototype-guided alignment increases Δ_{EO} from 10.86% to 14.28%, while removing decorrelation raises it to 17.72% and reduces AUC to 82.36%. Combining three components yields the best fairness and diagnostic performance.

Table 3. Ablation study of the proposed components.

Method	$\mathcal{L}_{disease} + \mathcal{L}_{attribute}$	$\mathcal{L}_{decorr} + \mathcal{L}_{dCor}$	$\mathcal{L}_{d-con} + \mathcal{L}_{proto}$	BACC $\uparrow$	AUC $\uparrow$	F1 $\uparrow$	Δ_{EO} $\downarrow$	Δ_{AUC} $\downarrow$
Ablation 1	✓			73.59	82.21	62.60	16.67	5.30
Ablation 2	✓	✓		75.43	83.10	64.49	14.28	5.21
Ablation 3	✓		✓	75.15	82.36	64.09	17.72	5.79
Full model	✓	✓	✓	75.47	83.36	65.34	10.86	5.01

4. CONCLUSION

We present a single-encoder framework for fair CXR classification across intersectional demographic groups. The framework disentangles disease and demographic representations through dual-level decorrelation and aligns disease representations using contrastive learning. Results across four CheXpert tasks demonstrate improved intersectional fairness at prediction and representation levels. These findings highlight the importance of reducing demographic dependence in learned representations for fair CXR diagnosis.

5. ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (NSFC) under Grant No. 82604003 and the XJTLU Research Development Fund (RDF-25-01-067). We acknowledge the CheXpert dataset used in this study.

6. REFERENCES

- [1] Jingfeng Yao et al., “EVA-X: a foundation model for general chest X-ray analysis with self-supervised learning,” *npj Digit. Med.*, vol. 8, 2025, Art. no. 678.
- [2] Zikang Xu, Jun Li, Qingsong Yao, Han Li, Mingyue Zhao, and S. Kevin Zhou, “Addressing fairness issues in deep learning-based medical image analysis: a systematic review,” *npj Digit. Med.*, vol. 7, 2024, Art. no. 286.
- [3] Ben Glocker, Charles Jones, Mélanie Roschewitz, and Stefan Winzeck, “Risk of bias in chest radiography deep learning foundation models,” *Radiol. Artif. Intell.*, vol. 5, no. 6, 2023, Art. no. e230060.
- [4] William Lotter, “Acquisition parameters influence AI recognition of race in chest X-rays and mitigating these factors reduces underdiagnosis bias,” *Nat. Commun.*, vol. 15, 2024, Art. no. 7465.
- [5] Yuzhe Yang, Haoran Zhang, Judy W. Gichoya, Dina Katabi, and Marzyeh Ghassemi, “The limits of fair medical imaging AI in real-world generalization,” *Nat. Med.*, vol. 30, pp. 2838–2848, 2024.
- [6] David Jacob Drexlin et al., “MeDi: Metadata-guided diffusion models for mitigating biases in tumor classification,” in *Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2025, pp. 379–388.
- [7] Luyang Luo, Xin Huang, Minghao Wang, Zhuoyue Wan, Wanteng Ma, and Hao Chen, “Mitigating medical dataset bias by learning adaptive agreement from a biased council,” *Med. Image Anal.*, vol. 105, 2025, Art. no. 103629.
- [8] Pengcheng Zhao and Xiaowei Ding, “DFCA: Disentangled feature contrastive learning and augmentation for fairer dermatological diagnostics,” in *Proc. 34th Int. Joint Conf. Artificial Intelligence (IJCAI)*, 2025, pp. 646–654.
- [9] Yicheng Gao, Jinkui Hao, and Bo Zhou, “FairREAD: Re-fusing demographic attributes after disentanglement for fair medical image classification,” *Med. Image Anal.*, vol. 107, 2026, Art. no. 103858.
- [10] Chenlang Yi et al., “AdFair-CLIP: Adversarial fair contrastive language-image pre-training for chest X-rays,” in *Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2025, pp. 13–23.
- [11] Moritz Hardt, Eric Price, and Nati Srebro, “Equality of opportunity in supervised learning,” in *Advances in Neural Information Processing Systems*, 2016, vol. 29.
- [12] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger, “Densely connected convolutional networks,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708.
- [13] Simon Reiß, Constantin Seibold, Alexander Freytag, Erik Rodner, and Rainer Stiefelhagen, “Decoupled semantic prototypes enable learning from diverse annotation types for semi-weakly segmentation in expert-driven domains,” in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 15495–15506.
- [14] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton, “Similarity of neural network representations revisited,” in *Proceedings of the 36th International Conference on Machine Learning*, 2019, vol. 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529, PMLR.
- [15] Jeremy Irvin et al., “CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison,” in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, pp. 590–597.
- [16] Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz, “Simple data balancing achieves competitive worst-group-accuracy,” in *Proc. Conf. Causal Learning and Reasoning (CLear)*, 2022, vol. 177 of *Proceedings of Machine Learning Research*, pp. 336–351.
- [17] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang, “Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization,” in *Proc. Int. Conf. Learning Representations (ICLR)*, 2020.
- [18] Yaroslav Ganin et al., “Domain-adversarial training of neural networks,” *J. Mach. Learn. Res.*, vol. 17, no. 59, pp. 1–35, 2016.
- [19] Zikang Xu, Shang Zhao, Quan Quan, Qingsong Yao, and S. Kevin Zhou, “FairAdaBN: Mitigating unfairness with adaptive batch normalization and its application to dermatological disease classification,” in *Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2023, pp. 307–317.