Robust Mitigation of Age-Dependent Confounding Effects via Sample-Difficulty Decorrelation

Nikhil Cherian Kurian[1,2⋆], Victor Caquilpan Parra[1], Abin Shoby[1], Luke Whitbread[1], and Lyle J. Palmer[1,2]

[1] Australian Institute for Machine Learning, Adelaide University, Adelaide, Australia
nikhil.kurian@adelaide.edu.au
[2] School of Public Health, Adelaide University, Adelaide, Australia

Abstract. Age-dependent performance disparities in medical image classification often arise because age acts as a confounder, linking imaging morphology with disease prevalence. In practice, disparities can manifest as overdiagnosis at ages where disease prevalence is higher and underdiagnosis at ages where prevalence is lower, and can become more pronounced under train–test shifts in the age distribution. Conventional mitigation approaches that enforce strict age invariance may suppress diagnostically meaningful information encoded in age. We therefore propose a robust framework that mitigates the effects of age-dependent confounding by targeting first-order, label-conditioned age–error trends rather than enforcing invariance. Following a warm-up phase, we characterize model-perceived difficulty and model its age-dependent trends in a label-conditioned manner. We decorrelate age from dominant age–difficulty trends using robust, Huber-weighted affinity weights, attenuating confounding driven shortcuts while preserving clinically meaningful, nonlinear age information. We further introduce an Age Coverage Score that scales the decorrelation penalty by mini-batch age variance to ensure stable optimization under limited age diversity. Across two radiology datasets, our approach reduces age-dependent true- and false-positive disparities with minimal AUC impact and remains robust to increasing train–test age distribution shifts.

Keywords: Confounding · Shortcut learning · Spurious correlations.

1 Background and Related Work

Deep neural networks in medical imaging often rely on shortcuts—readily available cues such as age—to reduce training error rather than learning true disease markers [6, 22]. Age can act as a confounder because it is associated with both imaging morphology (input appearance) and disease prevalence (label distribution), encouraging models to use age-linked signals instead of subtle pathological features [3, 19]. The resulting confounding effects appear as systematic

⋆ Corresponding author. Email: nikhil.kurian@adelaide.edu.au.

age-dependent shifts in operating characteristics, with true- and false-positive rates (TPR/FPR) varying across the age spectrum [8, 21]. Despite this, most age confounding–mitigation approaches discretize age or remove it from the feature space, with limited work examining how such schemes affect model performance [4]. In this context, enforcing strict age invariance is often undesirable, since age also encodes clinically meaningful variation, motivating approaches that selectively mitigate spurious age dependencies rather than removing age information entirely [5]. These challenges are further exacerbated under train–test shifts in patient age, which commonly accompany prevalence changes [7].

We propose a framework that mitigates age-dependent confounding effects by targeting spurious age-linked trends while preserving higher-order age information. Our approach applies a decorrelation that prevents simple age-based trends from distorting learning dynamics, effectively flattening the dominant age-related slope across the age spectrum. This "one-slope" strategy is less sensitive to the particular age distribution in the training set and remains effective under train–test age shifts. To handle limited age diversity within mini-batches, we incorporate an Age Coverage Score that adaptively scales the penalty based on batch-wise age variance. Evaluations on two chest X-ray benchmarks under varying train–test age distribution shifts show that our method reduces age-dependent TPR/FPR disparities while maintaining diagnostic accuracy.

2 Method

We consider a radiological finding prediction task and propose a regularization framework that modulates age-dependent trends in training dynamics. Rather than enforcing strict age invariance in representations, our approach targets label-conditioned linear age–error trends during learning. We quantify the training dynamics for each sample i via the gradient magnitude g_i. Under standard binary cross-entropy loss ($\mathcal{L}_{\text{BCE}}$), the logit-gradient magnitude is $g_i = |\partial \mathcal{L}_{\text{BCE}} / \partial \ell_i| = |p_i - y_i|$, where ℓ_i is the logit, $p_i \in [0, 1]$ is the predicted probability, and $y_i \in \{0, 1\}$ is the ground-truth label [18]. Thus, g_i is a bounded, label-conditioned proxy for the model's current prediction error, requiring no gradient hooks or explicit backward-pass interventions. This provides a bounded $[0, 1]$ proxy for model-perceived difficulty that remains computationally efficient.

2.1 Empirical Motivation

Following a model warm-up phase (one-epoch), we analyze the conditional expectation $\mathbb{E}[g \mid \text{Age}, y]$ and observe a strong linear dependence on age across the radiology datasets (Fig. 1), with Pearson correlation coefficients $|r| > 0.9$. Specifically, when disease is present ($y = 1$), samples from older patients are, on average, "easier" (lower g_i) for the model than those from younger patients; conversely, when disease is absent ($y = 0$), older patients represent more difficult samples. These trends mirror the empirical age-dependent confounding effects in these datasets—namely, reduced true-positive rates (under-diagnostic)

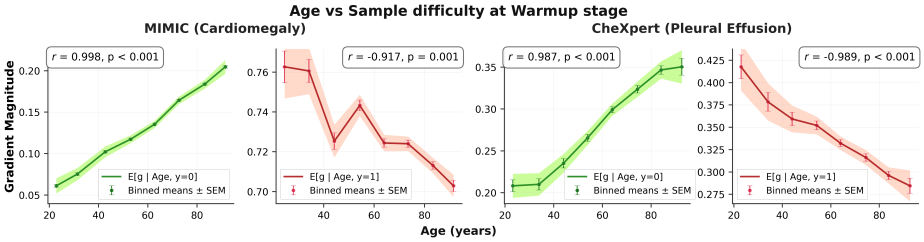

Fig. 1. Empirical age–difficulty trends at warm-up. Age-conditioned mean model-perceived difficulty g at warm-up, averaged over five seeds. Points show age-bin means (±SEM); curves show $\mathbb{E}[g \mid \text{Age}, y]$ (red: y=1, green: y=0), matching the observed lower TPR/FPR at younger ages and higher TPR/FPR at older ages [8, 21].

for younger patients and elevated false-positive rates (over-diagnostic) for older patients, suggesting that age is associated with systematic variation in prediction error and operating characteristics [8].

Although the strength and direction of this trend may vary across cohorts (e.g., pediatric vs. geriatric dominant), the consistent presence of monotonic age–difficulty relationships can distort training dynamics and induce systematic bias [8]. This motivates an intervention that suppresses dominant age-linked error trends associated with these disparities, rather than enforcing strict age invariance that may remove clinically meaningful information [22].

2.2 Weighted Slope Penalty

Motivated by the observed label-conditioned age–difficulty relationship, we model the interaction between normalized age z_i and model-perceived difficulty g_i as:

$$g_i \approx \alpha_{y_i} + \beta_{y_i} z_i, \tag{1}$$

where parameters are estimated independently for each label $y_i \in \{0, 1\}$. Rather than enforcing full statistical independence, we impose a weaker constraint by decorrelating z and g. This is implemented by penalizing the squared estimate of the age–difficulty slope β_{y_i} within each mini-batch. This approach attenuates first-order linear age–error trends without enforcing full age invariance.

Directly penalizing $\beta_{y_i}^2$ (decorrelation) is often unstable, as g_i inherently reflects clinical variability and annotation noise. Therefore, unweighted decorrelation risks penalizing highly informative samples or inducing optimization instability. We address this by estimating slopes via a Weighted Ordinary Least Squares (WOLS) formulation, where each sample is weighted by its *trend affinity*—a measure of alignment with the dominant age–difficulty relationship. Trend affinity is estimated once immediately after warm-up using label-conditioned Huber regression on the full training set and then kept fixed, since early gradients best reflect relative difficulty while later gradients shrink and yield noisier re-estimates [1]. Moreover, Huber regression residuals at convergence provide

per-sample weights equivalent to a WOLS solution [12, 9]. Specifically, for each sample i, we compute the residual r_i from the Huber fit:

$$r_i \;=\; g_i - \big(\hat{\alpha}_{y_i} + \hat{\beta}_{y_i} z_i\big), \tag{2}$$

which measures the deviation from the fitted trend. Influence weights w_i are then derived from a robust scale estimate δ_{y_i} of the residuals:

$$w_i \;=\; \begin{cases} 1, & |r_i| \le \delta_{y_i}, \\ \dfrac{\delta_{y_i}}{|r_i|}, & |r_i| > \delta_{y_i}. \end{cases} \tag{3}$$

Samples within a central residual band receive full weight, while those farther from the trend are smoothly downweighted. Consequently, mini-batch slope penalization primarily acts on samples participating in the systematic linear age–difficulty dependence, reducing the risk of unstable updates or training collapse when decorrelating across all gradients. In our experiments, δ_{y_i} was set to the median absolute deviation (MAD) of the residuals, a standard robust scale estimator in Huber-type regression and weighted least squares [12, 9]. It is used here as a robust default rather than a uniquely optimal scale choice, and controls how aggressively off-trend samples are downweighted.

2.3 Batch-level Slope Estimation

During optimization, we estimate the linear age–difficulty trend within each mini-batch $\mathcal{B}$, calculated separately for each class $y \in \{0, 1\}$ using the trend-affinity weights w_i. Let $\mathcal{B}_y = \{i \in \mathcal{B} \mid y_i = y\}$ denote the subset of batch samples with label y. The WOLS estimate of the batch-level slope, $\hat{\beta}_y$, is given by the ratio of the weighted age–difficulty covariance to the weighted age variance [10]:

$$\hat{\beta}_y \;=\; \frac{\sum_{i \in \mathcal{B}_y} w_i \,(z_i - \mu_{z,y})\,(g_i - \mu_{g,y})}{\sum_{i \in \mathcal{B}_y} w_i \,(z_i - \mu_{z,y})^2}, \tag{4}$$

where $\mu_{z,y}$ and $\mu_{g,y}$ are the weighted batch means of normalized age and model-perceived difficulty:

$$\mu_{z,y} \;=\; \frac{\sum_{i \in \mathcal{B}_y} w_i z_i}{\sum_{i \in \mathcal{B}_y} w_i}, \qquad \mu_{g,y} \;=\; \frac{\sum_{i \in \mathcal{B}_y} w_i g_i}{\sum_{i \in \mathcal{B}_y} w_i}. \tag{5}$$

We define the corresponding trend regularization term for each class as:

$$\mathcal{L}_{\text{slope},y} \;=\; \hat{\beta}_y^2. \tag{6}$$

By minimizing this penalty, the model is encouraged to reduce reliance on dominant linear age-based shortcuts linked to age during training. Because g_i is derived from the model output p_i, the penalty is fully differentiable, allowing for direct modulation of training dynamics.

2.4 Age Coverage Modulation

To account for the reliability of batch-wise slope estimation, we introduce an *Age Coverage Score* $C \in [0, 1]$, which quantifies the age diversity within a mini-batch. We define C as the normalized variance of the normalized age values $\{z_i\}_{i=1}^{N}$:

$$\mathrm{Var}(z) \;=\; \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})^2, \qquad C \;=\; \frac{\mathrm{Var}(z)}{\mathrm{Var}_{\max}}, \tag{7}$$

where N is the mini-batch size, $\bar{z}$ is the batch mean and $\mathrm{Var}_{\max} = 0.25$ is the theoretical maximum variance for $z \in [0, 1]$ [2]. Higher C indicates broader age coverage and more reliable slope estimation within a mini-batch. In the objective below, we compute this score within each label subset, yielding a label-conditioned coverage score C_y.

We incorporate this score into a coverage-aware objective:

$$\mathcal{L}_{\mathrm{total}} \;=\; \mathcal{L}_{\mathrm{BCE}} \;+\; \lambda \sum_{y\in\{0,1\}} C_y\, \mathcal{L}_{\mathrm{slope},y} \tag{8}$$

This modulation downweights decorrelation updates when age coverage is limited, improving robustness under age imbalance and train–test age distribution shifts. Consequently, this coverage-aware modulation ensures that the model only reduces the age–difficulty slope toward zero when the mini-batch provides a representative sample of the underlying age distribution. By basing the penalty on sufficient evidence of a trend, the framework maintains stable fairness mitigation that is inherently less sensitive to the specific demographic compositions encountered during train–test distribution shifts.

3 Experiment Details

3.1 Datasets and Distribution Shift

We evaluate our framework on the most prevalent findings in two benchmark radiology datasets: **CheXpert Pleural Effusion (CXP PE)** [13] and **MIMIC-CXR Cardiomegaly (MIMIC-CXR Car)** [14], which exhibit some of the strongest age-related unfairness [8]. To conduct experiments with shifted age distributions between training and testing environments, we first establish a strict patient-level split. We then apply an age-based exponential sampling scheme to generate refined splits with skewed age profiles. For a sample with normalized age $z \in [0, 1]$, the sampling weights for the training and testing pools (w_{tr} and w_{te}) are:

$$w_{tr} \propto e^{\gamma(z-b_{tr})}, \quad w_{te} \propto e^{-\gamma(z-b_{te})}, \tag{9}$$

where the training pivot b_{tr} is set at the 25th percentile and the testing pivot b_{te} at the 75th percentile of the dataset age distribution.

As shown in Fig. 2, we utilize three shift intensities $\gamma \in \{0, 4, 8\}$. As γ increases, the training set is drawn increasingly from a younger population while

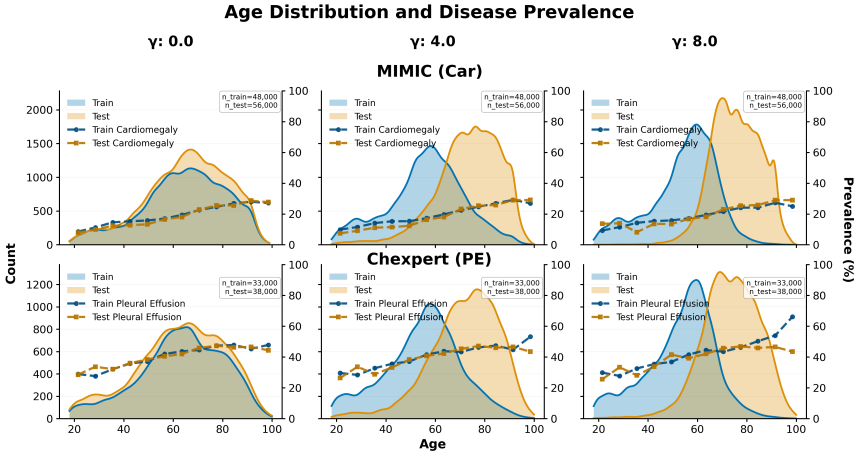

Fig. 2. Sampling-induced age distribution shifts. Training (blue) and test (orange) age densities for $\gamma \in \{0, 4, 8\}$. Larger γ yields stronger age shifts. Dashed lines show prevalence as a function of age in train (blue) and test (orange).

the test set shifts toward an older cohort. This creates a challenging stress test scenario with significant covariate shift, as the model is evaluated on a population with higher disease prevalence and increased morphological complexity than observed during training. Finally, 10% of the training pool is reserved as a distribution-consistent validation set for hyperparameter tuning.

3.2 Experimental Setting

We use DenseNet-121 [11] for all experiments across both datasets. Models are trained for 30 epochs on a single NVIDIA GeForce RTX 3090 with batch size 64 and initial learning rate 0.001 using Adam optimizer [15]. Results are aggregated over five random seeds. The test operating point is selected by maximizing Youden's J on the validation set [8, 5]. The regularization coefficient $\lambda = 1.2$ was selected on the validation set to balance the reduction in ΔSep_{10} against the AUC drop relative to ERM [17, 5].

3.3 Evaluation Metrics

To evaluate mitigation, we assess fairness improvement using a continuous-age analogue of Equalized Odds [24] that captures age-dependent TPR and FPR disparities at a fixed operating point selected on the validation set. Following prior work [3, 23], we use the *separation* metric to quantify systematic age-related performance disparities.

Age dependence is estimated by fitting logistic models of age separately to positive ($y = 1$) and negative ($y = 0$) samples, yielding coefficients s^+ and s^- that capture structured age-dependent variation in TPR and FPR, respectively.

Table 1. Age-separation metrics ($|S^+|, |S^-|, \Delta\text{Sep}_{10}$) across shift strengths $\gamma \in \{0, 4, 8\}$. Lower values indicate superior fairness. $|S^{\pm}|$ values are scaled by 10^{-2} and averaged over five runs with standard error. **Bold** denotes the best result; underlined the second best.

Method	Dataset	$\gamma = 0$			$\gamma = 4$			$\gamma = 8$			Avg ΔSep_{10} (%)
		$\lvert S^+\rvert$	$\lvert S^-\rvert$	ΔSep_{10}(%)	$\lvert S^+\rvert$	$\lvert S^-\rvert$	ΔSep_{10}(%)	$\lvert S^+\rvert$	$\lvert S^-\rvert$	ΔSep_{10}(%)	
ERM	MIMIC-CXR (Car)	2.31±0.04	3.58±0.03	32.21±0.27	2.11±0.05	3.49±0.06	32.26±0.20	2.04±0.02	3.54±0.06	32.31±0.21	32.26±0.22
ReSampled		2.00±0.05	3.40±0.03	30.90±0.36	2.42±0.08	3.75±0.07	36.13±0.31	2.69±0.04	3.75±0.04	38.04±0.35	35.02±0.32
Adv		1.31±0.02	2.62±0.06	21.73±0.32	1.95±0.05	3.44±0.03	30.95±0.30	2.04±0.06	3.54±0.04	32.22±0.26	28.30±0.26
JTT		1.42±0.03	2.92±0.05	24.23±0.25	1.67±0.03	3.31±0.07	28.27±0.31	1.97±0.02	3.32±0.06	30.28±0.28	27.60±0.28
GDRO		1.01±0.03	2.84±0.04	21.26±0.30	1.67±0.04	3.36±0.03	28.59±0.27	2.42±0.03	3.72±0.04	35.93±0.34	28.59±0.27
Our		**0.84±0.03**	**2.42±0.05**	**17.69±0.31**	**1.02±0.03**	**2.84±0.06**	**21.30±0.28**	**1.36±0.05**	**2.86±0.06**	**23.48±0.25**	**20.82±0.24**
ERM	CXP (PE)	1.35±0.02	1.89±0.03	17.59±0.14	1.30±0.03	1.99±0.06	17.89±0.14	1.59±0.02	1.76±0.05	18.21±0.19	17.90±0.17
ReSampled		1.00±0.02	1.33±0.03	12.30±0.17	0.92±0.03	1.47±0.05	12.69±0.16	1.30±0.01	1.53±0.02	15.58±0.21	13.52±0.21
Adv		0.86±0.03	1.37±0.06	11.82±0.15	1.06±0.03	1.39±0.07	13.01±0.18	1.41±0.02	1.84±0.06	17.60±0.23	14.14±0.19
JTT		0.73±0.03	1.50±0.05	11.82±0.24	1.10±0.04	1.40±0.04	13.28±0.21	**0.81±0.02**	1.69±0.07	13.29±0.24	12.80±0.21
GDRO		0.94±0.01	1.41±0.05	12.46±0.16	1.12±0.03	1.59±0.05	14.55±0.19	1.52±0.04	1.52±0.06	16.44±0.22	14.49±0.22
Our		**0.56±0.02**	**0.89±0.07**	**7.51±0.14**	**0.54±0.03**	**0.85±0.06**	**7.19±0.15**	0.88±0.03	**0.80±0.06**	**8.76±0.23**	**7.82±0.18**

We report $|s^+|$ and $|s^-|$ as the magnitudes of these effects. Further, the separation coefficient is defined as $\text{Sep} = \frac{1}{2}(|s^+| + |s^-|)$. For interpretability, we report fairness as the relative change per decade of age [3],

$$\Delta\text{Sep}_{10} = \exp(\text{Sep} \cdot \Delta a) - 1, \qquad \Delta a = 10 \text{ years}. \tag{10}$$

We further analyze the AUC (and ΔAUC)–ΔSep_{10} trade-off in the classical fairness–accuracy setting [17], particularly as γ increases.

3.4 Comparison with Existing Methods

Our approach is compared against Empirical Risk Minimization (ERM) and other baselines: JTT [16], which upweights samples misclassified in an initial training stage; Resampled Batch [3], which enforces a uniform age distribution per mini-batch; GroupDRO [20], which performs worst-group reweighting over age-defined groups (10-year age bins, matching the ΔSep_{10} metric) and Adversarial Training (Adv) [3], uses a gradient-reversal adversarial head trained with a mean squared error objective.

4 Results and Discussion

Table 1 reports ΔSep_{10}(%), defined as the percentage relative change in TPR and FPR per decade of age, across shift strengths $\gamma \in 0, 4, 8$ (lower is better). Across both datasets, our method achieves the lowest ΔSep_{10} at all γ, indicating consistently reduced age dependence.

On MIMIC-CXR (Car) and CXP (PE), our method attains average ΔSep_{10}(%) of 20.82% and 7.82%, yielding relative reductions of ∼35% and ∼56% over ERM (32.26%, 17.90%). With increasing age shift, several mitigation baselines degrade substantially; notably, age-invariant adversarial training on MIMIC-CXR worsens from 21.73% (γ=0) to 32.22% (γ=8), approaching ERM-level fairness and in

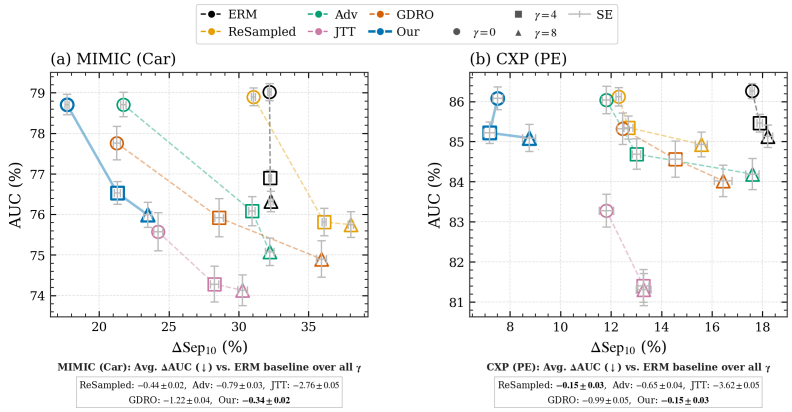

Fig. 3. AUC (%) vs. ΔSep_{10}(%) on two datasets. Marker shape encodes $\gamma \in \{0, 4, 8\}$; error bars show Standard Error. Our method (blue) achieves the best fairness–performance trade-off. ΔAUC relative to ERM AUC (mean $\pm$ SE, avg. over γ) below.

Table 2. Ablation study evaluated with ΔSep_{10}(%) and AUC (%) across shift strengths $\gamma \in \{0, 4, 8\}$, examining the contributions of trend affinity and coverage score. The Reference row corresponds to the full model with both components enabled.

Method	$\gamma = 0$		$\gamma = 4$		$\gamma = 8$		Avg	Avg
	ΔSep_{10}(%)	AUC (%)	ΔSep_{10} (%)	AUC (%)	ΔSep_{10} (%)	AUC (%)	ΔSep_{10} (%)	AUC (%)
W/O Trend Affinity	18.02±0.23	78.24±0.25	21.71±0.28	75.98±0.22	24.01±0.29	75.18±0.26	21.25±0.32	76.47±0.24
W/O Coverage Score	19.41±0.26	78.81±0.21	23.87±0.31	76.31±0.18	25.13±0.27	75.53±0.29	22.80±0.29	76.88±0.23
Reference	**17.69±0.23**	**78.99±0.21**	**21.30±0.26**	**76.53±0.25**	**23.48±0.28**	**75.98±0.27**	**20.82±0.24**	**77.17±0.24**

some cases matching or underperforming ERM. In contrast, our method degrades more gradually (17.69% $\rightarrow$ 23.48% on MIMIC-CXR) and remains comparatively stable on CXP (7.51% $\rightarrow$ 8.76%).

Fig. 3 shows the AUC–ΔSep_{10} trade-off. All methods exhibit expected AUC decline with increasing γ, but most fairness baselines incur large performance losses (e.g., JTT: -2.76 ± 0.05 on MIMIC-CXR; -3.62 ± 0.05 on CXP). Our method achieves the most favorable trade-off, consistently attaining the lowest ΔSep_{10} with the smallest AUC drop (-0.34 ± 0.02 on MIMIC-CXR; -0.15 ± 0.03 on CXP). Table 2 evaluates trend affinity and coverage on MIMIC-CXR. Removing either component degrades fairness and slightly reduces AUC, confirming complementarity. Eliminating coverage causes the largest fairness drop (Avg. ΔSep_{10}(%) = 22.80%), while removing trend affinity yields a smaller but consistent degradation (21.25%), both increasingly pronounced under stronger shift. Neither ablation matches the full model. While these results support the proposed components, the method targets first-order linear age–error trends and does not explicitly model nonlinear or multimodal effects. The parameters λ, δ_y, and warm-up duration remain task-dependent, and fixed post-warm-up affinities may miss later-emerging trends.

5 Conclusion

Age-dependent confounding is a critical source of systematic bias in medical image classification, with age-linked prediction-error trends associated with harmful diagnostic disparities under distribution shift. We address this by attenuating linear age–difficulty shortcuts without enforcing strict age invariance. Across two radiology datasets, our framework reduces age-dependent TPR/FPR disparities with minimal AUC impact and provides robust fairness mitigation under increasing train–test age distribution shifts.

Acknowledgements

This work was supported by GSK plc through a Responsible AI Research Grant. The authors also acknowledge the Australian Institute for Machine Learning for providing institutional support and research infrastructure.

Disclosure of Interests

Nikhil Cherian Kurian is supported by an unrestricted research grant from GSK plc. The remaining authors declare no competing interests.

References

1. Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M.S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al.: A closer look at memorization in deep networks. In: International conference on machine learning. pp. 233–242. PMLR (2017)
2. Bhatia, R., Davis, C.: A better bound on the variance. The american mathematical monthly **107**(4), 353–357 (2000)
3. Brown, A., Tomasev, N., Freyberg, J., Liu, Y., Karthikesalingam, A., Schrouff, J.: Detecting shortcut learning for fair medical ai using shortcut testing. Nature communications **14**(1), 4314 (2023)
4. Chu, C., Donato-Woodger, S., Khan, S.S., Shi, T., Leslie, K., Abbasgholizadeh-Rahimi, S., Nyrup, R., Grenier, A.: Strategies to mitigate age-related bias in machine learning: Scoping review. JMIR aging **7**, e53564 (2024)
5. Gao, Y., Hao, J., Zhou, B.: Fairread: Re-fusing demographic attributes after disentanglement for fair medical image classification. Medical Image Analysis p. 103858 (2025)
6. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nat. Mach. Intell. **2**(11), 665–673 (Nov 2020)
7. Giguere, S., Metevier, B., da Silva, B.C., Brun, Y., Thomas, P.S., Niekum, S.: Fairness guarantees under demographic shift. In: International Conference on Learning Representations (2022)
8. Glocker, B., Jones, C., Bernhardt, M., Winzeck, S.: Algorithmic encoding of protected characteristics in chest x-ray disease detection models. EBioMedicine **89** (2023)

9. Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A.: Robust Statistics: The Approach Based on Influence Functions. Wiley (1986)
10. Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H.: The elements of statistical learning: data mining, inference, and prediction, vol. 2. Springer (2009)
11. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
12. Huber, P.J.: Robust estimation of a location parameter. The Annals of Mathematical Statistics **35**(1), 73–101 (1964)
13. 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)
14. Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data **6**(1) (2019). https://doi.org/10.1038/s41597-019-0322-0, cited by: 1067; All Open Access, Gold Open Access, Green Open Access
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
16. Liu, E.Z., Haghgoo, B., Chen, A.S., Raghunathan, A., Koh, P.W., Sagawa, S., Liang, P., Finn, C.: Just train twice: Improving group robustness without training group information. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 6781–6792. PMLR (18–24 Jul 2021), https://proceedings.mlr.press/v139/liu21f.html
17. Menon, A.K., Williamson, R.C.: The cost of fairness in binary classification. In: Conference on Fairness, accountability and transparency. pp. 107–118. PMLR (2018)
18. Murphy, K.P.: Machine learning: a probabilistic perspective. MIT press (2012)
19. Ong Ly, C., Unnikrishnan, B., Tadic, T., Patel, T., Duhamel, J., Kandel, S., Moayedi, Y., Brudno, M., Hope, A., Ross, H., McIntosh, C.: Shortcut learning in medical AI hinders generalization: method for estimating AI model generalization without external data. NPJ Digit. Med. **7**(1), 124 (May 2024)
20. Sagawa*, S., Koh*, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks. In: International Conference on Learning Representations (2020), https://openreview.net/forum?id=ryxGuJrFvS
21. 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)
22. Xu, Z., Li, J., Yao, Q., Li, H., Zhao, M., Zhou, S.K.: Addressing fairness issues in deep learning-based medical image analysis: a systematic review. npj Digital Medicine **7**(1), 286 (2024)
23. Yang, Y., Zhang, H., Gichoya, J.W., Katabi, D., Ghassemi, M.: The limits of fair medical imaging ai in real-world generalization. Nature medicine **30**(10), 2838–2848 (2024)
24. Zafar, M.B., Valera, I., Rogriguez, M.G., Gummadi, K.P.: Fairness constraints: Mechanisms for fair classification. In: Artificial intelligence and statistics. pp. 962–970. PMLR (2017)