Fair Curriculum Learning for Concept Bottleneck Models in Dermatology

Matthew J. Cockayne*, Marco Ortolani, and Baidaa Al-Bander

School of Computer Science and Mathematics, Keele University, Keele, Staffordshire ST5 5BG, United Kingdom
m.j.cockayne@keele.ac.uk

Abstract. Black box deep learning models for dermatology show performance disparities across skin types, with darker-skinned patients experiencing lower diagnostic performance. Concept Bottleneck Models (CBMs) offer interpretability through intermediate concept predictions but lack explicit fairness mechanisms. We introduce Fair Curriculum CBM, a four-phase curriculum learning approach for in-training bias mitigation that, instead of relying on concept difficulty, orders training by fairness objectives, including balanced foundation, demographic parity, equalized odds, and performance parity. This progressive structure trains all concepts jointly while gradually introducing fairness constraints and adversarial debiasing directly, rather than treating fairness as a post-hoc adjustment. On SkinCon (3230 images, 6 Fitzpatrick types), Fair Curriculum CBM improves lowest-group F1 by 63% (0.270→0.441, $p<0.001$) and reduces performance gaps by 44% (0.361→0.203, $p=0.003$) compared to *difficulty-based* curriculum learning, with simultaneous 5.3% overall F1 gain ($p<0.001$). Paired t-tests comparing fairness-first to difficulty-based curriculum approaches show significant improvements on diverse skin types (Types II, V, VI: $p\leq0.005$) without degradation on lighter types. Results from 100 independent runs per model validate that fairness curriculum learning improves rather than constrains performance. Code available at: https://github.com/Matt-Cockayne/FairCBM.

Keywords: Fairness · Concept bottleneck models · Curriculum learning · Medical imaging · Dermatology · Responsible AI

1 Introduction

Artificial intelligence in dermatology shows systematic performance disparities across skin types, with darker-skinned patients experiencing significantly lower diagnostic performance [4]. These disparities stem from dataset bias (medical images datasets over-represent Fitzpatrick I–III), biased feature extraction (deep networks encode protected attributes), and evaluation gaps (overall performance metrics do not consider per-group performance). Fairness-aware training methods [23] often do not consider interpretability, the ability to understand model

* Corresponding author

decisions through intermediate reasoning. In high-stakes medical domains, interpretability and fairness are essential for clinical trust and error diagnosis.

Concept Bottleneck Models (CBMs) [14] address interpretability by introducing an intermediate concept layer between encoder and classifier, where interpretable concepts (e.g., morphological features in dermatology) mediate the mapping from input images to final predictions. This architecture enables concept-based explanations, expert intervention for concept correction, and failure mode debugging through concept analysis. Standard CBMs, however, lack explicit fairness mechanisms, resulting in disparate performance across demographic groups despite interpretable concept predictions.

Curriculum learning [2] structures training from easy to hard examples. Traditional *difficulty-based* curricula order concepts by prediction difficulty for learning efficiency. We reframe this by ordering training by fairness objectives rather than concept difficulty, prioritizing equity alongside interpretability.

We introduce Fair Curriculum CBM, combining CBM interpretability with fairness training. Contributions include: (1) a 4-phase *fairness-first* curriculum (balanced foundation, demographic parity, equalized odds, and performance parity); (2) CBM architecture with fairness loss optimization, and adversarial debiasing; and (3) validation showing 44% gap reduction, 5% F1 gain, and 63% lowest-group improvement compared to a *difficulty-based* curriculum, as well as direct and CBM variant classifiers, on SkinCon [5].

2 Related Work

Fairness in machine learning can be addressed through pre-processing [12], in-processing [23,17,1], or post-processing [8] strategies. Pre-processing methods reweight training data or oversample under-represented groups but risk discarding informative examples and cannot adapt fairness constraints during learning. Post-processing approaches adjust decision thresholds per group after training, offering flexibility but operating on fixed representations that may encode biases. In-processing methods optimize fairness constraints during training, jointly learning fair representations alongside predictive objectives. Adversarial debiasing [23] exemplifies this approach by training a discriminator to predict protected attributes from learned features while gradient reversal [7] encourages the encoder to learn group-invariant representations. Recent work addresses bias amplification [24] and dataset auditing [15]. However, adversarial methods face training instability when fairness constraints are introduced abruptly, motivating gradual constraint introduction.

Concept Bottleneck Models [14] enforce predictions through domain-specific, understandable concepts, enabling concept-based explanations and expert intervention [6], which builds clinical trust in medical domains. Extensions address sequential prediction [18], leakage mitigation [9], intervention [21], trustworthiness [10], medical classification [13,19,3], and theoretical concept set design [20]. Dermatology datasets with dense concept annotations, such as SkinCon [5], enable fine-grained model debugging through morphological attribute labeling. Recent

work addresses fairness in medical imaging [22], though integrating interpretability with fairness during training remains underexplored.

Curriculum learning [2] structures training progressively from easy to hard examples, with applications including example scheduling [11] and concept ordering in CBMs [18]. *Difficulty-based* Curriculum CBMs order concepts by prediction difficulty (training easier concepts like "color" before harder concepts like "texture") for learning efficiency. We leverage a curriculum's progressive structure to address adversarial training instability by ordering training by *fairness objectives* rather than task complexity.

3 Methodology

3.1 Problem Formulation

Given a dataset $\mathcal{D} = \{(x_i, c_i, y_i, a_i)\}_{i=1}^{N}$ where $x_i \in \mathbb{R}^{3\times H\times W}$ are input images, $c_i \in \{0,1\}^K$ are binary concept annotations ($K = 23$ morphological features), $y_i \in \{0,1\}$ are binary malignancy labels, and $a_i \in \{I, II, III, IV, V, VI\}$ are protected attributes (Fitzpatrick skin types), we learn a model $f : \mathcal{X} \to \{0,1\}$ that simultaneously achieves high predictive performance (F1 score), provides interpretable concept predictions, and ensures fairness across all groups such that $F1(a) \approx F1(a_i)$ for all a, a_i.

3.2 Fair Curriculum CBM

The architecture comprises four components (Fig. 1): an encoder extracting feature representations from images; a concept bottleneck predicting 23 interpretable morphological concepts from features; a binary classifier producing malignancy predictions from concept activations; and an adversarial discriminator predicting group membership from features via gradient reversal to learn group-invariant representations.

Unlike a standard *difficulty-based* Curriculum CBM which orders concepts by a given definition of difficulty, Fair Curriculum CBM orders training by fairness objectives while training all concepts jointly. The 4-phase curriculum progressively introduces fairness constraints by coordinating sampling strategies, fairness loss weighting, and adversarial debiasing across epochs Q1: balanced foundation, Q2: demographic parity, Q3: equalized odds with adversarial warmup, and Q4: performance parity with full adversarial debiasing.

The curriculum structure addresses three challenges. First, simultaneous introduction of fairness constraints and adversarial debiasing would cause training collapse [23]; progressive introduction enables incremental adaptation. Second, constraints follow a hierarchy of: representation fairness (balanced sampling), allocation fairness (DP), outcome fairness (EO), and performance fairness (gap minimization), where each level builds on prior constraints [8]. Third, constraint complexity increases from marginal (DP) to conditional (EO) to continuous (performance gap) metrics, mirroring curriculum learning's easy-to-hard progression.

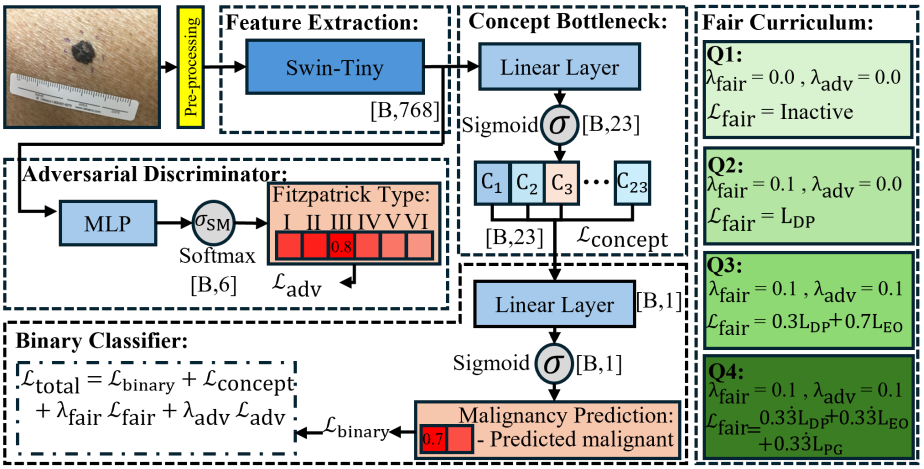

Fig. 1. Fair Curriculum CBM architecture with 4-phase training curriculum. The model coordinates sampling strategies, fairness loss weighting, and adversarial debiasing across phases while training all 23 concepts jointly.

Phase 1: Balanced Foundation (Q1). Equal samples per Fitzpatrick type prevent the encoder from learning group-specific shortcuts early in training when representations are most variable. No fairness penalty is applied, the model trains only on concept and binary classification losses ($\mathcal{L}_{\text{concept}}$ and $\mathcal{L}_{\text{binary}}$: binary cross-entropy over concept predictions and malignancy labels respectively) to establish a bias-free baseline from which fairness constraints can be introduced.

$$\mathcal{L}_{\text{Phase 1}} = \mathcal{L}_{\text{concept}} + \mathcal{L}_{\text{binary}} \tag{1}$$

Phase 2: Demographic Parity (Q2). DP equalizes positive prediction rates across groups via marginal distributions. We compute $\mathcal{L}_{\text{DP}}$ using soft predictions over batch samples:

$$\mathcal{L}_{\text{DP}} = \sum_{a,a'} |\mathbb{E}_{x\sim a}[\sigma(f(x))] - \mathbb{E}_{x\sim a'}[\sigma(f(x))]| \tag{2}$$

$$\mathcal{L}_{\text{Phase 2}} = \mathcal{L}_{\text{concept}} + \mathcal{L}_{\text{binary}} + \lambda_{\text{fair}}\mathcal{L}_{\text{DP}} \tag{3}$$

Phase 3: Equalized Odds (Q3). EO requires equal error rates across groups conditioned on labels, imposing stricter constraints than DP. Stratified sampling ensures equal samples per (group, label) stratum. Adversarial debiasing is introduced here as the model has established stable representations, with warmup preventing destabilization:

$$\lambda_{\text{adv}}(t) = \lambda_{\text{target}} \cdot \frac{t - 0.5T}{0.25T}, \quad t \in [0.5T, 0.75T] \tag{4}$$

Equalized odds computes $\mathcal{L}_{\text{EO}} = \sum_{y,a,a'} |\mathbb{E}[\sigma(f(x))|y,a] - \mathbb{E}[\sigma(f(x))|y,a']|$ over strata. The phase loss combines DP and EO with progressive adversarial

training:

$$\mathcal{L}_{\text{Phase 3}} = \mathcal{L}_{\text{concept}} + \mathcal{L}_{\text{binary}} + \lambda_{\text{fair}}(0.3\mathcal{L}_{\text{DP}} + 0.7\mathcal{L}_{\text{EO}}) + \lambda_{\text{adv}}(t)\mathcal{L}_{\text{adv}} \quad (5)$$

Phase 4: Performance Parity (Q4). While DP and EO ensure fairness in rates, Phase 4 directly optimizes performance gap via soft-F1: $\mathcal{L}_{\text{PG}} = \max_a \tilde{F1}_a - \min_a \tilde{F1}_a$. Error-driven sampling oversamples low-performing groups:

$$w_a = \begin{cases} \frac{1}{F1_a + \epsilon} & \text{if } F1_a \geq 0.1 \\ 0.5 & \text{otherwise} \end{cases} \quad (6)$$

Train data sampling proportions are recomputed every 5 epochs based on per-group performance. This affects only sampling proportions, not gradient computation or checkpoint selection. The final phase combines all fairness objectives with full adversarial debiasing:

$$\mathcal{L}_{\text{Phase 4}} = \mathcal{L}_{\text{concept}} + \mathcal{L}_{\text{binary}} + \lambda_{\text{fair}} \left(\frac{1}{3}\mathcal{L}_{\text{DP}} + \frac{1}{3}\mathcal{L}_{\text{EO}} + \frac{1}{3}\mathcal{L}_{\text{PG}} \right) + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} \quad (7)$$

The adversarial discriminator predicts group membership from encoder features using standard cross-entropy loss. Gradient reversal [7] (identity in forward pass, negated scaled gradient in backward pass) encourages the encoder to learn group-invariant representations that confuse the discriminator.

4 Experimental Setup

The SkinCon dataset [5] contains 3230 dermatological images annotated with 23 morphological concepts (papule, plaque, nodule, ulcer, etc.), binary malignancy labels, and Fitzpatrick skin type annotations (I-VI). We employ a stratified 80/10/10 split preserving both malignancy and Fitzpatrick distributions, yielding 2,584 training images, 323 validation images, and 323 test images.

We compare 5 model variants to isolate the effect of *fairness-first* curriculum design. *Direct* uses standard Swin-Tiny for direct classification without CBM or fairness constraints. *Standard CBM* incorporates a concept bottleneck but applies no curriculum or fairness mechanisms. *Curriculum CBM* orders concepts by difficulty without fairness constraints. Difficulty is determined by ranked concept F1 scores from standard CBM over 100 runs. *Fair Standard CBM* combines CBM with static fairness (fixed λ_{fair}) but no curriculum. *Fair Curriculum CBM* implements the 4-phase fairness-first curriculum with joint concept training.

We employ Swin-Tiny [16] pre-trained on ImageNet. Training uses Adam ($lr = 10^{-4}$, batch size 32, 100 epochs) with phase transitions at epochs 25, 50, 75. Hyperparameters: $\lambda_{\text{fair}} = 0.1$, $\lambda_{\text{adv}} = 0.01$, gradient clipping max norm 1.0. Each model trains with 100 independent random seeds using fixed stratified splits (500 runs total). Experiments run on Intel Xeon Gold 6430 CPUs (64 cores), 512 GB RAM, 4 NVIDIA L40 GPUs (48 GB) using CUDA 12.0.

Performance is measured by overall F1, per-group F1 (Types I–VI), lowest-group F1, performance gap (the difference between the highest and lowest per-group F1 across Fitzpatrick types), and demographic parity (DP) disparity. The checkpoint with the highest validation F1 is selected for test evaluation. Statistical significance uses paired t-tests with Cohen's d across 100 matched runs (same random seeds, initialization, and data splits). Significance threshold α=0.05.

5 Results

Table 1 presents results from 100 independent runs across 5 model variants. Fair Curriculum CBM outperforms all baselines across performance and fairness metrics. Compared to Curriculum CBM (*difficulty-based* ordering), Fair Curriculum CBM reduces performance gaps by 43.8% (0.361 → 0.203, p=0.003, Cohen's d=0.31) while improving overall F1 by 5.3% (p<0.001, d=-0.40), recall by 16.2% ($p<10^{-9}$, d=-0.70), and lowest-group F1 (LG-F1) by 63.3% (p<0.001, d=-0.43). DP disparity remains unchanged (p=0.86), indicating prediction rate balance maintenance while performance equity improves. Fair Curriculum CBM also outperforms Fair Standard CBM (static fairness constraints): +6.1% F1, +74.3% lowest-group F1, -47.7% performance gap, validating progressive curriculum introduction over static fairness application.

Figure 2 presents comprehensive performance analysis. Figure 2(A) shows Fair Curriculum CBM achieves highest F1 scores across all six Fitzpatrick types. Per-Fitzpatrick statistical tests (Fair Curriculum CBM vs. Curriculum CBM) reveal targeted equity improvements: Type II improves +8.5% (0.534→0.580, p=0.000142, Cohen's d=0.40), Type V improves +14.9% (0.558→0.642, p=0.005, d=0.29), and Type VI improves +63.2% (0.270→0.441, p=0.002, d=0.33). Critically, Types I, III, and IV show no significant degradation (p>0.13, Cohen's d<0.15), confirming fairness gains without performance sacrifice on lighter skin types.

Figure 2(B)demonstrates Fair Curriculum CBM leads across three complementary objectives: overall F1 (p<0.001), lowest-group F1 (p<0.001), and performance gap reduction (p=0.003). This constitutes a Pareto improvement, simultaneous gains across all three metrics, challenging the assumed fairness-performance tradeoff in this example. DP remaining stable (p=0.86) while performance gap decreases validates that the method improves performance equity without altering prediction rate balance.

Recall improvements (Table 1) mirror F1 gains, with Fair Curriculum CBM achieving 16.2% higher recall than Curriculum CBM, translating to fewer missed malignancy predictions across all skin types. This addresses the critical clinical concern of false negatives in diverse skin type populations.

Table 3 quantifies each phase's contribution across 100 runs per configuration. Phase 3 (equalized odds + adversarial debiasing) removal causes the largest performance drop (-6.1% F1) and substantial fairness degradation (gap increases 0.203→0.721), validating this as the most critical phase for performance. The large ablation effects and high standard deviations (particularly for LG-F1 and

Table 1. Performance and fairness across 5 models (100 runs). LG-F1: lowest-group F1 . Gap: performance gap. DP: demographic parity disparity. Bold indicates best.

Model	F1	Recall	LG-F1	Gap	DP
Direct	0.539 ± 0.081	0.495 ± 0.103	0.322 ± 0.044	0.267 ± 0.046	0.138 ± 0.042
Standard CBM	0.561 ± 0.076	0.504 ± 0.103	0.257 ± 0.042	0.379 ± 0.044	**0.136 ± 0.036**
Curriculum CBM	0.580 ± 0.074	0.538 ± 0.100	0.270 ± 0.043	0.361 ± 0.045	0.143 ± 0.039
Fair Std. CBM	0.576 ± 0.071	0.534 ± 0.109	0.253 ± 0.040	0.388 ± 0.043	0.140 ± 0.046
Fair Curr. CBM	**0.611 ± 0.088**	**0.625 ± 0.123**	**0.441 ± 0.046**	**0.203 ± 0.048**	0.142 ± 0.050

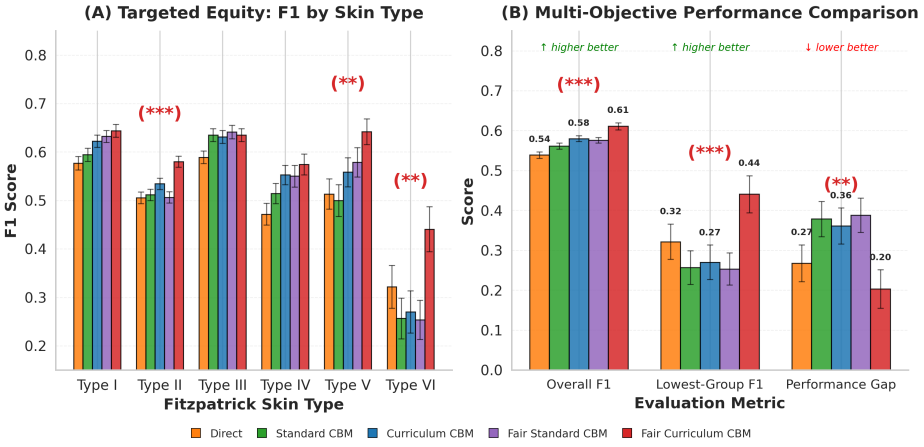

Fig. 2. Performance across 100 runs. **(A)** Per-Fitzpatrick F1 scores. **(B)** Multi-objective comparison. Significance markers indicate paired t-test results comparing Fair Curr. CBM vs. Curriculum CBM: (***) p<0.001, (**) p<0.01.

Gap) reflect training instability. Without fairness mechanisms some random seeds converge to fair solutions while others catastrophically fail on minority class groups. The full model's curriculum structure stabilizes this variability. Adversarial warmup during Phase 3 enables stable group-invariant representation learning, without curriculum structure adversarial methods causes training collapse.

Phase 4 (error-driven sampling) removal produces a large fairness degradation (gap increases to 0.708 ± 0.197) while maintaining reasonable performance (-1.4% F1), highlighting its efficacy as a fairness mechanism when training is stable. Error-driven oversampling of low-performing groups acts as hard example mining for fairness objectives, directly targeting lowest-group performance (0.441→0.147 when removed).

Phases 1-2 show modest individual impact (-5.2% and -3.8% F1 respectively), suggesting they establish a bias-free foundation enabling later optimiza-

Table 2. Per-Fitzpatrick paired t-tests: Fair Curriculum CBM vs. Curriculum CBM (100 paired runs). (***) p<0.001, (**) p<0.01, (*) p<0.05.

Type	Difficulty Curriculum	Fairness Curriculum	Difference	p-value	Cohen's d	Significance
I	0.622	0.644	+0.022	0.130	0.17	
II	0.534	0.580	+0.046	<0.001	0.40	***
III	0.631	0.635	+0.004	0.781	0.03	
IV	0.553	0.574	+0.021	0.316	0.10	
V	0.558	0.642	+0.084	0.005	0.29	**
VI	0.270	0.441	+0.171	0.002	0.33	**

Table 3. Ablation study: phase removal (100-run baseline). LG-F1: lowest-group F1. Gap: performance gap. DP: demographic parity disparity.

Configuration	F1	LG-F1	Gap	DP
Full Model	**0.611 ± 0.088**	**0.441 ± 0.046**	**0.203 ± 0.048**	0.142 ± 0.050
w/o Ph. 1 (bal. init)	0.579 ± 0.076	0.112 ± 0.020	0.718 ± 0.017	0.138 ± 0.041
w/o Ph. 2 (DP)	0.587 ± 0.065	0.115 ± 0.020	0.728 ± 0.019	0.133 ± 0.043
w/o Ph. 3 (EO+adv.)	0.574 ± 0.072	0.092 ± 0.018	0.721 ± 0.017	**0.129 ± 0.041**
w/o Ph. 4 (err.-drv.)	0.602 ± 0.095	0.147 ± 0.024	0.708 ± 0.020	0.132 ± 0.043
w/o Adversarial	0.592 ± 0.067	0.091 ± 0.018	0.756 ± 0.015	0.139 ± 0.042

tion rather than directly driving improvements. Removing Phase 1 degrades lowest-group F1 from 0.441 to 0.112 (-74.6%), demonstrating balanced initialization's importance for protecting vulnerable groups. The progressive structure prevents training instability, despite adversarial training, variance increases only 19% (std: 0.074→0.088) compared to Curriculum CBM baseline.

6 Discussion

Three mechanisms validate our framework design. First, balanced sampling prevents early bias encoding, establishing a fair baseline. Second, progressive constraint introduction (DP→EO→PG) prevents training collapse; Phase 3 ablation confirms that removing EO+adversarial debiasing increases gap from 0.203 to 0.721. Third, error-driven sampling targets lowest-performing groups, validating hard example mining for fairness. The Pareto improvement challenges assumed fairness-performance tradeoffs, while DP stability alongside gap reduction shows these metrics measure complementary aspects of equity.

Limitations include single-dataset evaluation, limited representation of Fitzpatrick Types V and VI, and omitted comparisons with dedicated in-processing fairness methods such as GroupDRO, LAFTR, and CFair. The contribution of this work is the curriculum scheduling structure, specifically the coordinated ordering of sampling strategies, fairness objectives, and adversarial activation across training phases, rather than fairness loss functions. The ablation study (Table 3) supports this design choice. The use of SkinCon is constrained by the requirement for joint concept and protected-attribute annotations. Comparable dermatology benchmarks with equivalent annotation density are limited

in availability, and results are calibrated to this controlled setting. This study reports methodological evidence and does not claim deployment readiness or patient-outcome impact. External validation, clinical assessment, and comparison against dedicated fairness baselines remain necessary future work.

7 Conclusion

Fair Curriculum CBM reframes curriculum learning from task-based to fairness-objective ordering, progressively introducing constraints across four phases. This achieves Pareto improvement over difficulty-based curricula while preserving interpretability, challenging assumed fairness-performance tradeoffs. Statistical validation across 100 runs confirms targeted improvements on diverse skin types without degradation on lighter types. Future work should explore external validation, alignment and stability.

Disclosure of Interests

The authors have no competing interests to declare.

References

1. Amini, A., Soleimany, A.P., Schwarting, W., Bhatia, S.N., Rus, D.: Uncovering and mitigating algorithmic bias through learned latent structure. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. pp. 289–295. ACM, Honolulu HI USA (2019). https://doi.org/10.1145/3306618.3314243
2. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proceedings of the 26th annual international conference on machine learning. pp. 41–48 (2009). https://doi.org/10.1145/1553374.1553380
3. Chowdhury, T.F., Phan, V.M.H., Liao, K., To, M.S., Nguyen, T., Phung, D., Venkatesh, S.: AdaCBM: An adaptive concept bottleneck model for explainable and accurate diagnosis. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 36–46. Springer (2024). https://doi.org/10.1007/978-3-031-72117-5_4
4. Daneshjou, R., Vodrahalli, K., Novoa, R.A., Jenkins, M., Liang, W., Rotemberg, V., Ko, J., Swetter, S.M., Bailey, E.E., Gevaert, O., et al.: Disparities in dermatology ai performance on a diverse, curated clinical image set. Science advances **8**(31), eabq6147 (2022). https://doi.org/10.1126/sciadv.abq6147
5. Daneshjou, R., Yuksekgonul, M., Cai, Z.R., Novoa, R., Zou, J.: SkinCon: A skin disease dataset densely annotated by domain experts for fine-grained model debugging and analysis. In: Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2022), https://proceedings.neurips.cc/paper_files/paper/2022/hash/HASH-Datasets_and_Benchmarks-Conference.html
6. Espinosa Zarlenga, M., Barbiero, P., Ciravegna, G., Marra, G., Giannini, F., Diligenti, M., Shams, Z., Precioso, F., Melacci, S., Weller, A., et al.: Concept embedding models: Beyond the accuracy-explainability trade-off. vol. 35, pp. 21400–21413 (2022), https://proceedings.neurips.cc/paper_files/paper/2022/hash/867c06823281e506e8059f5c13a57f75-Abstract-Conference.html

7. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International conference on machine learning. pp. 1180–1189. PMLR (2015), https://proceedings.mlr.press/v37/ganin15
8. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Advances in Neural Information Processing Systems. vol. 29 (2016), https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html
9. Havasi, M., Parbhoo, S., Doshi-Velez, F.: Addressing leakage in concept bottleneck models. In: Advances in Neural Information Processing Systems. vol. 35, pp. 23386–23397 (2022), https://proceedings.neurips.cc/paper_files/paper/2022/hash/944ecf65a46feb578a43abfd5cddd960-Abstract-Conference.html
10. Huang, Q., Song, J., Hu, J., Zhang, H., Wang, Y., Song, M.: On the concept trustworthiness in concept bottleneck models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 21161–21168 (2024). https://doi.org/10.1609/aaai.v38i19.30109
11. Jiang, L., Zhou, Z., Leung, T., Li, L.J., Fei-Fei, L.: Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In: International conference on machine learning. pp. 2304–2313. PMLR (2018), https://proceedings.mlr.press/v80/jiang18c
12. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowledge and information systems **33**(1), 1–33 (2012). https://doi.org/10.1007/s10115-011-0463-8
13. Kim, I., Kim, J., Choi, J., Kim, H.J.: Concept bottleneck with visual concept filtering for explainable medical image classification. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 225–235. Springer (2023). https://doi.org/10.1007/978-3-031-47401-9_22
14. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: International conference on machine learning. pp. 5338–5348. PMLR (2020), https://proceedings.mlr.press/v119/koh20a
15. Lafargue, V., Claeys, E., Loubes, J.M.: Fairness is in the details: Face dataset auditing. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Lecture Notes in Computer Science, vol. 16022, pp. 299–315. Springer (2025). https://doi.org/10.1007/978-3-032-06129-4_18
16. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021), https://openaccess.thecvf.com/content/ICCV2021/html/Liu_Swin_Transformer_Hierarchical_Vision_Transformer_Using_Shifted_Windows_ICCV_2021_paper
17. Madras, D., Creager, E., Pitassi, T., Zemel, R.: Learning adversarially fair and transferable representations. In: International Conference on Machine Learning. pp. 3384–3393. PMLR (2018), https://proceedings.mlr.press/v80/madras18a.html
18. Marconato, E., Passerini, A., Teso, S.: Glancenets: Interpretable, leak-proof concept-based models. Advances in Neural Information Processing Systems **35**, 21212–21227 (2022), https://proceedings.neurips.cc/paper_files/paper/2022/hash/85b2ff7574ef265f3a4800db9112ce14-Abstract-Conference.html
19. Pang, W., Ke, X., Tsutsui, S., Wen, B.: Integrating clinical knowledge into concept bottleneck models. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 241–251. Springer (2024). https://doi.org/10.1007/978-3-031-72083-3_23

20. Ruiz Luyten, M.: A theoretical design of concept sets: Improving the predictability of concept bottleneck models. In: Advances in Neural Information Processing Systems. vol. 37 (2024), https://proceedings.neurips.cc/paper_files/paper/2024/hash/b5a412531110b92961fa13c90938806a-Abstract-Conference.html
21. Shin, S., Jo, Y., Ahn, S., Lee, N.: A closer look at the intervention procedure of concept bottleneck models. In: International Conference on Machine Learning. pp. 31504–31520. PMLR (2023), https://proceedings.mlr.press/v202/shin23a.html
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), 268 (2024). https://doi.org/10.1038/s41746-024-01276-5
23. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. pp. 335–340 (2018). https://doi.org/10.1145/3278721.3278779
24. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2979–2989. Association for Computational Linguistics, Copenhagen, Denmark (2017). https://doi.org/10.18653/v1/D17-1323