Debiased Noise Editing on Foundation Models for Fair Medical Image Classification

Ruinan Jin[1,2], Wenlong Deng[1,2], Minghui Chen[1,2], and Xiaoxiao Li[1,2]

[1] The University of British Columbia [2] Vector Institute
ruinanjin@alumni.ubc.ca, {dwenlong,chenmh}@student.ubc.ca
xiaoxiao.li@ece.ubc.ca

Abstract. In the era of Foundation Models' (FMs) rising prominence in AI, our study addresses the challenge of biases in medical images while the model operates in black-box (e.g., using FM API), particularly spurious correlations between pixels and sensitive attributes. Traditional methods for bias mitigation face limitations due to the restricted access to web-hosted FMs and difficulties in addressing the underlying bias encoded within the FM API. We propose a D(ebiased) N(oise) E(diting) strategy, termed `DNE`, which generates `DNE` noise to mask such spurious correlation. `DNE` is capable of mitigating bias both within the FM API embedding and the images themselves. Furthermore, `DNE` is suitable for both white-box and black-box FM APIs, where we introduced G(reedy) (Z)eroth-O(rder) (`GeZO`) optimization for it when the gradient is inaccessible in black-box APIs. Our whole pipeline enables fairness-aware image editing that can be applied across various medical contexts without requiring direct model manipulation or significant computational resources. Our empirical results demonstrate the method's effectiveness in maintaining fairness and utility across different patient groups and diseases. In the era of AI-driven medicine, this work contributes to making healthcare diagnostics more equitable, showcasing a practical solution for bias mitigation in pre-trained image FMs. Our code is provided at https://github.com/ubc-tea/DNE-foundation-model-fairness.

Keywords: Truswothy Machine Learning · Fairness · Xray Classification.

1 Introduction

Using pre-trained models or encoders to convert complex input data into vector representations is widely used in computer vision and natural language processing (NLP) [8, 28]. This process transforms the original data into a representative low-dimensional hidden space, preserving information for downstream tasks. With the advancement of Foundation Models (FM), services like Google MedLM [20], Voyage.ai, and ChatGPT provide data embedding services, eliminating the need

Corresponding author: Xiaoxiao Li (xiaoxiao.li@ece.ubc.ca)

for specialized hardware or extensive training. Their effectiveness is especially notable in medical fields, addressing data scarcity and hardware constraints. Clinics can enhance these models with minimal effort by fine-tuning custom classifiers on compact embedding output from FM application programming interface (APIs), which provides machine learning (ML) service as a function without users to train their own models from scratch. However, the inherent biases in the APIs' training data and their model potentially harm marginalized groups by perpetuating gender, race, and other biases, a problem that has been exposed in the NLP field [3, 19]. The biased embedding can compromise models' performance on minority groups [5, 9]. Therefore, it is essential to develop new solutions to attain fairness in medical image embedding from pre-trained FM API.

Studies have been done to address the fairness issues in classification problems using ML models. These methods can be categorized into three approaches: 1)Model-based strategies update or remove bias-related model parameters to mitigate bias, for example, using adversarial methods [1, 12, 24]; or prune parameters that are significant for sensitive attributes (SAs) [25]; or update model to minimize the mutual information between target and SA representations using disentanglement learning [4]. However, the pretrained FM API services offer users very limited control over the model parameters [10, 11]. Thus model-based approaches are infeasible under the constraints. 2)Prediction calibration-based methods analyze the prediction probability distribution of a classifier and apply different thresholds to each subgroup to reduce the discrepancy among subgroups [6, 16, 18]. These methods need custom thresholds for various tasks, classes, and groups, which limits their generalizability. Moreover, a significant amount of validation data is needed to determine the thresholds, rendering them less efficient in medical imaging applications where data scarcity is an unavoidable problem. 3)Data-based strategies alleviate bias at the pre-training stage. Re-distribution and re-weighting methods [17, 18] address unfairness by adjusting the balance of subgroups. The effectiveness of redistributed training data is limited because it can only affect the classification head, not the pre-trained FM API embedding encoder. Recent data editing methods show strong ability to reduce disparity among subgroups by removing sensitive information from the input images [12, 23]. However, can leave the model unchanged, *they fail to address the underlying bias encoded within the black-box model, e.g., FM API*. Moreover, these approaches either depend on disease labels and require significant computational costs. Yao *et al.* introduce an image editing method independent on targeted downstream tasks via sketching [27], but this method suffers from subpar performance as not learnable.

In this work, we take two unique properties of pretrained FM API into consideration. Firstly, we aim to emphasize the need to maintain the flexibility of using FM without limiting it to a specific task while maintaining good utility, thus focusing on the learnable data editing-based method that is independent of the downstream task. We choose to edit on the image space rather than on the FM API embeddings since the former provides better control on medical image fidelity and enables interpretability on the applied editing (as shown in App. B). Secondly, given the constraint against modifying parameters in the pre-trained API, we commit to not changing the FM API model parameters, even with black-box access to the API. To the best of our knowledge, no effective and universal [1] method has been available to address this important, timely, but challenging research question: *How to remove bias on medical images when using their embeddings from a pre-trained FM API for various classification tasks?*

To answer this question, we propose the Debiased Noise Editing, `DNE`, where the resulted edit can be shared across all subjects to eliminate SA-related information for various disease classification tasks. Specifically, we start by using a pre-trained SA classifier, trained with the same FM API embeddings. Then, we optimize a set of learnable parameters (referred `DNE` noise) to be added to the images by confusing the SA classifier. Furthermore, we introduce a greedy zeroth-order optimization strategy, `GeZO`, when APIs restrict gradient propagation in the black-box setting. Lastly, we apply this `DNE` noise to input images to generate fair embeddings and achieve unbiased disease classification across various disease tasks. Extensive evaluations on disease classification tasks show our method's effectiveness in promoting fairness while preserving utility.

2 Method

2.1 Problem Setting

This section outlines the problem of fairness in a binary medical image classification task. Firstly, we define key variables: input images $x \in \mathcal{X}$, binary disease labels $y \in \mathcal{Y} = \{0, 1\}$, and sensitive attribute $a \in \mathcal{A} = \{1, ..., |\mathcal{A}|\}$ (i.e. $|\mathcal{A}| = 2$ for gender with male and female). As patients may not have sensitive and disease labels simultaneously, we denote images with sensitive attribute labels as $\{x_i, a_i\}^N = [X_A, A]$ and images with disease label as $\{x_j, y_j\}^M = [X_T, Y]$, where N and M are the number of samples. Then we denote the FM API as $\phi : \mathcal{X} \to \mathcal{Z}$, to obtain the image embedding $z = \phi(x)$; the disease classifier as $f : \mathcal{Z} \to \mathcal{Y}$ and SA classifier as $g : \mathcal{Z} \to \mathcal{A}$. *We summarize all notations in App. A.*

Fairness Issue: Fig. 1 (a) shows there exists an association between sensitive attributes $\mathcal{A}$ and targets $\mathcal{Y}$ (dashed line). In the Empirical Risk Minimization (ERM) [22], the model $f(\phi(\cdot))$ superficially consider $\mathcal{A}$ as a proxy of $\mathcal{Y}$, causing the issue of fairness. This phenomenon persists even when there is an equal number of data points for each class in both $\mathcal{Y}$ and $\mathcal{A}$, as demonstrated in Sec. 3.1. Our approach aims to address these disparities using *debiased noise edit*.

2.2 Debiased Noise Editing on Image for Fair Disease Classification

In this section, we present our image editing approach aimed at reducing bias via learning a noise tensor with the same dimensionality as the image, which

[1] We refer 'universal' as debiasing API embeddings for various classification tasks.

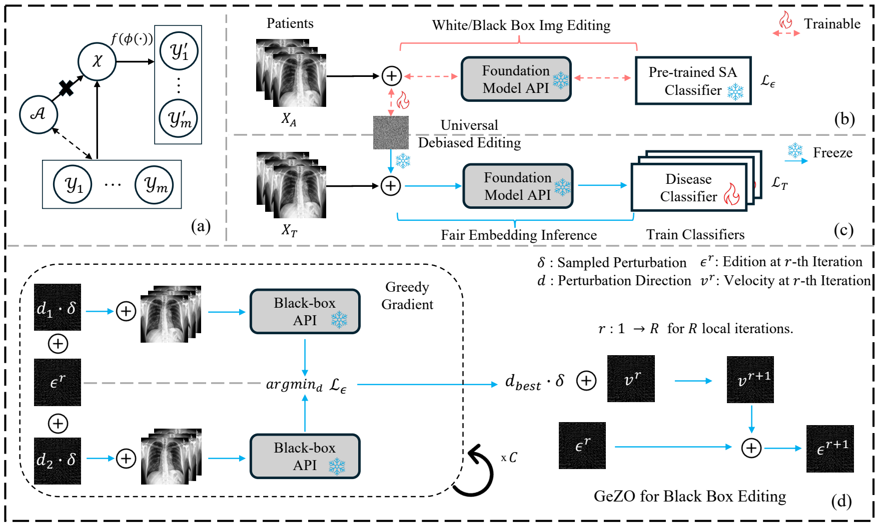

Fig. 1: Overview of debias noise editing pipeline. (a) We eliminate the spurious correlation by breaking the connection between SA $\mathcal{A}$ and image $\mathcal{X}$, ensuring the model relies solely on disease-related information $\mathcal{Y}$. (b) Training DNE noise deceives the pre-trained SA classifier in $\mathcal{A}$ using a frozen FM API and SA classifier, suitable for both white-box and black-box API scenarios based on gradient accessibility. (c) We demonstrate the use of DNE noise: users augment their images with this noise, extract embeddings via the FM API, and proceed to train fair disease classifiers. (d) The (G)reedy (Z)eroth-(O)rder (GeZO) black-box editing method, selects the optimal perturbation via the *Greedy Gradient* process when gradients is inaccessible. It tracks the best perturbation using velocity in each local iteration to update the DNE noise ϵ.

is called DNE noise. Then, we explain how to use this DNE noise for fair disease classification.

Debiased Noise Editing. Given the spurious correlation between SA $\mathcal{A}$ and disease $\mathcal{Y}$, we aim to remove this link through intervention on $\mathcal{X}$ by adding minimal editing vector ϵ, known as DNE, concealing the spurious correlations while preserving image utility. As shown in Fig. 1 (b), we first obtain a pre-trained SA classifier for medical images by either leveraging an existing one (if available) or using the collected $[X_A, A]$ to train the classifier g with cross-entropy loss. Then, we freeze the classifier and leverage gradient ascent to learn the minimal DNE noise ϵ, which is constrained by certain threshold to preserve image fidelity. ϵ is trained to deceive the SA classifier g without compromising useful information by solving the following optimization problem:

$$\mathcal{L}_\epsilon = -\left[\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{\texttt{CE}}(a_i, g(\phi(x_i+\epsilon)))\right] + \lambda||\epsilon||_2 \tag{1}$$

where λ regularizes the magnitude of ϵ. For example, larger λ enforces the small L2-norm of the ϵ in the process of gradient descent. Sec. 3.4 explores its effect in detail. Importantly, we also offer a zero-order strategy for black-box APIs, as detailed in Sec. 2.3.

Fair Disease Classification Having obtained the `DNE` noise ϵ to conceal the SA $\mathcal{A}$, we then freeze and add it to disease classification images X_T. As shown in Fig. 1 (c), this action guarantees that the FM generates a fair embedding devoid of sensitive information, denoted as $\hat{z}_i = \phi(x_i + \epsilon)$. Finally, these fair embeddings are leveraged by the subsequent disease classifier f for prediction. To train the classifier f, we employ a cross-entropy loss:

$$\mathcal{L}_T = \frac{1}{M}\sum_{i=1}^{M}\mathcal{L}_{\texttt{CE}}(y_i, f(\phi(x_i + \epsilon)), \tag{2}$$

where classifier f is the only trainable parameter and M is the size of X_T.

2.3 Greedy Zero-order Optimization for Black-box FM API

Zero-order (ZO) optimizations [2, 14], which avoid the need for gradient computations, offer distinct advantages in optimizing black-box models. However, these methods often exhibit slower convergence rates, particularly in training large FM models. In contrast, our proposed `DNE`, concentrating on the input space, offers a mitigation of this training efficiency issue of ZO optimization.

Inspired by ZO-SGD [21] and recent MeZO [15] that employ in-place perturbation updates, we propose a G(reedy) Z(ero-)O(rder) (`GeZO`) optimization specifically for efficient `DNE`. `GeZO` employs in-place perturbations [15, 21] and greedily updates with the gradient sign that achieves global optimal loss, thereby accelerating the optimization process. As shown in Fig. 1 (d), the core procedure of `GeZO` is the *Greedy Gradient* in the right box, where the gradient of `DNE`'s objective (Eq. 1) is estimated for each local iteration $r \in \{1, ..., |R|\}$ in `GeZO`. *Greedy Gradient* takes the `DNE` noise from the current local iteration (ϵ^r) and estimates its gradient by continuously adding minor perturbations (δ) in different directions (*e.g.*, d_1 and d_2) to it, *i.e.*, $d_1 \cdot \delta$ and $d_2 \cdot \delta$. It then greedily selects the best direction perturbation ($d_{\texttt{best}} \cdot \delta$) that results in the smallest objective in this process. Like MeZO [15], we integrate stochastic sampling for C times. Each time we calculate the objective using a subset of training data to avoid sucking to the local optimum. Once the best direction and perturbation are selected for each local iteration r, we add $d_{\texttt{best}} \cdot \delta$ to the current velocity, v^r. This velocity keeps track of the accumulated updating direction and magnitude for ϵ for each local iteration r, where the initial value for velocity v^1 is 0. As shown in the right part of Fig. 1 (d), v^{r+1} is updated by adding the best perturbation returned by the *Greedy Gradient*. Finally, the `DNE` noise ϵ^{r+1} is updated through adding the v^{r+1}. The local iterations $|R|$ is a hyper-parameter, where a smaller iteration increases the efficiency at the cost of the more biased gradient estimation. We investigate the effect of different $|R|$ in Sec. 3.4. In actual implementation, we also introduced a momentum to update the velocity to accelerate the convergence.

While the key idea is presented in this section, we provide a detailed step-by-step algorithm box for `GeZO` in App. D.

3 Experiement

3.1 Settings

Dataset. To demonstrate the generalizability of `DNE`. we adopt the CheXpert dataset [7], a chest X-ray dataset with multiple disease labels, to predict the binary label for *Pleural Effusion*, *Pneumonia* and *Edema* individually in chest radiographs. All the ablation studies are performed on *Pleural Effusion* classification task for space limit. We take gender bias (male and female) as an example due to its broad impact on society and medical imaging analysis. To demonstrate the effectiveness of bias mitigation methods, we follow [4] to amplify the training data bias for each disease by (1) firstly dividing the data into different groups according to the SA; (2) secondly calculating the positive rate of each subgroup; (3) sampling subsets from the original training dataset and increase each subgroup's bias gap (more positive sample in a subgroup). Then, we sample testing data with the same procedure, but achieve an equal subgroup bias gap. The detailed data distribution is shown in Table 3 in App. C.
Evaluation metrics. We use the classification accuracy to evaluate the utility of classifiers on the test set. To measure fairness, we employ *equal opportunity (EO)* [6] and *disparate impact (DI)* [13] metrics. EO aims to ensure equitable prediction probabilities across different groups, defined by the sensitive attribute a^j, for a given class y^j. It quantifies the disparity in true positive rates between groups: $\mathrm{EO}_{Y=y_1} = P(\hat{Y} = y^1|Y = y^1, A = a^1) - P(\hat{Y} = y^1|Y = y^1, A = a^2)$, where a smaller gap signifies greater equality of opportunity. DI evaluates the presence of indirect discrimination by measuring the ratio of positive predictions across different groups: $\mathrm{DI} = \frac{P(\hat{Y}=1|A=a^1)}{P(\hat{Y}=1|A=a^2)}$. DI closing to 1 indicates the minimal disparity in positive prediction rates between the groups. To align with EO metric, we employ the $|1 - \mathrm{DI}|$ to quantify the fairness in percentage, with smaller values denoting greater fairness.

3.2 Implementation Details

Architecture. In our implementation, all methods use the same architecture. To simulate the FM API, we utilize a publicly available self-supervised pre-trained ViT-base model [2], optimized on X-ray images, given its superior performance metrics among all other architectures [26]. Within our study, we fix the encoder of the ViT model and only fine-tune the classification head, simulating the configuration detailed in Sec. 2.1.
Debiased Noise Editing. For `DNE`, we first fine-tune the SA classifier, g, using the feature encoded by the FM ($\phi(x_i)$) with the training set of *Pleural Effusion*

[2] `https://github.com/lambert-x/Medical_MAE`

Table 1: Comparision of binary prediction of *Pleural Effusion, Pneumonia, and Edema*. We label the best performance in **bold** and the second-best performance with underline. All the values are in percentage.

Diseases	Pleural Effusion				Pneumonia				Edema			
	$EO_n \downarrow$	$EO_p \downarrow$	$\lvert$1-DI$\rvert \downarrow$	Acc $\uparrow$	$EO_n \downarrow$	$EO_p \downarrow$	$\lvert$1-DI$\rvert \downarrow$	Acc $\uparrow$	$EO_n \downarrow$	$EO_p \downarrow$	$\lvert$1-DI$\rvert \downarrow$	Acc $\uparrow$
ERM	40.5	57.0	58.5	<u>72.9</u>	70.0	70.0	74.5	59.6	42.0	39.0	42.0	74.5
Sketch [27]	44.0	52.0	57.5	66.3	64.0	61.0	72.6	56.3	44.0	**15.0**	**15.5**	67.0
Group DRO [18]	41.0	58.0	59.5	72.3	60.0	56.0	64.4	61.0	40.5	40.5	43.3	74.8
Batch Samp. [17]	41.5	50.5	51.8	74.3	64.0	72.0	76.6	60.5	39.0	43.0	44.8	**76.0**
BiasAdv [12]	39.0	54.5	58.2	71.1	<u>36.0</u>	54.0	77.1	61.0	39.0	33.5	36.6	74.9
DNE	**28.5**	**23.0**	**25.6**	**75.1**	38.0	**25.0**	<u>34.7</u>	<u>61.3</u>	**27.5**	17.0	19.7	<u>75.6</u>
DNE-GeZO	<u>37.0</u>	<u>25.0</u>	<u>28.5</u>	72.8	**35.0**	<u>27.0</u>	**33.7**	**61.5**	<u>34.0</u>	<u>15.5</u>	<u>17.2</u>	75.1

in Table 3. We update it using Adam with a learning rate (lr) of 0.0001 for 50 epochs. Second, we update the `DNE` through implementing Eq. 1 by initializing ϵ as a trainable PyTorch parameter with all entries initially set to zero. The parameter matches the input image dimensions of 224×224. The `DNE` can be optimized using classic gradient descent or `GeZO`. Finally, we fine-tune the disease classifier, f, with ϵ added on the input data as delineated in Eq. 2. We update it using AdamW with lr of 1.25×10^{-4} for 50 epochs. We provide the visualization of the magnitude of `DNE` as an interpretation of `DNE` in App. B Fig. 3, where `DNE` is smoothed by the Gaussian kernel. As shown, larger noises are added to the bottom to discriminate the gender-related features.

GeZO. Here, we keep all parts for `DNE` the same as above, except using `GeZO` to update the edit rather than the Adam, given that the gradient is not accessible. The only key parameter that we are interested in is the local iteration T, where we investigated it in Sec. 3.4. Furthermore, the detailed implementation of the algorithm and hyperparameter setting for `GeZO` can be found in App. D.

3.3 Comparison with Baselines

Quantitative Analysis. Table 1 summarizes the performances among different diseases and baselines. The baselines include *model-based* strategy, like biasAdv [3] [12]; data-based strategies, like batch sampling [17] with data redistribution and sketch [27] with data generation; and prediction calibration-based strategy, like Group DRO [18]. The details of these baselines are introduced in Sec. 1. The results indicate that the `DNE` effectively balances fairness and utility. Taking *Pleural Effusion* as the example to analyze, the white-box optimized `DNE` outperforms all baselines. The EO_p and DI decrease over 25% compared to all the baselines, an indication of less disparity between the two genders. Additionally, the utility not only maintains but also surpasses all other groups, e.g., `DNE`'s accuracy (Acc) is higher than 2.2% compared to ERM. Given we are using a balanced testing set, this increase is also a sign of better generalization.

[3] In BiasAdv's implementation, we treat FM API as a white box as it requires access to FM's gradient and there is no zeroth-order optimization for it yet.

Similar for *Pneumonia* and *Edema*, where `DNE` and `DNE-GeZO` takes the best two performance for almost all entries. Although learned through black-box optimization, our `DNE-GeZO` achieves comparable performance to the standard optimization used in `DNE`. This demonstrates not only the validity of `GeZO` but also the efficiency afforded by the small number of edition parameters [15]. In *Pneumonia*, `DNE-GeZO` even slightly surpasses the performance of `DNE` across most metrics. These findings affirm the efficacy of `DNE`, where it not only facilitates fair medical image embedding and training but also introduces better generalizability in downstream classification tasks.

3.4 Ablation Studies

Effect of Regularization Coefficient. To investigate the effect of different regularization coefficients' (λ) effect, we vary λ from 0 to 1. Fig. 2 (a) depicts the EO and the Acc for different λs. The metrics for ERM are labeled as horizontal dashed lines for convenience. As shown, both the fairness and utility metrics remain relatively stable for $\lambda < 1$, consistently surpassing the ERM baseline. However, as λ increases to 1, we observe a marked decline in accuracy below the ERM benchmark, along with a notable increase in the EO metric for both classes.

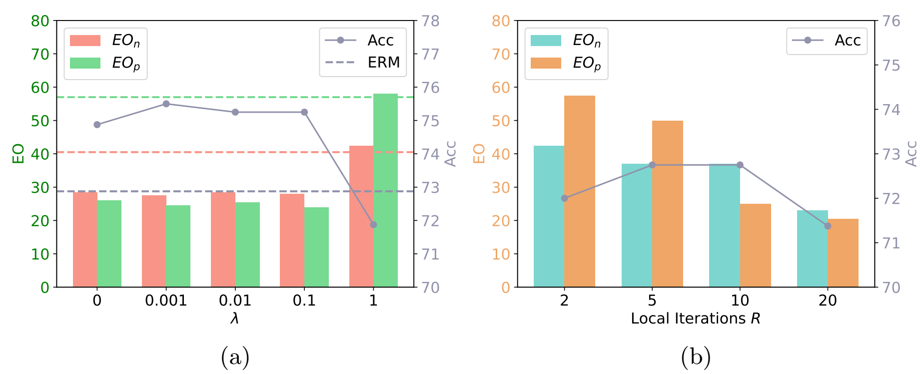

Fig. 2: Ablation study of our methods: (a) Effect of different λ; (b) Effect of different local epochs using `GeZO`.

Effect of Local Iterations in `GeZO`. In black-box API, total local iterations R in `GeZO` affect the optimization performance, wherein larger R leads to more accurate optimization at the cost of efficiency. Here, we examine the effect of changing R, as shown in Fig. 2 (b). For fairness metrics, increasing R from 2 to 20 significantly reduces the EO score for both classes, demonstrating a considerable debiasing impact with more local epochs. This effect is attributed to increased perturbation sampling that expands the search space with more local epochs, as introduced in Sec. 2.3. Meanwhile, accuracy remains relatively stable, with minor fluctuations between 72% and 73%.

4 Conclusion

In this study, we address a crucial, yet under-explored aspect of health equity—the inherent bias in FM API's usage for classification—through the introduction of *debiased noise editing*. `DNE` effectively masks bias-inducing pixels, enhancing fairness in API-generated embeddings. Furthermore, `GeZO` tackles the challenge of the inaccessibility of the gradient in black-box APIs by estimating gradients via perturbation. Future research will extend `DNE`'s application across various FM APIs and settings, aiming to solidify its role in promoting fairer machine learning practices.

Acknowledgments. This work is supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC), Compute Canada, and Vector Institute.

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

References

1. Adeli, E., Zhao, Q., Pfefferbaum, A., Sullivan, E.V., Fei-Fei, L., Niebles, J.C., Pohl, K.M.: Representation learning with statistical independence to mitigate bias. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2513–2523 (2021)
2. Chen, P.Y., Zhang, H., Sharma, Y., Yi, J., Hsieh, C.J.: Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In: Proceedings of the 10th ACM workshop on artificial intelligence and security. pp. 15–26 (2017)
3. Deng, W., Chen, B., Li, X., Thrampoulidis, C.: Content conditional debiasing for fair text embedding. arXiv preprint arXiv:2402.14208 (2024)
4. Deng, W., Zhong, Y., Dou, Q., Li, X.: On fairness of medical image classification with multiple sensitive attributes via learning orthogonal representations. In: International Conference on Information Processing in Medical Imaging. pp. 158–169. Springer (2023)
5. Glocker, B., Jones, C., Bernhardt, M., Winzeck, S.: Algorithmic encoding of protected characteristics in image-based models for disease detection. arXiv preprint arXiv:2110.14755 (2021)
6. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems 29 (2016)
7. 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)
8. Jaiswal, A., Liu, S., Chen, T., Wang, Z., et al.: The emergence of essential sparsity in large pre-trained models: The weights that matter. Advances in Neural Information Processing Systems 36 (2024)
9. Jin, R., Xu, Z., Zhong, Y., Yao, Q., Dou, Q., Zhou, S.K., Li, X.: Fairmedfm: Fairness benchmarking for medical imaging foundation models. arXiv preprint arXiv:2407.00983 (2024)
10. 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 pp. 1–9 (2024)
11. 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)
12. Lim, J., Kim, Y., Kim, B., Ahn, C., Shin, J., Yang, E., Han, S.: Biasadv: Bias-adversarial augmentation for model debiasing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3832–3841 (2023)
13. Lipton, Z., McAuley, J., Chouldechova, A.: Does mitigating ml's impact disparity require treatment disparity? Advances in neural information processing systems 31 (2018)
14. Liu, S., Chen, P.Y., Kailkhura, B., Zhang, G., Hero III, A.O., Varshney, P.K.: A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications. IEEE Signal Processing Magazine 37(5), 43–54 (2020)
15. Malladi, S., Gao, T., Nichani, E., Damian, A., Lee, J.D., Chen, D., Arora, S.: Fine-tuning language models with just forward passes. Advances in Neural Information Processing Systems 36 (2024)
16. 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)
17. Puyol-Antón, E., Ruijsink, B., Piechnik, S.K., Neubauer, S., Petersen, S.E., Razavi, R., King, A.P.: Fairness in cardiac mr image analysis: an investigation of bias due to data imbalance in deep learning based segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24. pp. 413–423. Springer (2021)
18. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. arXiv preprint arXiv:1911.08731 (2019)
19. Schramowski, P., Turan, C., Andersen, N., Rothkopf, C.A., Kersting, K.: Large pre-trained language models contain human-like biases of what is right and wrong to do. Nature Machine Intelligence 4(3), 258–268 (2022)
20. Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., et al.: Towards expert-level medical ques-

tion answering with large language models. arXiv preprint arXiv:2305.09617 (2023)
21. Spall, J.C.: Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. IEEE transactions on automatic control 37(3), 332–341 (1992)
22. Vapnik, V.: Principles of risk minimization for learning theory. Advances in neural information processing systems 4 (1991)
23. Wachinger, C., Rieckmann, A., Pölsterl, S., Initiative, A.D.N., et al.: Detect and correct bias in multi-site neuroimaging datasets. Medical Image Analysis 67, 101879 (2021)
24. Wadsworth, C., Vera, F., Piech, C.: Achieving fairness through adversarial learning: an application to recidivism prediction. arXiv preprint arXiv:1807.00199 (2018)
25. Wu, Y., Zeng, D., Xu, X., Shi, Y., Hu, J.: Fairprune: Achieving fairness through pruning for dermatological disease diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 743–753. Springer (2022)
26. Xiao, J., Bai, Y., Yuille, A., Zhou, Z.: Delving into masked autoencoders for multi-label thorax disease classification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3588–3600 (2023)
27. Yao, R., Cui, Z., Li, X., Gu, L.: Improving fairness in image classification via sketching. arXiv preprint arXiv:2211.00168 (2022)
28. Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., Zhang, K., Ji, C., Yan, Q., He, L., et al.: A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. arXiv preprint arXiv:2302.09419 (2023)