XTranPrune: eXplainability-aware Transformer Pruning for Bias Mitigation in Dermatological Disease Classification

Ali Ghadiri[1(✉)], Maurice Pagnucco[1], and Yang Song[1]

University of New South Wales, Sydney, Australia
a.ghadiri@unsw.edu.au

Abstract. Numerous studies have demonstrated the effectiveness of deep learning models in medical image analysis. However, these models often exhibit performance disparities across different demographic cohorts, undermining their trustworthiness in clinical settings. While previous efforts have focused on bias mitigation techniques for traditional encoders, the increasing use of transformers in the medical domain calls for novel fairness enhancement methods. Additionally, the efficacy of explainability methods in improving model fairness remains unexplored. To address these gaps, we introduce **XTranPrune**, a bias mitigation method tailored for vision transformers. Leveraging state-of-the-art explainability techniques, XTranPrune generates a pruning mask to remove discriminatory modules while preserving performance-critical ones. Our experiments on two skin lesion datasets demonstrate the superior performance of XTranPrune across multiple fairness metrics. The code can be found at https://github.com/AliGhadirii/XTranPrune.

Keywords: Bias mitigation · Explainable AI · Transformer pruning.

1 Introduction

Recent developments in deep learning-based models have demonstrated remarkable performance in disease diagnosis. Nevertheless, these black-box models cannot be trusted to be utilized in clinical settings if their trustworthiness is still questioned. The data-driven nature of deep learning models makes them vulnerable to learning patterns based on sensitive attributes such as demographic or ethnic information in the input in pursuit of better performance. This problem is more prevalent in the medical domain since these sensitive attributes are inherently embedded in the input image, increasing the odds of learning unjust patterns.

There has thus been a surging interest in investigating various ways to eliminate bias in medical deep learning models for various applications and data modalities such as brain Magnetic Resonance Imaging (MRI) [12,16,1], and skin lesion images [5,6]. Many of these debiasing methods try to either balance the input data according to the sensitive attribute or introduce new fairness-related

constraints to the training objective to promote just predictions. By convention, we can categorize these methods as *pre-processing* and *in-processing* techniques, respectively [20]. The major drawback of such approaches is that they require the current high-performing models to be retrained with new settings, which is time-consuming, computationally expensive, and incurs additional costs.

On the other hand, while current debiasing methods [21,10,14] are focused on convolutional neural networks, *Transformers* are now widely used in many medical computer vision tasks [15]. Given their complexity, it is more crucial than ever to find an efficient approach to eliminate the retraining hurdle proposed in the current debiasing methods to accomplish fairness. A viable solution to this problem is the third branch of debiasing techniques, i.e., *post-processing* methods, in which we try to calibrate the model to reduce the impact of unfair nodes [19,10]. The primary advantage of these methods is that they can overcome bias by *pruning* the discriminatory nodes of a pre-trained model rather than having to train it with fairness-aware constraints from scratch. However, the main limitation of existing methods is that they mainly make a general modification in all model nodes by introducing a constraint on the objective function to either adjust the model parameters or learn a fairness pruning mask. We advocate that the primary cause of bias in the model must be tackled with a more precise approach that focuses on identifying the *discriminatory nodes* to achieve more targeted bias mitigation.

Hereby, we introduce a simple and intuitive e**X**plainability-aware **Tran**sformer **Prun**ing (**XTranPrune**) method for bias mitigation. It utilizes an explainability method based on relevance propagation to prune transformer-based models for fairness enhancement. More precisely, we take advantage of the attribution vectors generated by an explainability method to identify the nodes in the transformer encoder that are causing the discriminatory predictions concerning the sensitive attribute, which are then pruned to mitigate bias. The process also factors in the contribution of nodes in the main classification task to avoid pruning the essential nodes for classification, minimizing the performance drop after pruning. The prime merit of using explainability methods to evaluate a node's contribution to the outcome over conventional approaches such as using the transformer's attention map or the nodes' gradient is that it provides a more precise representation of the contribution of each node in the final prediction, improving the accuracy of identifying nodes that cause unjust outcomes. Furthermore, instead of using a global constraint that affects all of the model's parameters, this approach will facilitate targeted debiasing according to the identified source of bias, by only adjusting the discriminatory nodes.

We have evaluated our method on two skin lesion datasets, which is a common benchmark for bias in classification, using a variety of fairness evaluation metrics. We also introduce a new fairness assessment metric called NFR based on the macro-averaged F1 score gap between the subgroups to provide a better evaluation of the model's performance. Our experiments demonstrate that **XTranPrune** outperforms all the state-of-the-art bias mitigation methods in dermatological data analysis in most of the metrics taken into account.

2 Methodology

2.1 Problem definition

Assume that each record of our dataset is represented by the input image x_i, its classification label y_i, and the associated sensitive attribute s_i such as gender, skin tone, age, etc. Our prime objective is to classify the label as accurately as possible while ensuring that the model's performance does not show a noticeable discrimination concerning each sensitive attribute. This means that the model should not rely on the features in the input related to the sensitive attribute, for example, the skin colour of the patient, to make its prediction. In the medical domain, two fundamental definitions have been proposed for fairness, group fairness and Min-Max fairness [22]. In the former, we aim to minimize the disparity in the models performance among subgroups, while in the latter our chief target is to maximize the worst performance of the model among subgroups, also known as worst-case scenario. We will showcase how XTranPrune effectively reduces bias according to both definitions.

2.2 XTranPrune

We introduce **XTranPrune**, a novel explainability-based pruning method to eliminate bias in vision transformers. Our method comprises two branches, each employing the Data-efficient Image Transformer (DeiT) [18] model as the image encoder. The Main Branch is responsible for the primary classification task, i.e., skin lesion diagnosis. The Sensitive Attribute (SA) Branch aims to detect sensitive attributes such as skin tone to facilitate *discriminatory nodes* identification. *Discriminatory nodes* are defined as those with significant contributions to detecting sensitive attributes in the SA branch while exhibiting low attributions in the main branch, thereby preserving classification performance.

Figure 1 illustrates the workflow of **XTranPrune**. Firstly, the main and SA branches are trained separately, with skin diagnosis and skin tone as target labels, respectively. The transformer explainability method is then applied to both networks to calculate nodes' *attribution vector*. The *attribution vector* of the main network is then filtered, selecting the least contributing nodes to the main classification task to generate the *Performance Mask*, protecting crucial nodes from pruning. Subsequently, *SA Attributions* are filtered using the *Performance Mask*, and the resulting *Masked SA Attributions* are used to create the *explainability-aware pruning mask*. This mask is iteratively applied to prune discriminatory nodes in the main branch, reducing the network's tendency for biased predictions. During inference, the pruned main branch network is used to classify input test images.

Explainability-aware Node Attribution Calculation. We utilize the state-of-the-art transformer explainability method [4] to generate the pruning mask by leveraging the calculated node attributions. This method has demonstrated superiority over conventional visual explanation methods such as GradCAM [13], and the Layer-wise Relevance Propagation (LRP) [2] in extensive experiments on

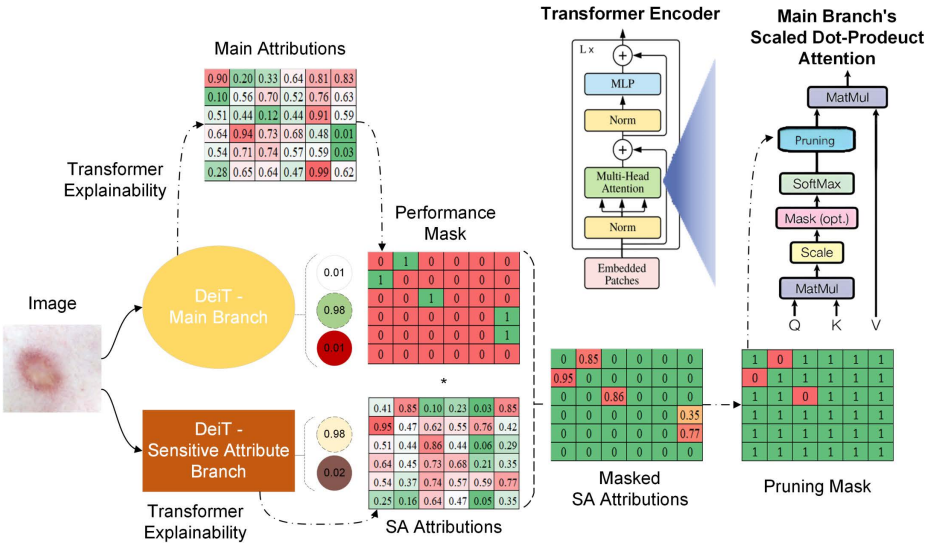

Fig. 1. Overview of **XTranPrune**: It consists of two branches, the *Main Branch* and the *SA Branch*. We utilise an explainability method to find the nodes' attribution in both branches. The calculated attributions allow us to generate the *Performance Mask* to keep the most important nodes, and then the *Pruning Mask* to prune the most discriminatory nodes in the *Main Branch*.

multiple datasets. Inspired by LRP, this method [4] adapts relevance propagation to the ViT architecture and introduces a new definition for node attribution. Specifically, it computes the element-wise product of the propagated relevancy matrix and the gradient of the respective node in the encoder.

Formally, for each encoder block b in ViT we derive its attention map $A^{(b)}$, gradients $\nabla A^{(b)}$, and relevance matrix $R^{(n_b)}$, where n_b is the layer corresponding to calculating the attention map using the *softmax* function. Let s be the total number of input tokens to the ViT, and h be the number of heads in each block. We calculate the attribution matrix of each block as $Attr^{(b)} \in R^{h \times s \times s}$:

$$Attr^{(b)} = \left(\nabla A^{(b)} \odot R^{(n_b)}\right)^{+} \tag{1}$$

The attribution matrix can be filtered to only retain positive contributions of the nodes in block b to the network's final prediction.

In **XTranPrune**, we consider all $s \times s$ nodes after the *softmax* function in block b, head h of the DeiT as *nodes*. The associated parameter vectors with these nodes in DeiT is the attention map $A^{(b)}$. We then represent the contribution of these nodes as $Attr^{(b)}$ using the explainability method and then prune these nodes according to the pruning algorithm as described in the following section.

Attribution-based Pruning. During each pruning iteration, both branches are fed with mini-batches of the training set to calculate node attributions using the explainability method. To enhance robustness, this process is repeated across

multiple batches, and average attributions are computed. Our pruning method leverages these attribution vectors to iteratively generate an explainability-aware pruning mask, aiming to identify the discriminatory nodes concerning the sensitive attribute while preserving nodes critical for the primary classification task in the main branch to maintain the model's performance.

To achieve this, we first generate a *Performance Mask* to filter nodes in the main branch based on attribution vectors, retaining nodes with maximal impact on classification performance. A hyperparameter, *retain_rate* controls the rate of node retention. Next, we compute the element-wise product of the *Performance Mask* and the attribution vector in the sensitive attribute branch to obtain the *Masked SA Attributions*. This filtered vector identifies nodes with the highest contribution to the sensitive attribute classification for pruning in the main branch network, reducing bias in predictions since the effect of discriminatory nodes in the network's predictions has been excluded. The *pruning_rate* hyperparameter determines the number of nodes pruned in the final Pruning Mask. The resulting vector after the *softmax* layer in each *Scaled Dot-Product Attention* module of the main branch network is pruned with this *Pruning Mask*. After pruning, we evaluate the Equality of Opportunity (EOM) metric on the validation set, iterating until no further progress in fairness is observed. EOM is chosen as we believe it is the most comprehensive fairness assessment metric. A detailed iteration of *XTranPrune* is provided in the supplementary material.

3 Experiments

3.1 Datasets and Evaluation Metrics

We have examined the effectiveness of our method for bias mitigation on two dermatological datasets including Fitzpatrick17k [8,7] and PAD-UFES-20 [11]. The Fitzpatrick17k dataset comprises 16,577 skin images of various organs, representing 114 skin conditions. We use the high-level 3-class labels, representing the severity of the condition. The PAD-UFES-20 dataset contains 2,298 images accounting for 6 diagnosis classes. In both datasets, skin tone is the sensitive attribute, which is annotated for each image with a Fitzpatrick scale ranging from 1 (light) to 6 (dark). A fair classification means that the method should achieve similar classification performance for different skin tones. Similar to other studies, the dataset is divided into training (80%) and testing (20%) sets, stratified by skin condition. Additional details about datasets and example images of different classes and skin tones can be found in the supplementary material.

To compare the performance of models in the main classification task, we compare the macro-averaged F1 score. Having imbalanced datasets, we opted for this metric to treat all classes equally, avoiding the frequent class to overpower the evaluation metric. We consider both fairness definitions in our experiments to provide a thorough evaluation. In terms of *Min-Max fairness*, we report the worst-case F1 score across subgroups. With *Group fairness*, we report 4 widely used fairness metrics, namely, Demographic disParity across Multiple subclasses (DPM), Equality of Opportunity across Multiple subclasses (EOM), Equalized

Opportunity (EOpp), Equalized Odds (EOdd). For metrics with binary subgroups only, we classify skin tones rated 1-3 as unprivileged and those rated 4-6 as privileged. Additionally, we propose a new fairness metric called Normalized F1 score Range (NFR), which measures the macro-averaged F1 score disparity across all sensitive attributes relative to their mean. NFR offers the advantage of jointly considering classification performance and fairness. In addition, using the F1 score instead of accuracy facilitates a more balanced performance assessment. Metrics formulas are provided in the supplementary material.

3.2 Results

We compare XTranPrune with the state-of-the-art bias mitigation methods specifically designed for medical data, including FairDisCo [6] that represents in-processing methods with the combination of contrastive and disentanglement learning, FairPrune [19], the state-of-the-art post-processing pruning method for bias mitigation, and FairME [5], the most recent multi-exit debiasing framework. In addition, we compare our method and the most recent benchmark in fairness evaluation, MEDFAIR [22]. Furthermore, within each bias mitigation category, we have selected top-performing candidates: LAFTR [9] for adversarial training, EnD [17] for disentanglement, and SWAD [3] for domain generalization methods. We also include two baseline models: the main branch DeiT model before pruning and a ResNet18 model, the baseline for the aforementioned debiasing methods. To combat dataset imbalance, we employ a weighted random sampler to maintain an even distribution of samples for every skin condition within a mini-batch. In both datasets, the SA Branch classifies subgroups into privileged and unprivileged, while the main branches classify skin conditions into 3 and 6 classes for Fitzpatrick17k and PAD-UFES-20 datasets, respectively.

Fitzpatrick17k. Our experiments on the Fitzpatrick17k dataset are summarized in Table 1. Regarding *Min-Max fairness*, our method exhibits a significant increase of nearly 11% in the worst-case F1 score among sensitive attribute subgroups compared to our baseline. Compared to other bias mitigation methods, XTranPrune demonstrates a notable improvement of 4.3% in this metric. In terms of *Group fairness*, XTranPrune enhances almost all included metrics by a

Table 1. Fitzpatrick17k dataset results.

Model	F1 score (%)	Worst-case F1 score (%)↑	DPM↑	EOM↑	EOpp0↓	EOpp1↓	EOdd↓	NFR↓
FairDisCo	75.97	64.81	0.478	0.622	**0.048**	0.142	0.104	0.295
FairPrune	61.55	57.86	0.497	0.707	0.138	0.094	0.232	0.125
FairME	73.12	60.66	0.579	0.668	0.100	0.134	0.221	0.202
EnD	68.82	64.64	0.507	0.718	0.122	0.091	0.150	0.137
LAFTR	64.88	57.25	0.476	0.626	0.138	0.149	0.287	0.157
SWAD	65.50	54.10	0.451	0.608	0.154	0.131	0.285	0.233
Baseline (Resnet18)	74.26	68.46	0.445	0.639	0.057	0.206	0.191	0.157
Baseline (DeiT)	76.06	58.58	0.527	0.628	0.092	0.133	0.195	0.314
XTranPrune	73.51	**69.13**	**0.586**	**0.790**	0.086	**0.066**	**0.095**	**0.114**

Table 2. PAD-UFES-20 dataset results.

Model	F1 score (%)	Worst-case F1 score (%)↑	DPM↑	EOM↑	EOpp0↓	EOpp1↓	EOdd↓	NFR↓
FairDisCo	62.82	33.85	0.018	0.410	0.180	1.653	1.657	0.999
FairPrune	45.46	18.43	**0.049**	0.454	0.490	1.288	1.393	1.663
FairME	55.68	27.36	0.009	0.376	0.372	1.254	1.595	1.252
EnD	63.08	28.19	0.018	0.365	0.201	1.777	1.729	1.172
LAFTR	57.82	29.33	0.021	0.351	0.323	1.697	1.720	1.142
SWAD	56.34	32.54	0.018	0.330	0.222	1.725	1.758	1.129
Baseline (Resnet18)	63.67	26.03	0.009	0.369	**0.117**	1.148	1.903	1.198
Baseline (DeiT)	62.89	55.84	0.009	0.562	0.493	1.504	1.345	0.592
XTranPrune	62.01	**57.03**	0.009	**0.624**	0.389	**1.141**	**0.909**	**0.587**

substantial margin compared to both our baseline and state-of-the-art bias mitigation methods. Regarding Eopp1, XTranPrune achieves less than half of our baseline value and outperforms the rest of the competing methods. Additionally, with our proposed fairness metric, NFR, XTranPrune achieves the lowest value compared to other models, indicating its effectiveness in reducing the performance gap among subgroups while maintaining overall performance.

PAD-UFES-20. A summary of our results on the PAD-UFES-20 dataset is depicted in Table 2. Incorporating vision transformers leads to a significant improvement in *Min-Max fairness*. Our debiasing approach further enhances the worst-case F1 score to 57.03%. In terms of *Group fairness*, XTranPrune outperforms the state-of-the-art by a significant margin, with a 0.21 increase in EOM. Similarly, the result for EOpp1 illustrates superior fairness compared to competing methods and our baselines. Additionally, XTranPrune achieves a notable improvement in EOdd, surpassing cutting-edge bias mitigation methods and the baseline by more than 0.7 and 0.4 units, respectively. Furthermore, achieving a lower NFR compared to other methods supports the effectiveness of XTranPrune in reducing the performance gap between subgroups without harming the overall performance of skin lesion classification. More experiments can be found in the supplementary material.

Overall, XTranPrune could outperform other methods concerning various fairness metrics. This suggests that using the explainability method enables this method to significantly enhance fairness in vision transformers by identifying the discriminatory nodes and applying a targeted pruning instead of imposing a general fairness constraint to the loss function.

Effect of pruning hyperparameters. We analysed the impact of three key hyperparameters in XTranPrune using the Fitzpatrick17k dataset. Initially, we set a *pruning rate* of 0.4 and a *number of batches* of 200 to examine the effect of the *retention rate*. Fig. 2(a) demonstrates a positive correlation between the *retention rate* and classification performance, plateauing around 0.8. Subsequently, with a fixed *retention rate* of 0.8, Fig. 2(b) depicts the influence of the *pruning rate* on performance. Increasing the *pruning rate* enhances the EOM but gradually reduces the F1 score, with higher values potentially harming fairness. We determined 0.4 as an effective *pruning rate* balancing performance and

Table 3. Ablation study on node attribution calculation method.

Method	F1 score (%)	Worst-case F1 score (%)↑	DPM↑	EOM↑	EOpp0↓	EOpp1↓	EOdd↓	NFR↓
Attention Map	73.33	54.07	0.506	0.590	0.087	0.102	0.137	0.363
Gradients	59.72	54.03	0.529	0.663	0.154	0.219	0.323	0.166
LRP	73.24	68.58	0.549	0.745	0.094	0.080	0.122	0.129
Our method	**73.51**	**69.13**	**0.586**	**0.790**	**0.086**	**0.066**	**0.095**	**0.114**

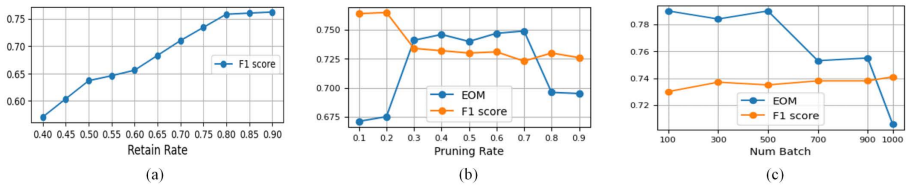

Fig. 2. Ablation Study on (a) the *retain rate* in the main branch, (b) the *pruning rate* used for generating the pruning mask, (c) the *number of batches* used to get the average attribute vectors in each pruning iteration.

fairness. Finally, in Fig. 2(c), we explored the impact of the *number of batches* in each pruning iteration. While higher values slightly improve classification performance, a notable fairness drop occurs after 500. Notably, the optimal hyperparameters for XTranPrune are *num batch* of 500, *retain rate* of 0.8, and *pruning rate* of 0.4. Setting the aforementioned values for the hyperparameters, we observed that XTranPrune prunes nearly 8% of the nodes in the main branch every iteration, that is around $477K$ parameters of the total $2.8M$ parameters we have in the 12 encoder blocks in DeiT.

Effect of using explainability method. Table 3 compares the effectiveness of various approaches to calculate each block b node attributions used to generate the pruning masks in XTranPrune such as using the learned attention maps ($A^{(b)}$), its gradients ($\nabla A^{(b)}$), and the LRP method [2], i.e., purely using node relevancy score ($R^{(n_b)}$). Our experiments demonstrate that incorporating an explicit explainability method for deriving the pruning masks is more effective since XTranPrune could substantially outperform the other methods in all of the examined fairness metrics. Ablation studies on the PAD-UFES-20 dataset can be found in the supplementary material.

4 Conclusion

This paper proposes a novel bias mitigation method tailored for vision transformers. Leveraging explainability techniques, XTranPrune effectively identifies and prunes discriminatory nodes while preserving classification performance. Through extensive experiments on two skin lesion datasets and comparisons with six other diverse methods, our method demonstrates superior performance

across various fairness metrics. XTranPrune's success highlights the superiority of explainability-based targeted debiasing over adding fairness-aware constraints, affecting the entire network, in improving model fairness. Future work could explore using non-LRP-based explainability methods to calculate node attribution and extend experiments to data modalities where the sensitive attribute is not visible.

Acknowledgments. This work was supported in part by the Google 2023 Award for Inclusion Research.

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. Binder, A., Montavon, G., Lapuschkin, S., Müller, K.R., Samek, W.: Layer-wise relevance propagation for neural networks with local renormalization layers. In: Artificial Neural Networks and Machine Learning–ICANN 2016: 25th International Conference on Artificial Neural Networks, Barcelona, Spain, September 6-9, 2016, Proceedings, Part II 25. pp. 63–71. Springer (2016)
3. Cha, J., Chun, S., Lee, K., Cho, H.C., Park, S., Lee, Y., Park, S.: Swad: Domain generalization by seeking flat minima. Advances in Neural Information Processing Systems **34**, 22405–22418 (2021)
4. Chefer, H., Gur, S., Wolf, L.: Transformer interpretability beyond attention visualization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 782–791 (2021)
5. Chiu, C.H., Chung, H.W., Chen, Y.J., Shi, Y., Ho, T.Y.: Toward fairness through fair multi-exit framework for dermatological disease diagnosis. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 97–107. Springer Nature Switzerland, Cham (2023)
6. Du, S., Hers, B., Bayasi, N., Hamarneh, G., Garbi, R.: Fairdisco: Fairer ai in dermatology via disentanglement contrastive learning. In: European Conference on Computer Vision. pp. 185–202. Springer (2022)
7. Groh, M., Harris, C., Daneshjou, R., Badri, O., Koochek, A.: Towards transparency in dermatology image datasets with skin tone annotations by experts, crowds, and an algorithm. arXiv preprint arXiv:2207.02942 (2022)
8. Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., Badri, O.: Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1820–1828 (2021)
9. 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)

10. Marcinkevics, R., Ozkan, E., Vogt, J.E.: Debiasing deep chest x-ray classifiers using intra-and post-processing methods. In: Machine Learning for Healthcare Conference. pp. 504–536. PMLR (2022)
11. Pacheco, A.G., Krohling, R.A.: The impact of patient clinical information on automated skin cancer detection. Computers in Biology and Medicine **116**, 103545 (2020)
12. Petersen, E., Feragen, A., da Costa Zemsch, M.L., Henriksen, A., Wiese Christensen, O.E., Ganz, M., Initiative, A.D.N.: Feature robustness and sex differences in medical imaging: A case study in mri-based alzheimers disease detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 88–98. Springer (2022)
13. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
14. 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)
15. Shamshad, F., Khan, S., Zamir, S.W., Khan, M.H., Hayat, M., Khan, F.S., Fu, H.: Transformers in medical imaging: A survey. Medical Image Analysis p. 102802 (2023)
16. Stanley, E.A., Wilms, M., Mouches, P., Forkert, N.D.: Fairness-related performance and explainability effects in deep learning models for brain image analysis. Journal of Medical Imaging **9**(6), 061102–061102 (2022)
17. Tartaglione, E., Barbano, C.A., Grangetto, M.: End: Entangling and disentangling deep representations for bias correction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13508–13517 (2021)
18. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021)
19. 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)
20. Xu, Z., Li, J., Yao, Q., Zhou, S.K.: Progress and prospects for fairness in healthcare and medical image analysis (May 2023), https://arxiv.org/abs/2209.13177v5
21. Zhang, H., Dullerud, N., Roth, K., Oakden-Rayner, L., Pfohl, S., Ghassemi, M.: Improving the fairness of chest x-ray classifiers. In: Conference on Health, Inference, and Learning. pp. 204–233. PMLR (2022)
22. Zong, Y., Yang, Y., Hospedales, T.: Medfair: Benchmarking fairness for medical imaging. arXiv preprint arXiv:2210.01725 (2022)