Are X-ray landmark Detection Models fair? A preliminary assessment and mitigation strategy

Roberto Di Via[1] Massimiliano Ciranni[1] Davide Marinelli[1] Allison Clement[3] Nikil Patel[3]
Julian Wyatt[3] Francesca Odone[1] Matteo Santacesaria[2] Irina Voiculescu[3] Vito Paolo Pastore[1*]
{[1]MaLGa-DIBRIS, [2]MaLGa-DIMA}, University of Genoa, Italy
[3]Oxford University, Department of Computer Science, UK

Abstract

Datasets used for benchmarking are always acquired with a view to representing different categories equally, with the best intentions to be fair to all. Whilst it is usually assumed that equal numerical representation in the training data leads to similar accuracy among demographic groups, so far, there has been next to no investigation or measurement of this assumption for the anatomical landmark detection task. In this work, we define what it means for anatomical landmark detection to be carried out fairly on different demographic categories, evaluating the fairness of models trained on two publicly available X-ray datasets that are known to be balanced, and showing how unfair predictions can uncover metadata attributes intended to be hidden. We further design a potential mitigation strategy in the landmark detection context, adapting a group optimization method typically employed for debiasing image classification models, obtaining a partial improvement in terms of per-keypoint fairness, while paving the way for further research in this field.

1. Introduction

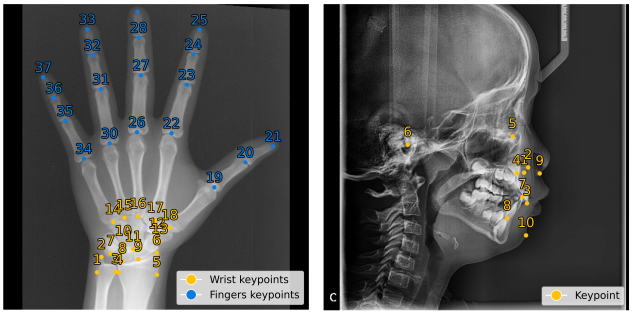
Precise and reliable anatomical landmark detection is critical for several clinical tasks [2, 8]. While the focus has been on overall accuracy [4, 9, 18] and confidence [10], few studies have addressed bias within these models [3]. Biased and unfair predictions in medical imaging can stem from non-representative training datasets or from models that inadvertently perform better for certain demographic sub-groups (i.e., age, gender, race). Limited literature exists on fairness assessment in landmark detection, and is mainly for face recognition datasets [7]. However, fairness in anatomical landmark prediction remains largely unexplored, despite its crucial clinical applications, such as diagnosis and surgical treatment planning [14]. Our work addresses this critical gap by establishing a protocol for assessing fairness in X-ray anatomical landmark prediction. Our main contribution is to show how fairness must be evaluated at *single keypoint* level (see Fig. 1), since global measures hide fairness issues that may only affect a specific subset of keypoints. To this end, we adapt a popular classification fairness metric for use with landmark detection, further investigating the relationship between landmark prediction and patient metadata (age, gender), we show how errors on keypoints can be used to infer sensitive attributes, potentially raising privacy concerns. After measuring the fairness issue, we propose a potential mitigation approach based on a group optimization method typically employed for debiasing image classification models [12, 17], that is GroupDRO [15]. Our results show a partial improvement in the fairness metrics with negligible degradation of the overall landmark detection accuracy. To our knowledge, we are the first to expose a potential lack of fairness in the context of anatomical landmark detection, which occurs even when the training data is carefully acquired in balanced categories. This work is put forward as a critical foundation for improving data acquisition makeup and for developing benchmarking criteria. At the same time, we aim to shed light on the necessity of developing an ad-hoc solution for improving fairness in the context of anatomical landmark detection.

2. Approach

2.1. Reference Datasets

Since a study on attribute bias requires plentiful raw images and metadata, of the publicly available contenders (described in [4, 19]), only the Digital Hand Atlas (DHA) [6] and the CephAdoAdu dataset are fit for purpose.
The **DHA dataset** (Fig. 1a) includes 909 radiographs (average size: 1563×2169 pixels) annotated with 37 landmarks. Among the available demographic attributes, we consider age and gender, which divide patients into groups large enough to allow a reliable assessment of group fairness. The dataset is balanced by design, with equal male and female

*Correspondence to: vito.paolo.pastore@unige.it



(a) Digital Hand Atlas (b) CephAdoAdu

Figure 1. Numbered ground truth landmarks annotations for considered datasets.

patients per age group and broadly similar numbers of patients per age group. Ages range from 9 to 18 years, but due to small per-group sizes, we cluster them into younger' (9–13 y.o.) and older' (14–18 y.o.). The **CephAdoAdu dataset** (Fig. 1b) is a new benchmark comprising cephalometric X-ray images across age groups. Our dataset version includes 350 adult and 350 adolescents X-ray images, and manually annotated with 10 key landmarks. The training protocol is age-balanced, with 400 images (including 40 for validation) for training and 300 for testing, ensuring even distribution of adult and adolescent cases.

2.2. Machine learning framework

We address landmark detection by framing it as a supervised pixel-wise classification task to produce output heatmaps. Specifically, we exploit a U-Net model to predict several landmarks at once, creating one heatmap for each landmark as separate output channels (n heatmaps, where n is the number of annotated landmarks). As per [10], we generate ground-truth heatmaps from a single pixel input annotation. Formally, $H_s(i,j) = \mathbb{1}\left(i = x_s \wedge j = y_s\right)$ where $H_s(i,j)$ is the heatmap value at pixel (i,j), the ground truth coordinates of the landmark (s) are at pixel (x_s, y_s), and $\mathbb{1}$ evaluates to 1 only when the condition is satisfied. The output heatmap intensities are in $[0,1]$, the hottest of which is the predicted landmark location.

2.3. Adapting fairness evaluation metrics

Since the dataset is designed to be balanced, it is assumed to be fair across all considered demographic categories (*gender* and *age*). Our goal is to identify the presence of group fairness issues [11] within the generic task of landmark detection. It is crucial to define an appropriate way of measuring whether a landmark detection model is *fair* (or *unfair*). A popular fairness metric for classification tasks is the *Demographic Parity* (DP), a.k.a. Statistical Parity [1, 5]. DP measures whether predicting a positive outcome is independent of a certain sensitive attribute. In a binary classification task, given a training dataset $\mathcal{D}=\{(x_1, y_1, g_1), \ldots, (x_n, y_n, g_n)\}$, where $y \in \{0,1\}$ is the target label, and g encodes a group ($0\rightarrow$Male, $1\rightarrow$Female), DP is satisfied when a positive outcome is equal across different demographic groups:

$$P\left(y{=}1 \mid g{=}0\right) \;=\; P\left(y{=}1 \mid g{=}1\right) \tag{1}$$

In this setting, this can be verified by computing the classifier's True Positive Rate $TPR{:=}\frac{TP}{TP+FN}$, separately for each group. The largest absolute TPR difference between group pairs provides an empirical measure of a classifier's fairness.

In landmark detection tasks, instead of being right or wrong, the prediction error is measured through the *Mean Radial Error* (MRE); this is the average distance between the predicted and the actual landmark positions, averaged over all landmarks. Here we use the Euclidean (L_2) distance. Another measure of the model's accuracy is the *Success Detection Rate* (SDR), reporting the proportion of predicted landmarks within a clinically acceptable distance threshold from the ground truth. In the case of n landmarks and a threshold ϕ, $SDR = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(MRE_i \leq \phi)$. Evidently, SDR can be translated as the TPR for a landmark prediction task: a True Positive is counted when a landmark is predicted within a threshold ϕ from the ground truth. Otherwise, the model is said to have missed the prediction (False Negative). In our analysis, we compute the DP for each keypoint (KP), to uncover potential fairness issues hiding in individual keypoint predictions. This results in a specific DP value for each landmark, calculated with respect to all available sensitive attributes.

2.4. Mitigating Fairness Issues

In this work, we provide an attempt at mitigating the emerged fairness issues, which seem to be related not only to demographic groups but also stemming from particular anatomical keypoints. As such, we opt to address the problem with a subpopulation characterization as fine-grained as possible. Given a dataset with K keypoints and G known demographic groups, we consider $K \times G$ possible subgroups. We encode them as an additional set of labels $\mathcal{G} = \{0, \ldots, G\}$. For an input image $x \in \mathbb{R}^{C\times W\times H}$ and per-keypoint ground-truth locations $y \in \mathbb{R}^{K\times 2}$, we provide a group label $g \in \mathcal{G}^K$, considering each keypoint separately, separating loss contributions from every G demographic group. For instance, in the case of the DHA dataset, we consider 37 keypoints and 4 demographic groups: {Young Males, Young Females, Old Males, Old Females}, for a total of 148 subgroups, where $g = 0$ denotes KP1 from Young Males, and $g = 148$ denotes KP37 in Old Females. Our subclass categorization allows us to frame the learning problem as

$$\hat{\theta} := \arg\min_{\theta\in\Theta} \max_{g\in\mathcal{G}} \left\{ \mathbb{E}_{(x,y)\sim\hat{P}_g} \left[\ell(\theta; (x,y))\right] \right\} \tag{2}$$

where θ are the Unet parameters to be optimized. This objective, known as GroupDRO [15], is originally intended for mitigating spurious correlations and improving worst-group generalization in image classification settings. We provide a customized adaptation of the method for landmark detection, fine-tuning the original model trained without any mitigation strategy with the objective in Equation 2.

3. Experiments

3.1. Experiment details

Experiments utilize a U-Net with an ImageNet pre-trained DenseNet121 encoder. Images are padded and resized to 512×512 pixels maintaining aspect ratio and normalized to $[0, 1]$. Optimization employs AdamW for up to 200 epochs with early stopping. The learning rate starts at 10^{-3}, adjusted by Exponential scheduler. The batch size is 8 with a gradient accumulation of 8. For SDR computation, ϕ is set to 2 mm, as it is the more restrictive threshold generally reported [4]. We compute metrics by converting pixel distances to millimeters: for CephAdoAdu, we use a pixel resolution of 0.1 mm; for DHA, we assume a 50 mm distance between wrist endpoints, as proposed by [13].

3.1.1. Baseline computation

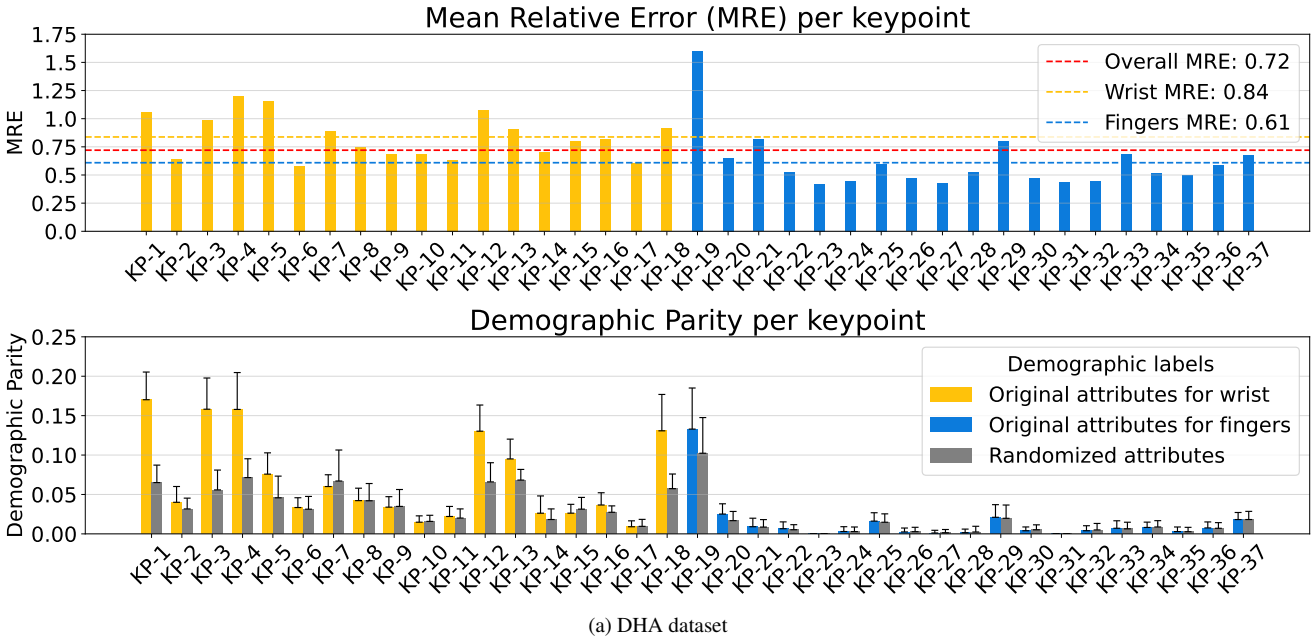
As a baseline, we perform 10 different hold-outs of the data, preserving the balance between the demographic groups. For each hold-out, we train a model on the training set and evaluate it on the held-out test set. Finally, we report the mean and standard deviation for each evaluation metric across the ten runs. Fig. 2a (top) shows the average MRE values for each keypoint of the DHA dataset across the ten runs. Wrist keypoints (KP1 to KP18) generally exhibit higher MRE values, indicating that they are somewhat more challenging to detect, with several keypoints far exceeding the overall average. Finger keypoints have comparatively better performance, though some keypoints (KP19, KP36) still exhibit high MRE. The overall MRE across all keypoints in the 10 performed runs is $0.72\,mm$, with the MRE for the wrist keypoints being slightly higher $(0.84\,mm)$ compared to MRE for finger keypoints $(0.61\,mm)$. Fig. 2b shows the same analysis replicated for the CephAdoAdu dataset. In this dataset, the overall MRE is $1.13\,mm$, with some keypoints harder to detect than the average (e.g., KP4, KP6 and KP8). For both datasets, the high variability in MRE across keypoints propagates in the SDR computations and highlights the importance of considering each keypoint individually, as relying solely on metrics averaged across the keypoints could easily hide any detection issue potentially correlated to demographic groups.

3.1.2. Fairness assessment

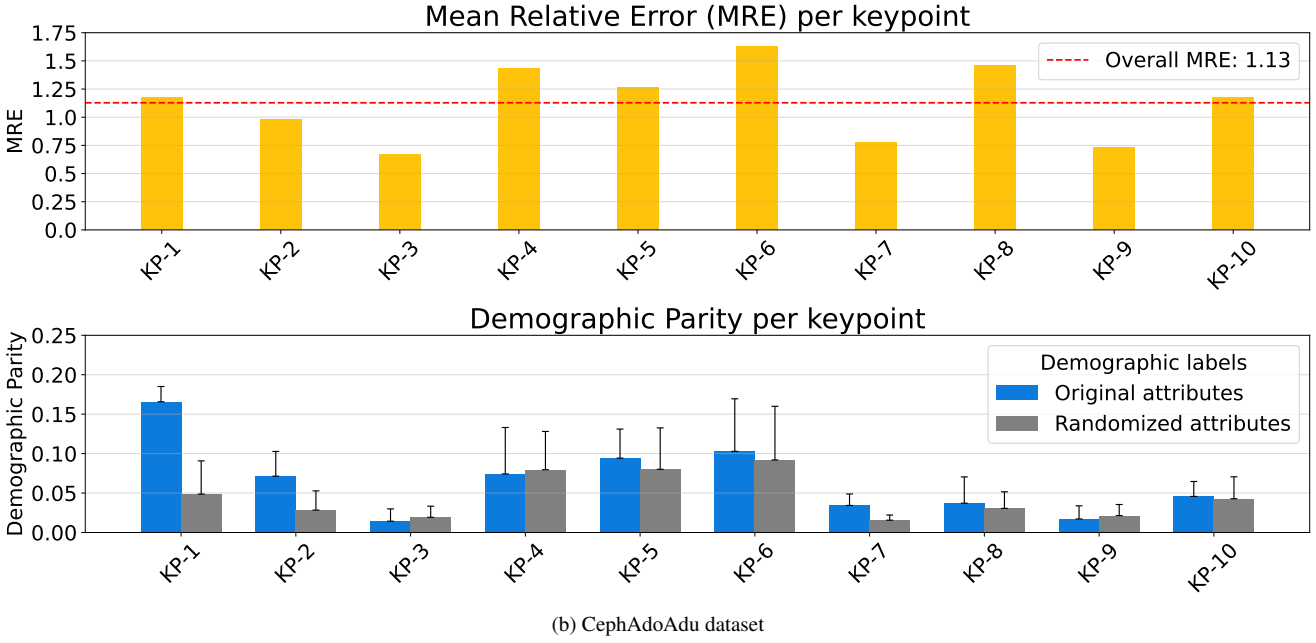
First, we compute the adapted fairness metric with respect to specific sensitive attributes over the 10 hold-outs previously introduced and averaged across all the available keypoints. For the DHA dataset, we get an overall maximum DP equal to 0.045 ± 0.009, while for the CheAdoAdu dataset, an average value of 0.080 ± 0.006 is obtained. Such relatively low values would suggest that our models are fair to the sensitive attributes. However, motivated by our previous results, whilst identifying peaks and troughs in the MRE for specific keypoints, we deepen the fairness evaluation by framing our analysis as a per-keypoint problem. Fig.2a (bottom) and Fig.2b (bottom) show the same analysis on the sensitive attributes for single keypoints for our two datasets. Analyzing these figures reveals significant variations in DP values across keypoints with respect to different sensitive attributes. Starting from the DHA dataset, for example, KP1 and KP3 (ulna) show a DP value of 0.20, meaning a maximum gap of 20% in the SDR across demographic groups. Specifically, this maximum difference arises between *female* patients in the two *age* groups. Interestingly, this doesn't align with corresponding medical research [16], who find no statistically significant difference between ages in males and females, ultimately suggesting dataset bias. Moreover, wrist keypoints show a DP value on average much higher than fingers, suggesting higher fairness in finger regions and underscoring the need for individual keypoint analysis over averaging. To further investigate the statistical significance of the obtained results, for each of our ten runs, we perform a metadata attribute randomization experiment. Specifically, we shuffle the sensitive attributes with a probability of 50% for each sample, obtaining by construction a test evaluation uncorrelated with metadata attributes. We report the corresponding average DP per keypoint in orange in Figs. 2a and 2b. For the DHA dataset, the most unfair wrist keypoints exhibit a DP significantly higher than the attribute-randomized counterpart, while finger keypoints, unaffected by fairness issues, maintain similar values across both settings. In the CephAdoAdu dataset, KP4, KP5, and KP6 show higher demographic parity but not significantly above the randomized counterpart. Notably, KP1 has a DP of 0.17, exceeding the randomized experiment, suggesting a potential fairness issue. Finally, to further support the obtained results, excluding that identified fairness issues are a consequence of a specific or suboptimal model, we report a comparison with available State-Of-The-Art (SOTA) and perform an ablation study on the Unet encoder in Table 1. Different models show similar values for the average DP across keypoints and roughly similar average MREs, being competitive with SOTA.

3.1.3. Results on fairness mitigation

Fig. 3a shows the results of our model fine-tuned with the GroupDRO objective in terms of DP for the DHA dataset (top) and CephAdoAdu (bottom). Regarding the DHA dataset, the proposed mitigation approach brings a general decreasing of DP across keypoints. However, the



(a) DHA dataset



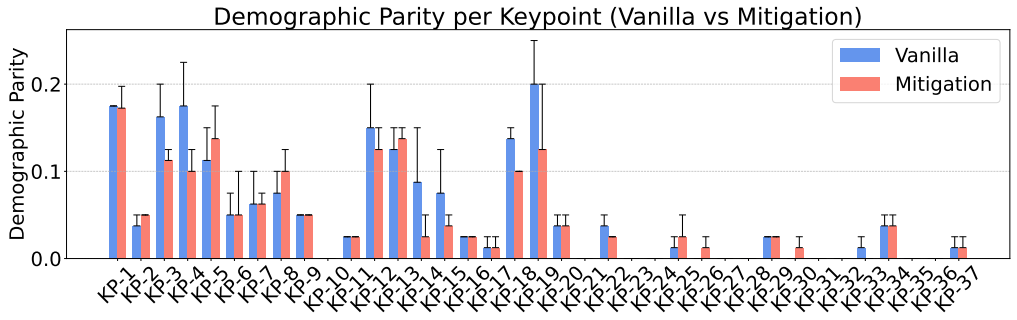
(b) CephAdoAdu dataset

Figure 2. Assessment of keypoint prediction errors and demographic parity. (a) Top panel: MRE for 37 DHA dataset keypoints, with higher error in wrist vs. finger keypoints. Bottom panel: Demographic Parity measurements with original and randomized attributes across folds. (b) Top panel: MRE for 10 CephAdoAdu dataset keypoints. Bottom panel: Demographic Parity measurements.

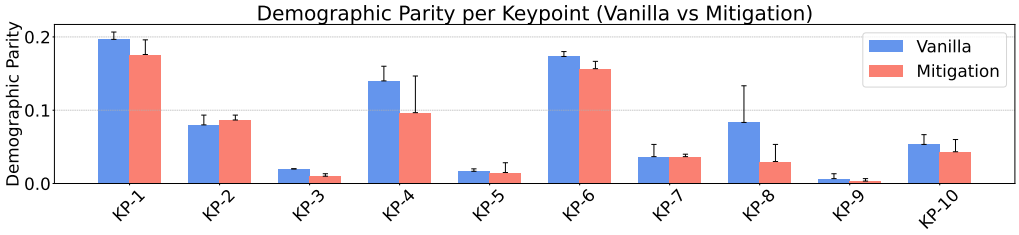
fairness issue is not entirely solved, with some wrist keypoints yet presenting a final DP higher than 0.10 (e.g., KP1, KP11, KP18 and KP19). A similar trend is observed in the CephAdoAdu dataset, with some keypoints improving their DP (e.g., KP4 and KP8). Again, the fairness issue is not entirely solved, with KP1 still presenting a DP higher than 0.15. Importantly, the mitigated models roughly preserve the average MRE across keypoints for both datasets, with a maximum drop of 0.07 and 0.04 for the CephoAdoAdu and the HDA dataset, respectively.

Table 1. Top. Comparison of state-of-the-art results for anatomical landmark detection in X-ray images. Bottom. Ablation on specific Unet backbone.

Methods	CephAdoAdu						Digital Hand Atlas					
	MRE ↓	SDR(%) ↑				DP ↓	MRE ↓	SDR(%) ↑			DP Wrist ↓	DP Fingers↓
	(mm, std.)	2mm	2.5mm	3mm	4mm	(avg. KPs)	(mm, std.)	2mm	4mm	10mm	(avg. KPs)	(avg. KPs)
SCN [13]	1.73 (1.06)	82.97	90.40	93.37	96.57	-	0.66	94.99	99.27	99.99	-	-
GU2Net [20]	1.69 (0.91)	80.33	88.13	91.47	95.57	-	0.84	95.40	99.35	99.75	-	-
CeLDA [19]	1.05 (0.33)	89.13	93.60	96.17	98.67	-	-	-	-	-	-	-
Ours (resnet50)	1.13 (0.04)	85.90	91.43	94.43	97.50	0.062 (0.004)	0.80 (0.02)	96.47	99.16	99.69	0.076 (0.006)	0.034 (0.007)
Ours (vgg19)	1.10 (0.01)	85.77	90.83	93.93	96.97	0.065 (0.003)	1.04 (0.20)	95.64	98.48	99.25	0.080 (0.004)	0.029 (0.002)
Ours (densenet121)	1.12 (0.04)	86.97	91.50	94.57	97.73	0.081 (0.006)	0.76 (0.06)	97.05	99.52	99.88	0.073 (0.011)	0.017 (0.008)



(a) DHA dataset



(b) CephAdoAdu dataset

Figure 3. Per key-point Demographic Parity (DP) on vanilla and fairness mitigated models for the (a) DHA and (b) CephAdoAdu datasets.

Table 2. 5-fold classification accuracy of a CNN trained on images and an RF trained on MREs, for various attributes.

Dataset	Sensitive Attribute	Filtered Attribute	CNN image Classifier	RF MRE-based Classifier
DHA	Age	male	0.53 ± 0.08	0.68 ± 0.05
		female	0.56 ± 0.07	0.64 ± 0.12
	Gender	young	0.56 ± 0.07	0.73 ± 0.13
		old	0.55 ± 0.08	0.72 ± 0.07
CephAdoAdu	Age	None	0.59 ± 0.16	0.64 ± 0.05

3.1.4. Privacy-related issues

Our results show a correlation between the MRE on specific keypoints and metadata attributes. Here, we evaluate if such an undesired correlation is strong enough to infer the sensitive attribute from the computed errors, potentially leading to a privacy issue. Specifically, for the CephAdoAdu dataset we consider the only available attribute (age). For the DHA dataset, where we have two sensitive attributes (gender and age), we further filter data according to a specific metadata attribute (*young/old* and *female/male* respectively), reported as *Filtered attribute* in Table 2. This approach prevents attribute mixing, isolating each sensitive attribute's contribution. Thus, we train a Random Forest (RF) classifier, exploiting the MREs corresponding to each keypoint as features and the target sensitive attribute as labels. We perform a 5-fold cross-validation, replicating the same experiments considering the test MRE across all keypoints as input features. Table 2 summarizes the obtained results. For both datasets, the MREs across keypoints bring an average test accuracy much higher than a random guess, with a maximum value of 0.75 for the sensitive attribute *age* in the *female* filtered attribute for the HDA and 0.64 for the

CephAdoAdu dataset. To ensure that these results are not a simple consequence of the sensitive attribute being inferred from the images, we train a CNN directly on the X-ray images with the same folds. As we can see in Table 2, we obtain an accuracy close to random guessing, further proving that the results are an actual consequence of the fairness issue.

4. Conclusions and future work

Despite the best intentions to acquire and anonymise patient data, we uncover concerns around the varying performance of landmark detection models known to be performing well. Privacy can be compromised through unintentional lack of fairness in such model–data pairs. Further work is required to understand this phenomenon better, potentially requiring the acquisition of new datasets and experimenting with different proportions of subjects in each demographic category, with a view to stabilising demographic parity. In this work, we adapt a typical mitigation strategy for image classification model debiasing, obtaining a partial mitigation of the described phenomenon. Despite promising results, our work eventually aims to highlight the necessity of designing ad-hoc methods (e.g., involving domain practitioners to define proper anatomical priors) for mitigating unfairness in anatomical landmark detection, potentially paving the way for multiple future investigations.

References

[1] Richard J Chen, Judy J Wang, Drew FK Williamson, Tiffany Y Chen, Jana Lipkova, Ming Y Lu, Sharifa Sahai, and Faisal Mahmood. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature biomedical engineering*, 7(6):719–742, 2023. 2

[2] Allison Clement, Abhinav Singh, and Irina Voiculescu. Landmark-based screening: Femoral head coverage and graf classification in infant developmental dysplasia of the hip. In *European Conference on Computer Vision (ECCV)*. Woman in Computer Vision (WiCV) Workshop, Springer Cham, 2024. 1

[3] Li David, Lin Cheng Ting, Sulam Jeremias, and Yi Paul H. Deep learning prediction of sex on chest radiographs: a potential contributor to biased algorithms. *Emergency Radiology*, 29(2):365–370, 2022. © 2022. American Society of Emergency Radiology. 1

[4] Roberto Di Via, Matteo Santacesaria, Francesca Odone, and Vito Paolo Pastore. Is in-domain data beneficial in transfer learning for landmarks detection in x-ray images? In *IEEE International Symposium on Biomedical Imaging, ISBI 2024, Athens, Greece, May 27-30, 2024*, pages 1–5. IEEE, 2024. 1, 3

[5] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012. 2

[6] Arkadiusz Gertych, Aifeng Zhang, James W. Sayre, Sylwia Pospiech-Kurkowska, and H. K. Huang. Bone age assessment of children using a digital hand atlas. *Comput. Medical Imaging Graph.*, 31(4-5):322–331, 2007. 1

[7] Leonardo Iurada, Silvia Bucci, Timothy M. Hospedales, and Tatiana Tommasi. Fairness meets cross-domain learning: a new perspective on models and metrics. *CoRR*, abs/2303.14411, 2023. 1

[8] Yankun Lang, Xiaoyang Chen, Hannah H. Deng, Tianshu Kuang, Joshua C. Barber, Jaime Gateno, Pew-Thian Yap, and James J. Xia. Dentalpointnet: Landmark localization on high-resolution 3d digital dental models. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2022 - 25th International Conference, Singapore, September 18-22, 2022, Proceedings, Part II*, pages 444–452. Springer, 2022. 1

[9] Juneja Mamta, Garg Poojita, Kaur Ravinder, Manocha Palak, Prateek, Batra Shivam, Singh Pradeep, Singh Shaswat, and Jindal Prashant. A review on cephalometric landmark detection techniques. *Biomedical Signal Processing and Control*, 66:102486, 2021. 1

[10] James McCouat and Irina Voiculescu. Contour-hugging heatmaps for landmark detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20597–20605, 2022. 1, 2

[11] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54 (6):1–35, 2021. 2

[12] Vito Paolo Pastore, Massimiliano Ciranni, Davide Marinelli, Francesca Odone, and Vittorio Murino. Looking at model debiasing through the lens of anomaly detection. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 2548–2557, 2025. 1

[13] Christian Payer, Darko Stern, Horst Bischof, and Martin Urschler. Integrating spatial configuration into heatmap regression based cnns for landmark localization. *Medical Image Anal.*, 54:207–219, 2019. 3, 5

[14] WR Proffit, HW Fields, and DM Sarver. Contemporary orthodontics: Elsevier health sciences. *Philadelphia, USA*, 2006. 1

[15] Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. 1, 3

[16] E Sayit, A Tanrivermis Sayit, M Bagir, and Yüksel Terzi. Ulnar variance according to gender and side during aging: An analysis of 600 wrists. *Orthopaedics & Traumatology: Surgery & Research*, 104(6):865–869, 2018. 3

[17] Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33: 19339–19352, 2020. 1

[18] Roberto Di Via, Francesca Odone, and Vito Paolo Pastore. Self-supervised pre-training with diffusion model for few-shot landmark detection in x-ray images. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV*

2025, Tucson, AZ, USA, February 26 - March 6, 2025, pages 3886–3896. IEEE, 2025. 1

[19] Han Wu, Chong Wang, Lanzhuju Mei, Tong Yang, Min Zhu, Dinggang Shen, and Zhiming Cui. Cephalometric landmark detection across ages with prototypical network. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2024 - 27th International Conference, Marrakesh, Morocco, October 6-10, 2024, Proceedings, Part V*, pages 155–165. Springer, 2024. 1, 5

[20] Heqin Zhu, Qingsong Yao, Li Xiao, and S. Kevin Zhou. You only learn once: Universal anatomical landmark detection. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part V*, pages 85–95. Springer, 2021. 5