

## Dataset Distribution Impacts Model Fairness

**Citation for published version (APA):**

Raumanns, R., Schouten, G., Pluim, J. P. W., & Cheplygina, V. (2024). Dataset Distribution Impacts Model Fairness: Single Vs. Multi-task Learning. In E. Puyol-Antón, G. Zamzmi, A. Feragen, A. P. King, V. Cheplygina, M. Ganz-Benjaminsen, E. Ferrante, B. Glocker, E. Petersen, J. S. H. Baxter, I. Rekik, & R. Eagleson (Eds.), *Ethics and Fairness in Medical Imaging: Second International Workshop on Fairness of AI in Medical Imaging, FAIMI 2024, and Third International Workshop on Ethical and Philosophical Issues in Medical Imaging, EPIMI 2024, Held in Conjunction with MICCAI 2024, Marrakesh, Morocco, October 6–10, 2024, Proceedings* (pp. 14–23). (Lecture Notes in Computer Science (LNCS); Vol. 15198). Springer. [https://doi.org/10.1007/978-3-031-72787-0\\_2](https://doi.org/10.1007/978-3-031-72787-0_2)

**Document license:**

TAVERNE

**DOI:**

[10.1007/978-3-031-72787-0\\_2](https://doi.org/10.1007/978-3-031-72787-0_2)

**Document status and date:**

Published: 13/10/2024

**Document Version:**

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

**Please check the document version of this publication:**

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

**General rights**

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

[www.tue.nl/taverne](http://www.tue.nl/taverne)

**Take down policy**

If you believe that this document breaches copyright please contact us at:

[openaccess@tue.nl](mailto:openaccess@tue.nl)

providing details and we will investigate your claim.



# Dataset Distribution Impacts Model Fairness: Single Vs. Multi-task Learning

Ralf Raumanns<sup>1,2(✉)</sup>, Gerard Schouten<sup>1</sup>, Josien P. W. Plum<sup>2</sup>,  
and Veronika Cheplygina<sup>3</sup>

<sup>1</sup> Fontys University of Applied Science, Eindhoven, The Netherlands  
[ralf.raumanns@gmail.com](mailto:ralf.raumanns@gmail.com)

<sup>2</sup> Eindhoven University of Technology, Eindhoven, The Netherlands

<sup>3</sup> IT University of Copenhagen, Copenhagen, Denmark

**Abstract.** The influence of bias in datasets on the fairness of model predictions is a topic of ongoing research in various fields. We evaluate the performance of skin lesion classification using ResNet-based CNNs, focusing on patient sex variations in training data and three different learning strategies. We present a linear programming method for generating datasets with varying patient sex and class labels, taking into account the correlations between these variables. We evaluated the model performance using three different learning strategies: a single-task model, a reinforcing multi-task model, and an adversarial learning scheme. Our observations include: 1) sex-specific training data yields better results, 2) single-task models exhibit sex bias, 3) the reinforcement approach does not remove sex bias, 4) the adversarial model eliminates sex bias in cases involving only female patients, and 5) datasets that include male patients enhance model performance for the male subgroup, even when female patients are the majority. To generalise these findings, in future research, we will examine more demographic attributes, like age, and other possibly confounding factors, such as skin colour and artefacts in the skin lesions. We make all data and models available on GitHub.

**Keywords:** Skin lesions · Bias · Fairness · Multi-task learning · Adversarial learning

## 1 Introduction

Deep learning has shown many successes in medical image diagnosis [3, 12, 30], but despite high overall performance, models can be biased against patients from different demographic groups [1, 15, 22]. Bias and fairness are becoming an active topic in medical imaging, with studies focusing for instance on skin lesions [1, 17], chest x-rays [22] and brain MR scans [27]. Examples of sensitive attributes

---

**Supplementary Information** The online version contains supplementary material available at [https://doi.org/10.1007/978-3-031-72787-0\\_2](https://doi.org/10.1007/978-3-031-72787-0_2).

include age, sex or race. For skin lesion classification, the Fitzpatrick skin type is often studied [4, 17, 31, 34].

Fairness studies typically include baselines showing bias between groups, and/or propose methods to improve fairness. The methods are based on sampling or weighting strategies during training [17], and/or introducing training strategies that try to debias the methods to rely on the sensitive attributes, such as adversarial methods [1]. For instance, Yang and colleagues [35] developed an adversarial debiasing framework to reduce biases in hospital location and patient ethnicity. Similarly, Wu et al. introduced FairPrune [34], a method for pruning parameters based on their significance to both privileged and unprivileged groups, focusing on sex and skin tone. Moreover, Bevan and Atapour-Abarghouei [5] employed various strategies to limit bias in skin lesion images, specifically targeting discrepancies arising from medical instruments, surgical markings, and rulers. Popular datasets include ISIC skin lesion datasets [9–11, 18, 29, 32] and Fitzpatrick-17K [16, 17]. Researchers either use already provided data splits for evaluation, or split the data ratios of patients with a specific demographic attribute, for example male vs female patients.

Our current study builds on two crucial insights from other topics in medical imaging: multi-task learning and shortcut learning [13, 25]. Firstly, some studies use demographic attributes within multi-task learning settings; for example, [23]. Here the attributes are *reinforcing* the diagnosis during optimization. This is at odds with the more recent adversarial strategies [1, 2] where models are encouraged to NOT predict the sensitive attribute. Secondly, there are correlations between demographics and demographic attributes and shortcut learning, including, for example, imaging devices and surgical markers [5, 6, 15, 21, 33]. In such cases, simply splitting the data according to a specific attribute can create imbalance in terms of the other attributes, thus the observed (un)fairness could be due to the attributes that were not considered.

**Our contributions** are as follows:

1. We propose using the linear programming (LP) approach for skin lesions retrieved from the ISIC archive [9–11, 18, 29, 32] via the gallery browser [20]. This method gives more control over patient subset assignment. It adjusts the proportions of selected dataset attributes while keeping others constant.
2. We systematically study two strategies that handle the demographic variable in different ways: a reinforcing multi-task strategy [24, 28] and an adversarial strategy [1, 2, 8]
3. We evaluate our models using overall and subgroup Area Under the Curve (AUC) based on sex, and show that:
  - Models perform better for male subgroups in the male-only and lightly skewed male patient experiments. In the balanced dataset and lightly skewed female patient experiments, there is no significant difference between the subgroups. However, in the lightly skewed female patient scenario, the adversarial model performs better for male patients.
  - Models trained exclusively on female patients exhibit a positive difference in performance for female patients.

- The base model reveals a significant sex bias, performing worse for female patients, except when trained exclusively on female patients.
  - The reinforcement model has no significant effect on sex bias.
  - The adversarial model significantly reduces sex bias in scenarios involving only female patients.
  - Eliminating model bias is challenging, with significant performance gaps observed in datasets with skewed sex distributions.
4. We make all data and models available on <https://github.com/raumannsr/data-fairness-impact>.

## 2 Methods

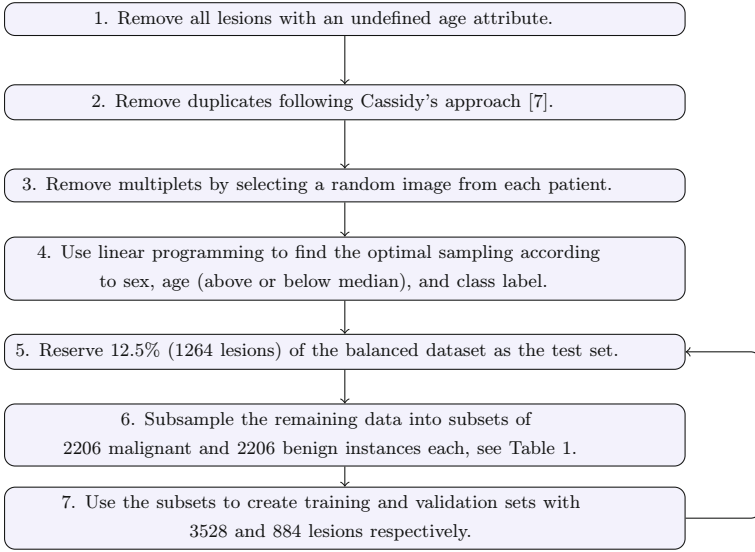
**Construction of Datasets.** We trained and validated our models on datasets with different female (F) to male (M) patient ratios, equal numbers of malignant and benign lesions, and equal number of patients below and above 60 (median age) for each sex. We refer to the datasets as M100 (100% male patients), F25M75 (25% female patients, 75% male patients), F50M50, F75M25, and F100 are defined analogously. We evaluate the models using a balanced test set mirroring F50M50.

We used the ISIC archive’s [9–11, 18, 29, 32] gallery browser [20], which had 81,155 dermoscopic images of skin lesions, some with age and sex metadata. We queried the archive for “dermoscopic” images diagnosed as “benign” or “malignant” for all ages and both sexes. This gave us 71,035 images (62,439 benign, 8,596 malignant), which we processed using the steps in Fig. 1 (see Appendix for more details).

**Linear Programming for Optimal Dataset Construction.** We have developed a method to create diverse dataset compositions using linear programming, a common mathematical optimisation technique. The goal is to maximise the number of instances of skin lesions within defined constraints, as we express below:

$$\begin{array}{ll}
 \text{Find a vector } x & \text{(decision variables)} \\
 \text{that maximises } f = x_1 & \text{(objective function)} \\
 \text{subject to } a_{i1}x_1 + a_{i2}x_2 + \dots + a_{in}x_n \leq b_i & \text{(constraints)} \\
 \text{for } i = 1, \dots, 13 & \\
 \text{and } x_j \geq 0. & \text{(non-negativity constraints)} \\
 \text{for } j = 1, \dots, 14 &
 \end{array}$$

- The decision variables  $(x_1, \dots, x_{14})$  denote specific categories, like benign lesions in female patients aged 60 and above. See Appendix for more.
- The objective function of the LP model is designed to maximise the number of malignant instances  $x_1$ . There are fewer malignant instances than benign ones in the ISIC archive, and the goal is to achieve a balance between the two.



**Fig. 1.** Steps for filtering lesions and creating test, training and validation sets. Steps 5 through 7 are repeated using 5 different seeds in a cross-validation setup.

- The constraints limit the solution by setting specific limits for each group and maintaining ratios between these groups. Group examples include all benign lesions, all female patients over 60, and all male patients under 60. The primary constraint ensures an equal number of malignant and benign lesions ( $x_1 - x_2 = 0$ ). See Appendix for more.
- Due to non-negativity constraints, decision variables cannot be negative.

**Table 1.** Datasets are distributed amongst malignant, benign, male patients (M), and female patients (F) categories for both training and validation.

	M100	F25M75	F50M50	F75M25	F100
Malignant (M/F)	2206 (2206/0)	2941 (2206/735)	4412 (2206/2206)	3235 (809/2426)	2426 (0/2426)
Benign (M/F)	2206 (2206/0)	2941 (2206/735)	4412 (2206/2206)	3235 (809/2426)	2426 (0/2426)

Within set constraints, the optimal solution maximises malignant lesions and assigns value to decision variables. To find this solution, we created a unique LP model for each dataset. Table 1 shows the result of the LP model for the different datasets.

**Models.** We used the ResNet50 model [19] in three ways, which include:

- The single-task baseline model enhanced with two fully connected layers has a sigmoid activation function and binary cross-entropy loss ( $L_c$ , see Eq. 1).

We did not use class weights but rather the actual distribution represented by the training dataset. We fine-tuned the model through a grid search of three varying learning rates and batch sizes, selecting the combination that yielded the highest performance across all experiments.

- The multi-task reinforcing model, with three added layers to the convolutional base, produces two outputs: one for classification and the other for the binary sex attribute. We employed binary cross-entropy loss ( $L_c$ ) and a sigmoid activation function for both heads, giving equal weight to both losses.
- The multi-task adversarial model was implemented following the methodology [1,2], using a network with a shared feature encoder and two classifier heads. One classifier targeted skin cancer classification; the other predicted confounding variables like sex or age. We used the ResNet architecture to compare performance with baseline and reinforcing models equitably. We trained the skin cancer classifier and encoder with cross-entropy loss ( $L_c$ ) and optimised the bias predictor with binary cross-entropy loss ( $L_c$ ). To diminish the confounder predictiveness of the encoded feature, we adversarially adjusted the encoder using a third loss ( $L_{br}$ ), setting  $\lambda$  as the penalty for accurate demographic predictions. We used a lambda ( $\lambda$ , see Eq. 2) value of 5, as in [1], to assess subgroup performance and set the penalty for correct predictions of the target demographic parameter.

To summarise, we use these loss functions:

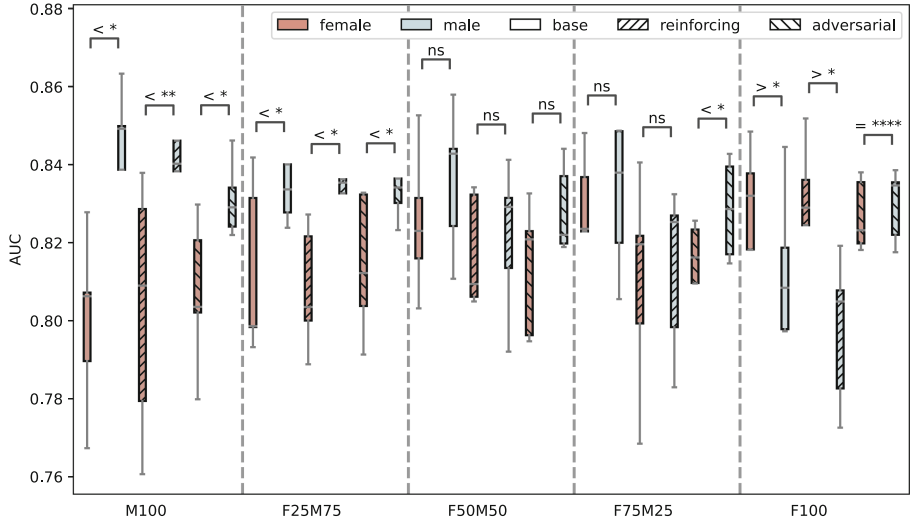
$$L_c = - \sum_{i=1}^n [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (1)$$

$$L_{br} = \lambda L_c \quad (2)$$

where  $n$  represents the number of lesions, and  $y_i$  and  $\hat{y}_i$  denote respectively the prediction and the expected outcome for lesion  $i$ .

We pre-trained all networks with ImageNet and resized images to  $384 \times 384$  for ResNet50’s input size. During model training, we used data augmentation techniques, ran up to 40 epochs with a batch size of 20, and set a learning rate of  $2.0e-5$ . We stopped training if no significant improvement occurred after 10 consecutive epochs to avoid overfitting. We implemented our baseline and reinforcing models in Keras with the TensorFlow backend [14] and our adversarial model in PyTorch [26].

**Evaluation.** For the purpose of an in-depth evaluation, we generated five distinct instances of each dataset: M100, F25M75, F50M50, F75M25 and F100. Each instance was created with a unique seed to ensure diversity and robustness in our evaluation. Each seed corresponds to a balanced test set to allow fair comparisons between dataset-model instances. Furthermore by using a balanced test set, we ensure that the results are not skewed towards any specific subgroup. We evaluate AUC overall and for male and female subgroups for each model architecture and dataset combination.



**Fig. 2.** The AUC score varies based on data splits ranging from only male patients (M100) to only female patients (F100). We show base, reinforcing and adversarial model performance for female and male patient subgroups. Significance per Mann-Whitney U test (as used in [22]) is denoted by \*\*\*\* ( $P \leq 0.0001$ ), \*\*\* ( $0.0001 < P \leq 0.001$ ), \*\* ( $0.001 < P \leq 0.01$ ), \* ( $0.01 < P \leq 0.1$ ), and not significant (ns) ( $P > 0.1$ ). < indicates lower AUCs, > higher AUCs, and = comparable AUCs for female patients.

### 3 Results

Figure 2 shows the impact of dataset distributions on three learning strategies, reporting AUC scores overall and for both sexes.

**Sex-Specific Training Data Yields Better Results.** Models perform better for male subgroups in the male-only and lightly skewed male patient experiments. In the balanced dataset and lightly skewed female patient experiments, there is no significant difference between the subgroups. However, in the lightly skewed female patient scenario, the adversarial model performs better for male patients. An exception is observed when the training datasets consist only of female patients; in such cases, there is a positive difference in AUC scores favouring female patients. Thus, our models seem more attuned to male patients in mixed-sex training sets, irrespective of the percentage of female patients. The best results are achieved when both sexes are trained exclusively on their respective data. Despite this, a male subgroup bias is apparent as the results for female patients are significantly worse than for male patients when trained exclusively on their data.

**Base Model Reveals Sex Bias.** The base model shows a significant sex bias in performance. When only male patients are involved the base model reveals a

substantial performance gap between male and female patients. The results are significantly worse for female patients in male-skewed scenarios, except for the female-only experiments. Interestingly, the base model performs better for female patients than the adversarial and reinforcing models in the F75M25 dataset scenario. It is worth noting that the performance gap between the subgroups is not as apparent when the dataset includes male and female patients, unlike in the experiment that only involved male patients.

**Reinforcement Model has No Significant Effect on Sex Bias.** Training the model only on male patients increases AUC score variability and reduces performance differences between both sexes. The reinforcement model does not significantly affect sex bias.

**Adversarial Model Eliminates Sex Bias in Cases Involving Only Female Patients.** The adversarial model reduces sex bias significantly in scenarios with only female patients but is less effective in other scenarios, often favouring male patients. Its performance varies across experiments and datasets.

## 4 Discussion and Conclusions

We studied model and subgroup performance across datasets to identify the influence of bias in datasets on fairness of model predictions. We used linear programming (LP) to create various datasets with controlled male-female ratios. This was done to systematically evaluate the performance of three different learning strategies using ResNet-based CNNs. Other fairness and bias studies that require a flexible method to create datasets under certain constraints could benefit from this universal LP technique.

Our study shows that eliminating bias is challenging. The adversarial model architecture is able to reduce sex bias in a female-only context but fails for other datasets. Other model approaches do not show convincing results with respect to bias reduction.

Skewed sex distributions still show a performance gap between male and female patients. Our experiments demonstrate that the adversarial model better corrects sex bias in female-only datasets and not in male-only datasets, possibly due to other confounding and/or unidentified factors. Further research is needed on this issue.

As expected the base model shows a sensitivity for sex bias, possibly due to overfitting. The reinforcing and adversarial models both having a form of regularisation (to counter overfitting), are potentially able to reduce sex bias compared to the base model. However in our experiments we only see a bias correction for adversarial models for female-only experiments.

Our outcomes show that sex-related information influences prediction tasks. Future research should determine which specific sex-related factors are essential to ensure fairness across different subgroups.

In contrast to categorical data like patient sex, where the groups are clearly defined, this is not possible or only partially possible with continuous data like

age, which could lead to somewhat arbitrary subgroups. Therefore, we started with the demographic attribute sex and will continue similar research with the non-categorical age attribute.

Further, we have identified the following directions for future work:

- Investigate whether using “early stopping” per task in a multi-task model reduces subgroup bias.
- Explore the impact of integrating segmentation with a classifier on sex-based disparities in identifying skin lesions.
- Study the roles of factors like skin colour and image artefacts in model fairness for different subgroups.
- Investigate the impact of shortcut learning on model fairness.

In conclusion, while we progress towards fairness, further advancements are needed to ensure consistent and equitable performance across various data distributions.

**Acknowledgments.** We gratefully acknowledge financial support from the Netherlands Organization for Scientific Research (NWO), grant no. 023.014.010.

**Disclosure of Interests.** The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

## References

1. Abbasi-Sureshjani, S., Raumanns, R., Michels, B.E., Schouten, G., Cheplygina, V.: Risk of training diagnostic algorithms on data with demographic bias. In: MICCAI LABELS Workshop, LNCS, vol. 12446, pp. 183–192. Springer (2020). [https://doi.org/10.1007/978-3-030-61166-8\\_20](https://doi.org/10.1007/978-3-030-61166-8_20)
2. Adeli, E., et al.: Representation learning with statistical independence to mitigate bias. *IEEE Winter Conf. Appl. Comput. Vis.* **2021**, 2512–2522 (2021)
3. Bejnordi, B.E., et al.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* **318**(22), 2199–2210 (2017)
4. Benčević, M., Habijan, M., Galić, I., Babin, D., Pižurica, A.: Understanding skin color bias in deep learning-based skin lesion segmentation. *Comput. Methods Programs Biomed.* **245**, 108044 (2024)
5. Bevan, P.J., Atapour-Abarghouei, A.: Skin deep unlearning: artefact and instrument debiasing in the context of melanoma classification. *arXiv preprint arXiv:2109.09818* (Apr 2023)
6. Bissoto, A., Valle, E., Avila, S.: Debiasing skin lesion datasets and models? Not so fast (2020)
7. Cassidy, B., Kendrick, C., Brodzicki, A., Jaworek-Korjakowska, J., Yap, M.H.: Analysis of the ISIC image datasets: usage, benchmarks and recommendations. *Med. Image Anal.* **75**, 102305 (2022)

8. Chu, Z., Rathbun, S.L., Li, S.: Multi-Task adversarial learning for treatment effect estimation in basket trials. In: Flores, G., Chen, G.H., Pollard, T., Ho, J.C., Naumann, T. (eds.) Proceedings of the Conference on Health, Inference, and Learning. Proceedings of Machine Learning Research, vol. 174, pp. 79–91. PMLR (2022)
9. Codella, N., et al.: Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the international skin imaging collaboration (ISIC) (2019)
10. Codella, N.C.F., et al.: Skin lesion analysis toward melanoma detection: a challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC) (2018)
11. Combalia, M., et al.: BCN20000: dermoscopic lesions in the wild (2019)
12. Esteva, A., et al.: Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**(7639), 115–118 (2017)
13. Geirhos, R., et al.: Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2**(11), 665–673 (2020)
14. Géron, A.: Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. “O’Reilly Media, Inc.” (Oct 2022)
15. Gichoya, J.W., et al.: AI recognition of patient race in medical imaging: a modelling study. *Lancet Digit. Health* **4**(6), e406–e414 (2022)
16. 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. *Proc. ACM Hum.-Comput. Interact.* **6**(CSCW2), 1–26 (2022)
17. Groh, M., et al.: Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1820–1828 (Apr 2021)
18. Gutman, D., et al.: Skin lesion analysis toward melanoma detection: a challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC) (2016)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
20. ISIC archive. <https://gallery.isic-archive.com>. Accessed 7 June 2024
21. Jiménez-Sánchez, A., Juodelyte, D., Chamberlain, B., Cheplygina, V.: Detecting shortcuts in medical images—a case study in chest x-rays. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), pp. 1–5. IEEE (2023)
22. 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. *Proc. Natl. Acad. Sci.* **117**(23), 12592–12594 (2020)
23. Liu, X., Shi, J., Zhou, S., Lu, M.: An iterated Laplacian based semi-supervised dimensionality reduction for classification of breast cancer on ultrasound images. In: International Conference of the IEEE Engineering in Medicine and Biology Society, vol. 2014, pp. 4679–4682 (2014). <https://doi.org/10.1109/EMBC.2014.6944668>
24. Marques, S., Schiavo, F., Ferreira, C.A., Pedrosa, J., Cunha, A., Campilho, A.: A multi-task CNN approach for lung nodule malignancy classification and characterization. *Expert Syst. Appl.* **184**, 115469 (2021)
25. Nauta, M., Walsh, R., Dubowski, A., Seifert, C.: Uncovering and correcting shortcut learning in machine learning models for skin cancer diagnosis. *Diagnostics (Basel)* **12**(1), 40 (2021)
26. Paszke, A., et al.: Others: an imperative style, high-performance deep learning library. *Adv. Neural. Inf. Process. Syst.* **32**, 8026–8037 (2019)

27. Petersen, E., et al.: Feature robustness and sex differences in medical imaging: a case study in MRI-based alzheimer’s disease detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 88–98. Springer (2022). [https://doi.org/10.1007/978-3-031-16431-6\\_9](https://doi.org/10.1007/978-3-031-16431-6_9)
28. Raumanns, R., Schouten, G., Joosten, M., Pluim, J.P.W., Cheplygina, V.: Enhance (enriching health data by annotations of crowd and experts): a case study for skin lesion classification. *Machine Learning for Biomedical Imaging* **1**, 1–26 (2021). <https://doi.org/10.59275/j.melba.2021-geb9>, <https://melba-journal.org/2021:020>
29. Rotemberg, V., et al.: A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific Data*; London **8**(1), s41597–021 (2021)
30. Saha, A., et al.: Artificial intelligence and radiologists in prostate cancer detection on MRI (PI-CAI): an international, paired, non-inferiority, confirmatory study. *Lancet Oncol.* (2024)
31. Seth, P., Pai, A.K.: Does the fairness of your Pre-Training hold up? Examining the influence of Pre-Training techniques on skin tone bias in skin lesion classification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 570–577 (2024)
32. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions (2018)
33. Willeminck, M.J., et al.: Preparing medical imaging data for machine learning. *Radiology* **295**(1), 4–15 (2020)
34. Wu, Y., Zeng, D., Xu, X., Shi, Y., Hu, J.: FairPrune: achieving fairness through pruning for dermatological disease diagnosis. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2022, pp. 743–753. Springer Nature Switzerland (2022). [https://doi.org/10.1007/978-3-031-16431-6\\_70](https://doi.org/10.1007/978-3-031-16431-6_70)
35. Yang, J., Soltan, A.A.S., Eyre, D.W., Yang, Y., Clifton, D.A.: An adversarial training framework for mitigating algorithmic biases in clinical machine learning. *NPJ Digit. Med.* **6**(1), 55 (2023)