MELBA Journal Machine Learning for Biomedical Imaging

Effect of Demographic Bias on Skin Lesion Classification

Ralf **Raumanns** [1,3,4], Gerard Schouten [2], Veronika Cheplygina [4], Josien P.W. Pluim [3]

1 Fontys University of Applied Science, Venlo, The Netherlands
2 Fontys University of Applied Science, Eindhoven, The Netherlands
3 Eindhoven University of Technology, Eindhoven, The Netherlands
4 IT University of Copenhagen, Denmark

Abstract

The influence of bias in datasets on the fairness of model predictions is a topic of ongoing research in various fields. In this study, we evaluate the performance of skin lesion classification using ResNet-based convolutional models, focusing on the impact of demographic bias in training data, particularly variations in patient sex and age. We use a linear programming method to generate datasets with controlled demographic characteristics, allowing systematic investigation of bias effects. Three distinct learning strategies are evaluated: a single-task model, a reinforcing multi-task model, and an adversarial learning scheme.

Our sex-based analysis indicates that sex-specific training datasets optimise model performance. Notably, including male patients in the training data improved performance for the male subgroup, even in female-majority cases. Reinforcing and adversarial learning schemes narrowed or eliminated bias gaps in balanced and female-majority datasets. However, these strategies proved less effective in male-majority settings, where models continued to perform better for males than females. The two learning schemes showed marginal bias reduction compared to the baseline model in predominantly male patient populations.

Age-based analysis demonstrates comparable baseline performance across the three model approaches, with performance declining across age categories. Younger groups consistently achieve the highest performance, regardless of training data distribution. Although balanced training yields optimal results for the youngest age category, performance decreases in older categories.

We find that sex biases arise mainly from data imbalances, while age biases consistently favour younger groups regardless of distribution. These distinct mechanisms require targeted mitigation strategies. Our work aims to advance equitable AI in medical imaging by addressing these specific sources of disparity.

Additionally, cross-dataset validation on two external datasets revealed that domain shifts notably affect performance and demographic bias patterns.

The source code and models are available on GitHub:
`https://github.com/raumannsr/demographic-fairness-extended`.

Keywords

Skin lesions, Bias, Fairness, Multi-task learning, Adversarial learning, Cross-dataset analysis

Article informations

`https://doi.org/10.59275/j.melba.2026-4156`

Volume 2026, Received: 2025-04-21, Published 2026-05-29

Corresponding author: ralf.raumanns@fontys.nl

Special issue: Special issue on Fairness of AI in Medical Imaging (FAIMI)
Guest editors: Veronika Cheplygina, Aasa Feragen, Andrew King, Ben Glocker, Enzo Ferrante, Eike Petersen, Esther Puyol-Antón, Melanie Ganz-Benjaminsen

1. Introduction

Deep learning has shown many successes in the diagnosis of medical images, as demonstrated by several studies (Saha et al. (2024); Esteva et al. (2017); Bejnordi et al. (2017)), but despite the high overall performance, models can be biased against patients from different demographic groups, a concern highlighted in recent work (Abbasi-Sureshjani et al. (2020); Larrazabal et al. (2020); Gichoya et al. (2022b)). Bias and fairness have therefore become central research

topics in medical imaging, with studies focusing, for example, on skin lesions (Abbasi-Sureshjani et al. (2020); Groh et al. (2021)), chest radiographs (Larrazabal et al. (2020)) and brain magnetic resonance imaging (Petersen et al. (2022)). Sensitive attributes commonly examined are age, sex, or race. For the classification of skin lesions, the Fitzpatrick skin type is often studied (Seth and Pai (2024); Benčević et al. (2024); Groh et al. (2021); Wu et al. (2022)).

While deep learning models continue to advance diagnostic capabilities, their fairness remains a significant concern because model performance is fundamentally tied to the quality and representativeness of the training data, as well as the model's ability to mitigate any bias embedded in the training dataset.

Although bias and fairness in AI for medical imaging have gained attention, prior studies have often examined individual demographic factors in isolation, typically within a single imaging modality or without systematic control over data distributions. A comprehensive evaluation comparing how these demographic attributes, when systematically skewed, influence model performance across different learning strategies (single-task, reinforcing, and adversarial) is lacking. Moreover, the relative effectiveness of debiasing approaches across specific demographic subgroups, particularly under extreme distributional imbalances, remains unexplored. Additionally, the utility of auxiliary demographic prediction heads as fairness indicators has not been systematically assessed.

In this paper, we define dataset bias (also known as representation bias) strictly as demographic bias, meaning any systematic imbalance in age, sex, or other protected attributes within the training set. Such imbalances lead to unbalanced learning and performance gaps between subgroups. We examine both demographic and model bias, measuring how controlled skews in the training data affect performance, and testing multi-task learning strategies designed to mitigate model bias. Using a balanced test set, we quantify the degree to which demographic bias propagates to model bias and identify the most effective approaches for equitable skin-lesion classification across age and sex groups.

This manuscript substantially extends our FAIMI 2024 workshop paper (Raumanns et al. (2025)). The workshop paper evaluated five distributions of male/female patients (sex demographics) with three learning strategies (one single-task and two multi-task models). The evaluation focused on overall and subgroup-specific performance to assess whether training data distribution biases manifested in results when tested on a balanced test set.

Extending our FAIMI 2024 workshop paper, we present the following contributions:

1. We extend our linear programming (LP) method to control age subgroups in addition to sex, introducing five age groups and three skewed age distributions.
2. We systematically evaluate two bias mitigation strategies (reinforcing multi-task and adversarial) across various age and sex subgroups. By presenting both overall and subgroup-specific metrics, we determine how each strategy performs under different conditions. This includes two new sex-distribution scenarios, namely predominantly male and predominantly female patients, which enable a more granular evaluation of the models.
3. Beyond the internal hold-out validation, we extend our external validation from the prior study. We introduce a new dermatoscopic skin-lesion dataset in this work. Alongside the retained smartphone dataset, the datasets facilitate testing across diverse geographical regions, acquisition methods, and demographic groups.
4. We analyse the auxiliary age-prediction head to assess its utility as a fairness indicator.

2. Related work

We revisit prior studies on demographic bias and fairness in medical imaging, highlighting how earlier work has examined demographic disparities, bias mitigation techniques such as multi-task and adversarial learning, and the limitations that motivate our more systematic analysis.

Understanding representation bias Demographic bias in medical imaging, referring to performance disparities across protected attributes (such as biological sex, race, age, and skin tone), has been extensively studied, revealing how these imbalances can cause unfair or discriminatory outcomes in healthcare. Glocker et al. showed that a widely used chest radiography foundation model actually encodes protected attributes, like biological sex and race, leading to statistically significant performance gaps across those subpopulations (Glocker et al. (2023)). Vaidya et al. reported that deep learning pathology models exhibit racial bias, as demonstrated on large publicly available cancer imaging datasets (Vaidya et al. (2024)).

Demographic bias in machine learning manifests in various forms, with representation bias being particularly significant in healthcare. Representation bias occurs when certain demographic groups are underrepresented in training data, leading to reduced model performance for these groups (Larrazabal et al. (2020)). This differs from bias caused by inherent anatomical or physiological differences between groups, though these can contribute to representation bias when they affect data collection, for example, clinical protocols that exclude pregnant patients for safety reasons (Seyyed-Kalantari et al. (2021)).

Sies et al. assessed a market-approved skin cancer CNN and documented a male predominance in the training data. Despite this imbalance, performance on a balanced test set showed no statistically significant sex-related disparity, suggesting the extensive training set mitigated the imbalance effect (Sies et al. (2022)). Conversely, even with deliberately balanced datasets, intrinsic anatomical differences can still generate bias. Klingenberg et al. demonstrated this by showing that a CNN trained on a sex-balanced MRI cohort for Alzheimer's detection performed markedly better in female patients than males, underscoring that demographic bias can arise from physiological factors rather than merely data imbalance (Klingenberg et al. (2023)).

Understanding how representation bias influences model performance is essential for building fair systems. By identifying underrepresented populations or those with the poorest performance, targeted corrections can be applied to the dataset. Moreover, understanding these effects provides insights for designing future datasets, allowing researchers to avoid similar problems early on.

Role of demographics The role of demographics in medical AI is multifaceted. Some demographic variations reflect genuine biological differences that models should take into account; for example, patient characteristics such as age and sex significantly influence the predictive precision of health markers, such as blood pressure, in retinal image analysis (Gerrits et al. (2021)). Deep learning models can extract demographic characteristics, such as sex and age, directly from medical images, such as chest X-rays, with high accuracy (Gichoya et al. (2022a);Jones and Glocker (2025)). This capability offers applications in forensic investigations, aiding identification and uncovering novel anatomical landmarks for sex and age determination (Yi et al. (2021)).

However, it is crucial to distinguish these valid demographic correlations from problematic representation bias, which often relates to data collection practices rather than physiological differences. Our research addresses this by deliberately building datasets with specific demographic imbalances, helping us determine whether performance disparities are due to true physiological factors or simply to data collection.

Addressing bias Research on fairness usually involves baseline studies that demonstrate bias between groups and/or suggest methods to enhance fairness. These approaches mainly tackle representation bias through sampling or weighting strategies during training (Groh et al. (2021)). Alternatively, they implement architectural techniques that prevent models from depending on sensitive attributes, such as adversarial learning (Abbasi-Sureshjani et al. (2020)). For example, Yang et al. developed an adversarial framework to mitigate biases arising from hospital location and patient ethnicity (Yang et al. (2023)). Wu et al. introduced FairPrune, which trims parameters based on their importance to both privileged and unprivileged groups (Wu et al. (2022)). Other methods focus on data augmentation. Stanley et al. proposed a synthetic bias framework for brain MRI. They showed that simple sample reweighting effectively reduces hidden biases (Stanley et al. (2024)). Ktena et al. demonstrated that diffusion-generated synthetic images improve fairness across histopathology, chest X-ray, and dermatology datasets (Ktena et al. (2024)).

Commonly used datasets for studying demographic bias in skin lesion classification include the ISIC skin lesion datasets (Gutman et al. (2016); Codella et al. (2018, 2019); Tschandl et al. (2018); Combalia et al. (2019); Rotemberg et al. (2021)) and Fitzpatrick-17K (Groh et al. (2021, 2022)). However, researchers typically rely on pre-provided data splits or stratify by a single demographic attribute (e.g., male vs female). Crucially, these methods often fail to control for the interplay between attributes, treating sex and age as independent variables rather than managing their joint distribution. Our linear programming approach, however, explicitly enforces constraints on both sex and age simultaneously, ensuring that specific subgroups (such as older males or younger females) are accurately represented according to the desired ratios.

Bias mitigation approaches Our current study builds on two crucial insights from medical imaging: multi-task learning and shortcut learning (Geirhos et al. (2020); Nauta et al. (2021)). The reinforcing model uses multi-task learning, which trains multiple related tasks simultaneously, aiming to enhance the model's generalisability while also reducing bias in two ways. Firstly, when the same hidden layers support multiple related tasks (Ruder (2017)), the network must learn features that work across various contexts. Benefiting one task over another is not the goal, as this would diminish performance on other tasks. Joint training of the tasks improves generalisability, as each task regularises the others (Caruana (1993)). Secondly, training a multi-task model requires more diverse data than a single-task model; in addition to medical image data, demographics are also included in the training. This learning approach exposes the model to a broader range of data, which may help reduce the influence of patterns that are prominent only in a subset of the data. Having more data lets multi-task models build stronger, more general features that work across several tasks, helping prevent overfitting (Zhang and Yang (2022)). This suggests that incorporating an auxiliary task, specifically addressing potential bias factors such as age and sex, alongside the primary binary classification (malignant or not), could help mitigate bias.

In addition to the standard multi-task approach, our study also employs an adversarial model approach, a special variant of multi-task learning. As previously demonstrated

(Adeli et al. (2021); Abbasi-Sureshjani et al. (2020)), this strategy reduces output bias to some degree through adversarial training. The model aims to minimise bias by decreasing the mutual information between learned features and the protected attribute, employing a negative-squared Pearson correlation loss for age and binary cross-entropy for sex.

We use the auxiliary head's performance as a diagnostic tool for the reinforcing model. Moderate to high accuracy confirms that the demographic signal has been learned and that regularisation is active, serving as a direct indicator that the debiasing mechanism is functioning properly.

Studies have explored different approaches to handling demographic attributes in model training. Some use demographics within multi-task learning settings (Liu et al. (2019)), where attributes reinforce diagnosis during optimisation. This contrasts with more recent adversarial strategies (Adeli et al. (2021); Abbasi-Sureshjani et al. (2020)) that specifically aim to reduce representation bias by preventing models from predicting sensitive attributes. Additionally, representation bias can be confounded by correlations between demographics and imaging characteristics, leading to shortcut learning. These characteristics include variations in imaging devices (such as different scanner types or image acquisition protocols) and technical artefacts such as surgical markers or medical instruments (Willemink et al. (2020); Jiménez-Sánchez et al. (2023); Gichoya et al. (2022b); Bissoto et al. (2020)). For example, Bevan and Atapour-Abarghouei specifically demonstrated how these technical artefacts can introduce bias in the classification of skin lesions, developing methods to identify and mitigate their impact (Bevan and Atapour-Abarghouei (2023)). In such cases, addressing representation bias requires considering multiple confounding factors, as balancing data for one demographic attribute may leave other sources of bias unaddressed.

We aim to comprehensively evaluate AI-based skin-lesion classification models across demographic groups, with a particular focus on identifying and mitigating representation bias. Whereas Sies et al. examined a market-approved CNN in its uncontrolled training set and considered only sex bias (Sies et al. (2022)), we deliberately construct subsets with exact sex and age ratios to study how the combination of demographic skews affect performance.

3. Methods

To evaluate the impact of demographic imbalance in training data on skin-lesion classification, we conducted two parallel experiments: one manipulating the distribution of patient sex and another modifying the age distribution. In the sex-based analysis, we created datasets with different male-to-female ratios. In the age-based analysis, we built datasets with skewed age profiles, favouring younger, older, or balanced age groups, while keeping a 1:1 sex ratio. Both analyses followed the same methodological pipeline, using linear programming, with the only difference being the demographic attribute constrained during dataset creation. We first describe the data collection and preprocessing steps, then outline the model architectures and evaluation methods.

3.1 Data

We used three skin-lesion datasets: a curated ISIC subset (for training, validation, and internal testing), plus PAD-UFES-20 and DERM7PT (for external testing only). The ISIC subset was derived from the full archive after preprocessing (Section 3.1.1), with controlled demographic distributions for sex and age. Figure 1 illustrates representative samples. Dermoscopic images (ISIC and DERM7PT, respectively, in the left and middle panels) show greater detail and subsurface structures, particularly with polarised dermoscopy, potentially improving diagnostic accuracy (Kittler et al. (2002)). Smartphone images (PAD-UFES-20) exhibit greater variation in lighting, angle, and background.

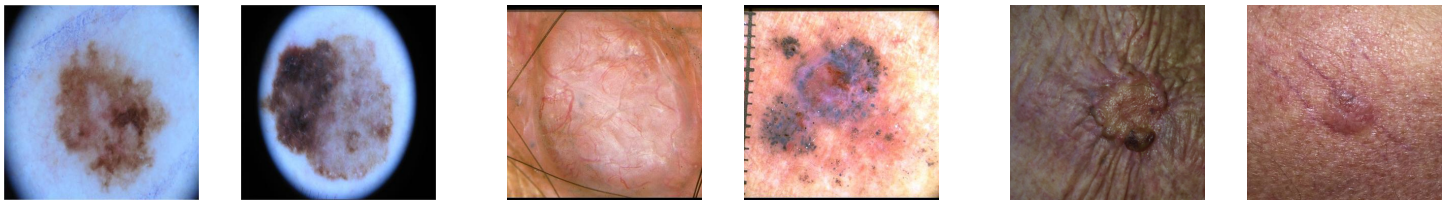
3.1.1 Collection and preprocessing

ISIC based dataset We used the ISIC archive's gallery browser (Gutman et al. (2016); Codella et al. (2018, 2019); Tschandl et al. (2018); Combalia et al. (2019); Rotemberg et al. (2021); ISIC2024), which contained 81,155 dermoscopic images of skin lesions with associated age and sex metadata. The archive was queried for dermoscopic images with diagnoses of "benign" or "malignant" in all age groups and both sexes, yielding 71,035 images (62,439 benign, 8,596 malignant). After data collection, we performed several preprocessing steps to ensure data quality. First, we removed cases lacking age attribute values, leaving 70,843 lesions (62,291 benign and 8,552 malignant). We then removed duplicate images by comparing MD5 hash-values, following Cassidy's method (Cassidy et al. (2022)). After duplicate elimination, 69,982 lesions remained (61,472 benign and 8,510 malignant). Finally, we identified multiple images of the same patient (multiplets) using patient ID attributes and excluded them, resulting in 35,884 lesions (28,810 benign and 7,074 malignant). This removal reduces bias by preventing a single patient from disproportionately influencing the model and eliminates the risk of data leakage across train/validation/test splits. Among the benign lesions, 13,207 were from female patients and 15,603 from male patients. Among malignant lesions, 3,012 were from female patients and 4,062 from male patients.

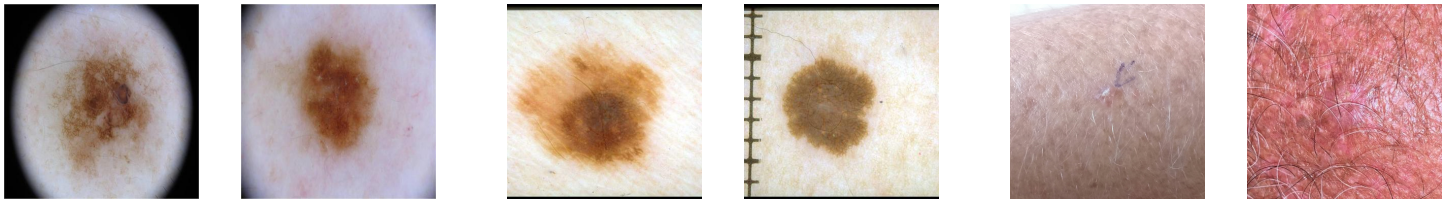
PAD-UFES-20 based dataset For external validation, we used the PAD-UFES-20 dataset (Pacheco et al. (2020)),

Table 1: Overview of the curated skin-lesion datasets used in this study.

Dataset	Modality	Total samples	Malignant (female/male)	Benign (female/male)	Age info	Geographic origin
ISIC	Dermoscopic	35,884	7,074 (3,012/4,062)	28,810 (13,207/15,603)	Yes	Global
PAD-UFES-20	Smartphone	1,179	833 (401/432)	346 (198/148)	Yes	Brazil
DERM7PT	Dermoscopic	1,011	294 (160/134)	718 (362/355)	No	Italy



(a) F. malignant (b) M. malignant (e) F. malignant (f) M. malignant (i) F. Malignant (j) M. malignant



(c) F. benign (d) M. benign (g) F. benign (h) M. benign (k) F. benign (l) M. benign

Figure 1: Comparison of skin lesion images: The left panel shows ISIC dermoscopic images, the middle panel presents four representative lesions DERM7PT dermoscopic images, and the right panel displays PAD-UFES-20 smartphone-captured images. In each panel, the top row contains malignant lesions from a male (M.) and a female (F.) patient, while the bottom row shows benign lesions from male and female patients, illustrating the visual characteristics across the different sources.

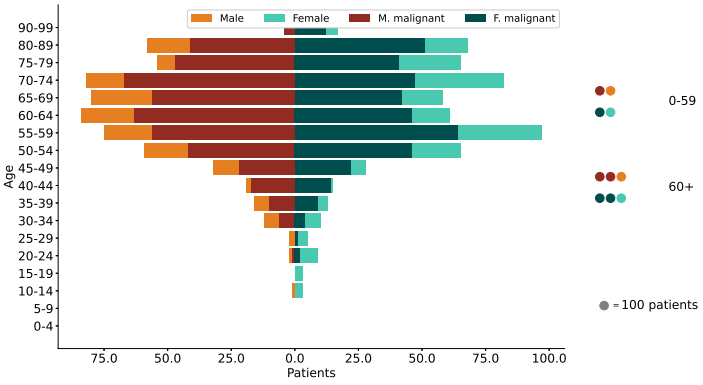


Figure 2: Age distribution across sex and diagnosis type in the curated **PAD-UFES-20** dataset, showing the breakdown between male and female patients with benign versus malignant skin lesions. For a detailed breakdown of lesion counts, see the Appendix C.

which comprises clinical skin-lesion photographs taken with smartphones from patients in Brazil. The original collection comprised 2,298 records. To prepare the data for cross-validation, we performed a sequence of cleaning operations to ensure completeness, consistency, and nonredundancy. First, we removed all entries lacking a sex label (804 rows in total), leaving a fully sex-annotated cohort (741 male and 753 female patients). No records were missing age information, eliminating the need for further imputation. Next, we excluded any lesion without an associated biopsy result, thereby ensuring that every sample used for model evaluation had a definitive pathological ground truth; this filtering reduced the set to 1,179 cases. We then consolidated diagnostic labels into two broad categories: malignant (Melanoma, Basal Cell Carcinoma, Squamous Cell Carcinoma) and benign (Actinic keratosis, Nevus, and Seborrheic keratosis). Finally, we checked for duplicate entries representing the same patient-lesion pair and found none. The final curated dataset comprises a malignant subset of 432 male and 401 female patients, and a benign subset of 148 male and 198 female patients, resulting in a dataset of 1,179 unique records. Figure 2 illustrates the age distribution stratified by sex for both malignant and benign cases in the curated dataset.

DERM7PT based dataset We used the publicly released dermoscopic collection (Kawahara et al. (2018)), comprising 1,011 cases originally curated for the Interactive Atlas of Dermoscopy by Argenziano et al. (Argenziano et al. (2000)). Each case includes a dermoscopic image, a clinical image, patient metadata, and eight labels (seven 7-point checklist criteria plus diagnosis). Sex metadata is available for all samples, though age information is absent. We grouped

diagnostic codes into two categories: benign lesions (nevi, dermatofibromas, lentigines, melanoses, vascular lesions, and seborrheic keratoses) and malignant lesions (basal cell carcinoma and all melanoma subtypes, including in situ, invasive, and metastatic). The malignant subset comprises 160 female and 134 male patients, while the benign subset includes 362 female and 355 male patients. For analysis, we utilised only the dermoscopic images.

Table 1 provides an overview of the curated datasets used in this study, including their modalities, sample sizes, geographic origins, and demographic distributions.

3.1.2 Dataset creation

We developed a method (Raumanns et al. (2025)) to create diverse dataset compositions using linear programming (LP), a standard mathematical optimisation technique. Our pipeline consists of two steps: (1) generating a demographically controlled subset using an LP model, and (2) splitting it into training, validation, and hold-out test sets. We applied this exact pipeline to every experiment, whether adjusting the male-to-female patient ratio or reshaping the age-group distribution. We chose LP over random sampling. LP exactly satisfies multiple demographic constraints (sex, age, lesion diagnosis) while selecting the largest possible subset meeting those ratios. Random sampling approximates the target distribution but often discards or duplicates rare cases. It cannot guarantee specific subgroup constraints (e.g., dark-skinned males aged 50–60 with malignant lesions). The accurate and reproducible control provided by LP is therefore essential for rigorous bias and fairness analysis. It can also be easily extended with additional constraints that random sampling cannot accommodate. In what follows, we describe the specific steps used to create the datasets for each age- and sex-related experiment.

Dataset composition using linear programming The goal of the LP model is to maximise the number of instances of skin lesions within defined constraints, as we express below:

$$
\begin{aligned}
&\text{Find a vector } x \text{ (decision variables)} \\
&\text{that maximises } f = x_1 \text{ (objective function)} \\
&\text{subject to } a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n \leq b_i \text{ (constraints)} \\
&\qquad \text{for } i = 1, \ldots, N_i \\
&\text{and } x_j \geq 0 \text{ (non-negativity constraints)} \\
&\qquad \text{for } j = 1, \ldots, N_j
\end{aligned}
$$

The model has N_j decision variables $(x_1, \ldots, x_{N_j})$ and N_i constraints. Each decision variable corresponds to specific categories (e.g., benign lesions in female patients aged > 60 years). The objective function maximises the count of malignant instances x_1. In the ISIC archive, there are fewer malignant instances than benign ones, and the goal is to achieve a balance between the two. The constraints enforce bounds on individual groups and maintain inter-group ratios. Representative groups include all benign lesions, all females over 60 years old, and all males under 60 years of age. A key constraint maintains class balance by ensuring an equal number of malignant and benign lesions ($x_1 - x_2 = 0$). Non-negativity constraints prohibit negative values for all decision variables. The complete LP formulation is detailed in Appendix A.

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 2 shows the result of the LP model for the different datasets. We adopted a procedure to obtain the final solution for each distribution, consisting of the following steps. First, we solved the LP model to identify the optimal composition of a balanced test set while maximising the number of malignant lesions. From this balanced set, we reserved one-eighth as a hold-out test set. Second, we recalibrate the upper-bound constraints using the lesion counts observed in the hold-out set. With these updated bounds, we resolved the LP model to derive the final solutions for the various distributions. Third, after obtaining solutions from the LP model, we determined the minimum number of malignant instances in all datasets. Fourth, we scaled each dataset proportionally to the minimum value, preserving demographic distributions while ensuring comparability.

Sex distribution analysis We created seven distinct training and validation datasets to analyse sex-related biases with varying patient ratios of female (F) to male (M). Each of the seven dataset instances was created using a distinct random seed. For each seed, we first created a hold-out test set; the remaining data were then shuffled and split strictly into an 80% training subset and a 20% validation subset while preserving the original demographic ratios. There was no overlap of lesions between the training and validation sets for any given seed, preventing overlapping samples from inflating performance. We maintained the same number of malignant and benign lesions in all datasets and balanced age distributions with the same numbers of patients below and above 60 years (median age) for each sex.

The datasets consisted of a M100 set (100% male patients), a F100 set (100% female patients), a F95M5 set (95% female, 5% male patients), a F75M25 set (75% female,25% male patients), a F50M50 set (50% female, 50% male patients), a F25M75 set (25% female, 75% male patients), a F5M95 set (5% female, 95% male patients), and a separate balanced test set that matches the distribution of the F50M50 (equally-split) dataset.

Figure 4 illustrates the age distributions across the sex-

Table 2: **ISIC-based** datasets are distributed amongst malignant, benign, male patients (M), and female patients (F) categories for both training and validation. Bold value indicates the minimal malignant-lesion count.

	M100	F5M95	F25M75	F50M50	F75M25	F95M5	F100
Malignant (M/F)	**2206** (2206/0)	2322 (2206/116)	2941 (2206/735)	4412 (2206/2206)	3235 (809/2426)	2554 (128/2426)	2426 (0/2426)
Benign (M/F)	2206 (2206/0)	2322 (2206/116)	2941 (2206/735)	4412 (2206/2206)	3235 (809/2426)	2554 (128/2426)	2426 (0/2426)

based training datasets and the balanced test set.

Age distribution analysis Similar to the sex-distribution analysis, we built the training, validation, and test sets for the age-focused experiments using an LP model (see Appendix B for full details). We defined three age-distribution schemes, each spanning the same five age brackets. Table 3 shows the definition and proportions of the five age brackets (A_1–A_5) for each of the three schemes. In the YOUNGER scheme, the majority of samples come from the youngest age brackets, with the proportion gradually decreasing toward older age groups. The BALANCED scheme allocates samples uniformly across all five brackets. Finally, the OLDER scheme is the inverse of the younger-skewed arrangement, concentrating most samples in the oldest age groups. Each distribution enforces a strict 1:1 balance between malignant and benign lesions and a 1:1 balance between male and female patients. For each scheme, we generated five independent instances using five different seeds. For each seed, we divided the data into non-overlapping training and validation sets and reserved a holdout test set with a uniform age category distribution (based on the BALANCED scheme). See Table 4 for the count of skin lesions in each split.

Table 3: Proportion of the five age groups (A_1–A_5) across the three schemes (YOUNGER, BALANCED, and OLDER) for the **ISIC-based** data. The age brackets are defined as follows, where a represents the patient's age in years: $A_1 = 0 \leq a \leq 50$, $A_2 = 51 \leq a \leq 60$, $A_3 = 61 \leq a \leq 70$, $A_4 = 71 \leq a \leq 80$, and $A_5 = a \geq 81$.

	A_1	A_2	A_3	A_4	A_5
YOUNGER	0.35	0.30	0.20	0.10	0.05
BALANCED	0.20	0.20	0.20	0.20	0.20
OLDER	0.05	0.10	0.20	0.30	0.35

Table 4: Division of the in **ISIC-based** dataset into training and validation subsets for all three age-distribution schemes. The numbers shown in the diagram indicate the count of skin-lesion images in each split.

	Training	Validation	Testing
YOUNGER	4740	1192	
BALANCED	4120	1040	1020
OLDER	2348	600	

3.2 Model

Using our carefully constructed datasets, we implemented three different architectures based on the ResNet50 model (He et al. (2016)). We selected ResNet50 for its proven performance in medical imaging and widespread adoption (Xu et al. (2023)), enabling meaningful study comparisons. These architectures evaluate different approaches to handling demographic information:

The single-task baseline model The single-task baseline model, enhanced with two fully connected layers, uses a sigmoid activation function and binary cross-entropy loss.

The multi-task reinforcing model The multi-task "reinforcing" model with three layers added to the convolutional base, produces two outputs: one for classification and another for the demographic attribute (either sex or age, depending on the specific experiment). Please note that we use the term reinforcing here in the sense of "strengthening influence", not in the reinforcement learning (RL) sense. When the attribute is sex (binary: male/female), we used a binary cross-entropy loss (L_c) and a sigmoid activation function for both heads. For age, which is treated as a continuous variable, we replace the binary loss with a mean-squared error loss applied to the normalised age value. Both heads (primary and demographic) receive equal weighting in the overall objective, while the classification head continues to use its standard loss function. When we use the auxiliary head for age prediction, we first normalise the age labels to the unit interval ([0, 1]) using the minimum and maximum ages observed in the dataset. During inference, we denormalise the sigmoid output back to the original age scale. Because the admissible age range is fixed and known a priori, this normalisation-denormalisation procedure preserves the semantic meaning of the prediction while keeping an identical model architecture for both auxiliary tasks.

The multi-task adversarial model The multi-task adversarial model was implemented following the methodology of Adeli et al. and Abbasi-Sureshjani et al. (Adeli et al. (2021); Abbasi-Sureshjani et al. (2020)), using a network with a shared feature encoder and two classifier heads. One classifier targeted skin cancer classification; the other predicted confounders such as sex or age. We used the ResNet architecture to compare performance with baseline and reinforcement models in a systematic way. We trained the skin-cancer classifier and its encoder with a stan-

dard cross-entropy loss (L_c). The choice of bias-predictor head loss (L_{bp}) depended on the demographic variable under study: for age-distribution experiments we employed an age bias predictor head whose loss was defined as the negative-squared Pearson correlation coefficient loss (this worked for protected attributes that are continuous or ordinal (Adeli et al. (2021)), while for sex-distribution experiments we used L_{bp} as a binary cross-entropy loss, reflecting the binary nature of the attribute.

To reduce the predictiveness of the encoded features, we adversarially adjusted the encoder using a third loss term (L_{br}), with λ governing the penalty for accurate demographic predictions: $L_{br} = \lambda L_{\text{bp}}$, following the practice of Abbasi-Sureshjani and colleagues (Abbasi-Sureshjani et al. (2020)).

Model training parameters We optimised the single-task model using a grid search over random seeds, learning rates, and momentum values, selecting the combination that yielded the highest validation performance across all experiments. Subsequently, these optimised parameters were then applied to the reinforcement and adversarial models for a fair comparison. The optimal hyperparameters, selected via validation set performance, are as follows:

Pre-training:	ImageNet
Input size:	384×384 pixels
Max epochs:	40
Batch size:	20
Learning rate:	2.0×10^{-5}

To mitigate overfitting and improve model robustness, we implemented an early stopping technique (patience = 10 epochs) and data augmentation techniques. Following prior implementations in the literature, we implemented our baseline and reinforcement models in Keras with the TensorFlow backend (Géron (2022)), while our adversarial model was implemented in PyTorch (Paszke et al. (2019)) to maintain consistency with existing adversarial learning frameworks.

4. Experiments

4.1 Demographic bias evaluation

Defining test, training and validation bundles We took steps to ensure that any observed performance differences between models are attributable to the models themselves and the underlying data distributions, rather than to inconsistencies in how we split the data. By fixing a single random seed for each run, we use the same hold-out test set across distributions, making cross-distribution comparisons meaningful. Importantly, we deliberately construct the hold-out test set to be balanced across relevant subgroups (such as age or sex). This balanced test set removes bias toward any particular segment and allows us to isolate the effect of bias in the training data itself. We acknowledge, however, that measuring "true" behaviour across the entire population would require a test set whose distribution matches the expected real-world population; our balanced test set is chosen specifically to evaluate bias rather than to predict real-life performance. When we generate the training and validation partitions with the same seed, we ensure that these splits faithfully reflect the target distribution. By repeating the entire process with several different seeds, we reduce the influence of random variation in the data. Implementation is illustrated in Figure 3. Our experimental workflow enables both within-distribution and cross-distribution model comparisons.

Sex distribution For a thorough evaluation, we created five bundles for each of the seven distributions: F100, F5M95, F25M75, F50M50, F25M75, F5M95, and M100, resulting in 35 bundles (seven distributions $\times$ five seeds). We assessed AUC overall and within male and female subgroups for each learning strategy and dataset combination. Using three learning strategies on 35 bundles, we conducted 105 experiments to evaluate model performance within and across all seven distributions. Notably, multi-task models cannot unlearn constant protected attributes (as in M100 and F100 experiments). These edge cases serve to stress-test the training pipeline's stability when demographic attributes are absent.

Age distribution In our age experiments, we conducted comprehensive evaluations using age-stratified data sets. We generate three distinct data distributions, across five age categories ($A_1 - A_5$): YOUNGER (predominantly young patients), BALANCED (evenly distributed age categories), and OLDER (predominantly elderly patients). For each configuration, we evaluated the three model architectures. Performance was measured using AUC scores, with results visualised across different age distributions and model architectures. We conducted 45 experiments in total (15 bundles and three learning strategies) to evaluate model performance within and across the three distributions.

4.2 Reinforcing model: Auxiliary head analysis

We evaluate the auxiliary prediction head of the reinforcing model to serve two purposes: (1) assessing whether the multi-task architecture successfully learns the demographic signal, and (2) providing a mechanistic explanation for bias mitigation in the primary task. Specifically, if the auxiliary head fails to learn the demographics, the reinforcing model loses its regularisation effect, which may explain observed failures in bias mitigation. An auxiliary-head analysis was omitted for the adversarial model because the original net-

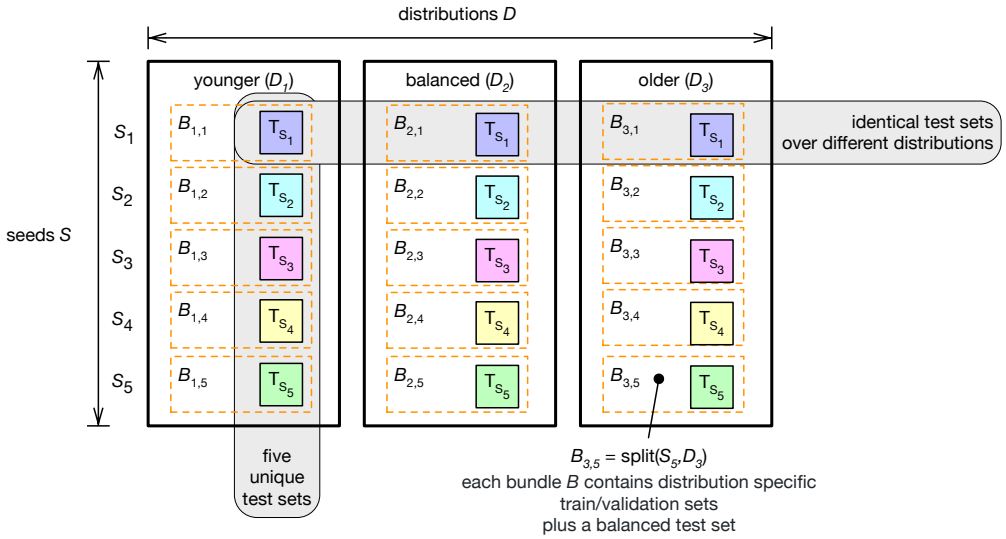


Figure 3: Experimental workflow (age analysis; sex analysis is analogous): for each distribution D, we build five independent bundles B, each split into mutually exclusive test, training, and validation sets with a fixed seed S. The same test set is shared by all distributions for a given seed, and training/validation splits use the same seed. The distribution properties of the training/validation splits are consistent with the distribution under test. All three models are evaluated on this test set, allowing within- and cross-distribution comparisons, and the process is repeated with multiple seeds to reduce random effects.

work implementation did not provide demographic outputs.

Evaluation of the sex-prediction head We evaluated the auxiliary sex-prediction head of the multi-task reinforcing model using two complementary metrics. First, we computed the AUC to assess discriminative ability. Second, we measured the Brier score (Rufibach (2010)) separately for male and female patients. We interpret the Brier score (β) for binary classification as follows: strong calibration for $0 \leq \beta < 0.05$, moderate for $0.05 \leq \beta < 0.15$, weak for $0.15 \leq \beta < 0.25$, poor for $0.25 \leq \beta < 0.35$, and very weak for $\beta \geq 0.35$. Lower Brier scores indicate tighter alignment between predicted probabilities and observed outcomes, whereas higher values reflect increasingly poor calibration. We do not evaluate the two edge cases (F100, M100) with a training set containing only one sex; the auxiliary head cannot learn a useful decision boundary.

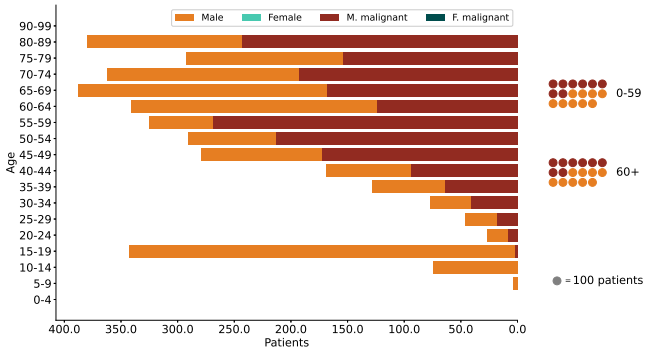
Evaluation of the age-prediction head For the auxiliary age-prediction head of the multi-task reinforcing model, we used mean absolute error (MAE) as our primary performance measure. MAE computes the average absolute difference between the predicted age and the ground-truth age across all test samples. To disclose any systematic biases throughout the age spectrum, we compute MAE for the five age categories (A_1–A_5, see Table 3). Furthermore, to assess the quality of the age predictions, we computed the Pearson correlation coefficient (ρ) between the predicted ages and the ground-truth ages. This metric quantifies the linear relationship between the two variables and complements the MSE by indicating how well the model captures age trends. We interpret the Pearson correlation coefficient as follows: strong correlation for $0.5 \leq \rho < 1$, moderate correlation for $0.3 \leq \rho < 0.5$, and weak correlation for $0 \leq \rho < 0.3$. Only correlations with $p < 0.05$ were retained for further analyses.

4.3 Cross-dataset evaluation

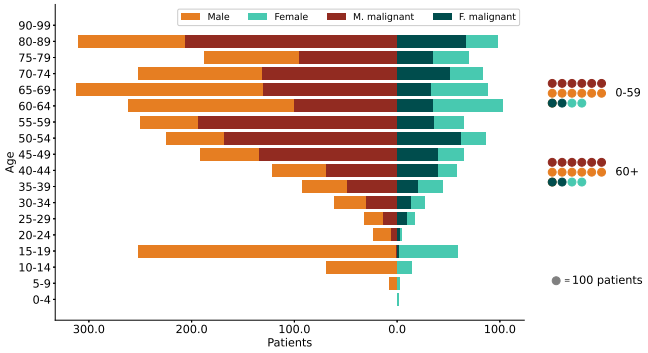
To validate our findings and assess generalisability, we performed additional experiments using two external skin-lesion datasets: PAD-UFES-20 (both sex and age) and DERM7PT (no age information). Using the saved weights from our previously trained models (base, reinforcement, and adversarial), we evaluated them on the external datasets without any further fine-tuning, using the same evaluation metrics. This cross-dataset validation approach provides insight into the robustness and transferability of our models across different patient populations and data collection contexts.

4.4 Exploratory performance assessment

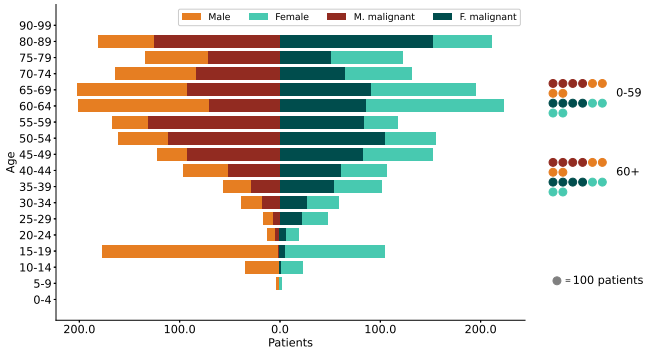
We did not conduct formal statistical testing because our study is exploratory, and the number of observations is minimal. Applying p-value–based tests to the model performances would yield unstable estimates that provide little trustworthy insight into whether any actual effect exists. Consequently, we refrained from labeling results as "significant" or "non-significant." As Amrhein et al. highlighted, such dichotomous labeling often leads to misinterpretation,



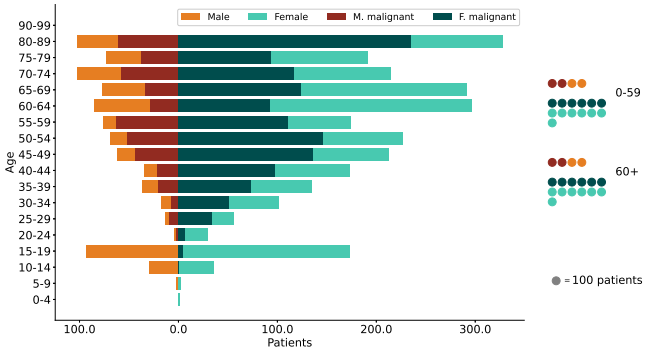
(a) M100



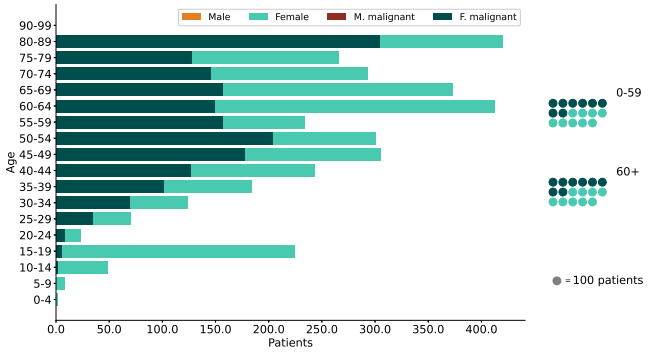
(b) F25M75



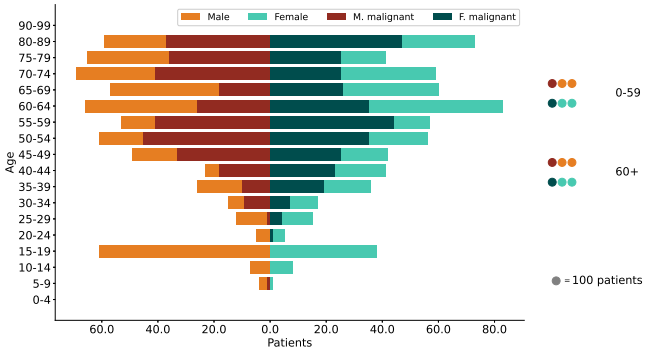
(c) F50M50



(d) F75M25



(e) F100



(f) Balanced testset

Figure 4: Age distributions in the ISIC-based datasets range from M100 to F100 (a to e), with a balanced test set (f). The training sets (3,528 records, approximately 80% of the total) and corresponding validation sets (880 records, approximately 20%) span distributions from M100 to F100 and maintain similar population compositions. The total of 4,412 records ($2 \times 2,206$ lesions) reflects the inclusion of both malignant and benign cases, with the base value of 2,206 taken from Table 2. Test sets contain 1,264 records. The visualised distributions correspond to seed value 1970; distributions for other seeds are equivalent.

and non-significant findings are frequently mistaken for evidence of no effect (Amrhein et al. (2019)). Instead, we present AUC values as boxplots and provide descriptive comparisons across subgroups, allowing readers to assess the magnitude and direction of any observed differences.

5. Results

5.1 Sex-specific model assessment

Figure 5 shows the performance of the model in sex distributions using box plots of AUC metrics for three types of models (base, reinforcement, adversarial). Figure 6 shows the impact of dataset distributions on three learning strategies, reporting AUC scores for both sexes.

Comparable performance at model level Our analysis in Figure 5 demonstrates that all three learning strategies achieve similar levels of effectiveness, showing only slight performance variations. Across the three model architectures, the accuracy scores maintain a consistent range between 0.79 and 0.85. Although there are minor variations between the approaches, none clearly outperforms the others.

Sex-specific training data yields better results Figure 6 shows that all models perform better for male patients in male-only (M100), predominantly male (F5M95) and lightly male-skewed (F25M75) scenarios. The reinforcing and base models show the most pronounced performance gap in the predominantly male dataset. The base and reinforcing show equal performance between subgroups in the balanced, lightly female-skewed (F75M25) and predominantly female (F95M5) scenarios. The base model shows better performance for female patients in the female-only (F100) scenario, while the reinforcing and adversarial models show equal performance between male and female patients. Thus, our models appear more attuned to male patients in mixed-sex training sets, regardless of the percentage of female patients. The best results are achieved when both sexes are trained exclusively on their respective data. We hypothesise that a model trained on a single-sex dataset may specialise in those sex-specific cues and often attains higher accuracy. In mixed-sex training, the network must accommodate both distributions, which can lead to a modest performance dip, typically favouring the more dominant signal.

Base model reveals sex bias We found substantial sex bias in the performance of the base model (see Figure 6). In the male-only and predominantly male scenarios, we observed a substantial performance gap between male and female patients. In the female-only scenario, there is also a performance gap; however, this is less pronounced. We found that the base model performed comparably for male and female patients across balanced, lightly skewed, and predominantly female experiments. We assume the base model binds onto the most prevalent cues in the training set. When the data consist of only one sex or are heavily skewed, it reveals sex-specific visual patterns (e.g., hair density, skin texture) that aid classification, performing well for the majority sex but poorly for the minority. In a balanced, only mildly skewed female dataset, both sexes are equally represented, forcing the model to rely on lesion-intrinsic features for discrimination, leading to comparable accuracy for males and females.

Reinforcement model partially successful in sex bias mitigation When we trained the model on male majority data, we observed performance disparities between the sexes (see Figure 6). With balanced training data, the reinforcement model successfully mitigates sex-based bias. Notably, this same bias reduction effect is observed in female-majority training sets as well. We hypothesise that the reinforcing multi-task model can reduce sex bias only when its auxiliary sex-prediction head receives sufficiently informative female patient-related signals. In the only-male and predominantly-male scenarios, the encoder overfits to male-specific cues because the auxiliary head lacks enough female examples to learn a meaningful discriminator. Conversely, with balanced or mostly-female settings, the auxiliary head can learn a reliable sex classifier; its loss then regularises the shared encoder towards sex-invariant representations, thereby reducing bias.

Adversarial model reduces sex bias in predominantly female training scenarios. The adversarial model reduces sex bias in scenarios with predominantly female patients but is less effective in other scenarios, often favouring male patients. Its performance varies between experiments and datasets (see Figure 6). We suspect that complex anatomical confounders, such as variations in body hair distribution or skin texture, continue to act as strong proxies for sex in mixed populations, resisting the adversarial removal of these features.

5.2 Age-specific model assessment

Figure 7 presents a comparative analysis using box plots to illustrate AUC metrics across YOUNGER, BALANCED, and OLDER datasets. For a more granular understanding, Figure 8 provides an age-stratified evaluation using five distinct age categories (A_1–A_5), demonstrating how each model architecture performs when trained on differently distributed datasets and evaluated against a balanced test set.

Comparable overall model performance Looking at the overall performance of the model in all experiments (Figure 7), the adversarial model shows the highest variance

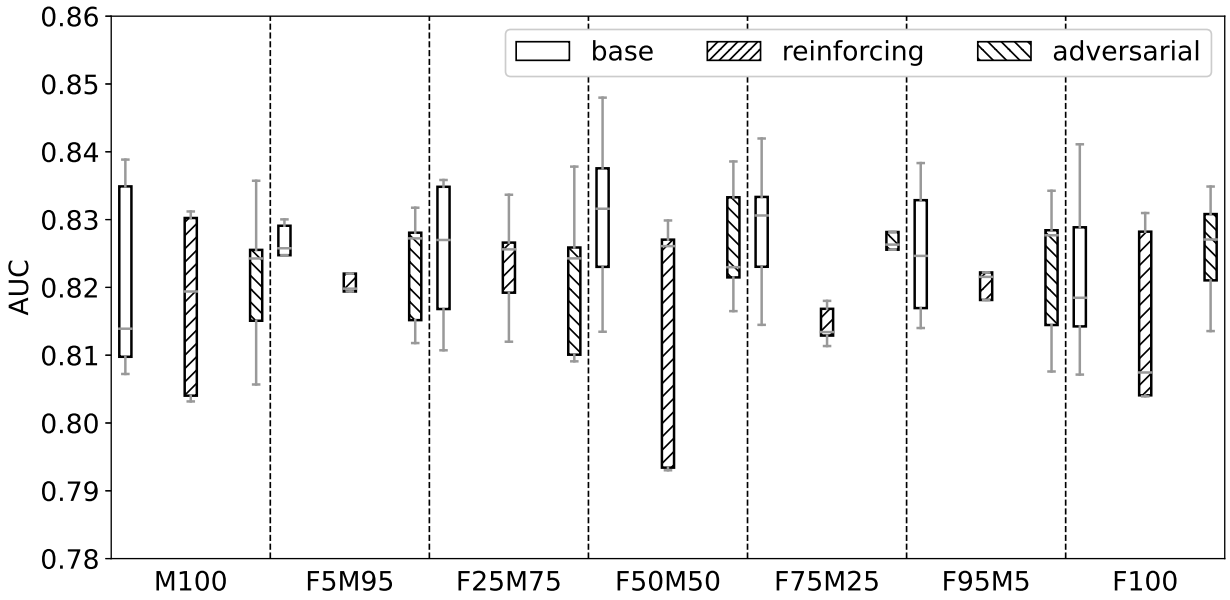


Figure 5: Comparison of model performance across sex distributions using datasets generated from the curated **ISIC dataset**. Box plots display AUC metrics for three model architectures (base, reinforcing, and adversarial) trained and validated on sex-biased datasets (M100, F5M95,F25M75, F50M50, F75M25, F95M5, F100) and evaluated on a balanced dataset.

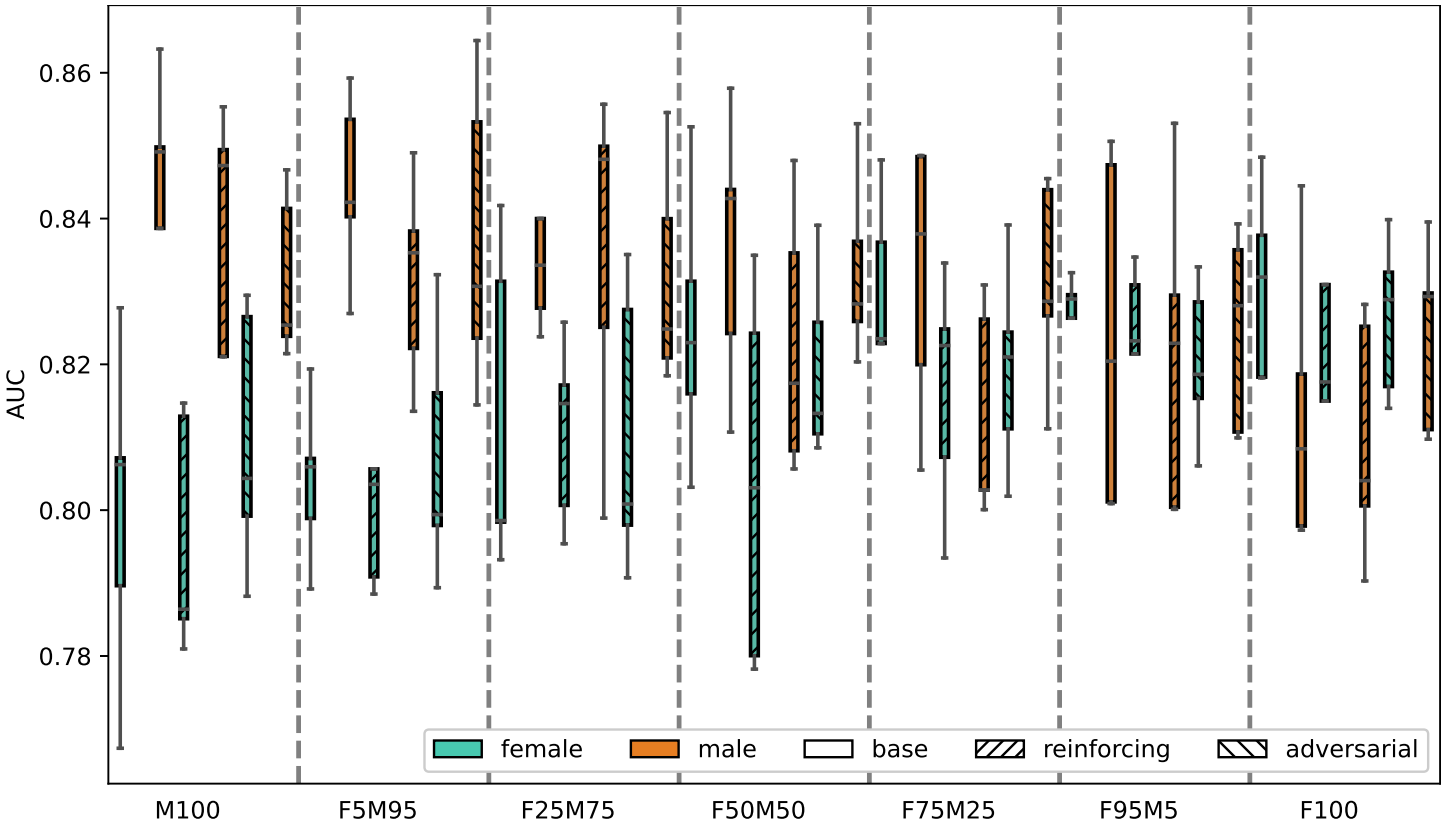


Figure 6: The AUC score varies based on data splits ranging from only male patients (M100) to only female patients (F100) in the **ISIC dataset**. We show base, reinforcing and adversarial model performance for female and male patient subgroups.

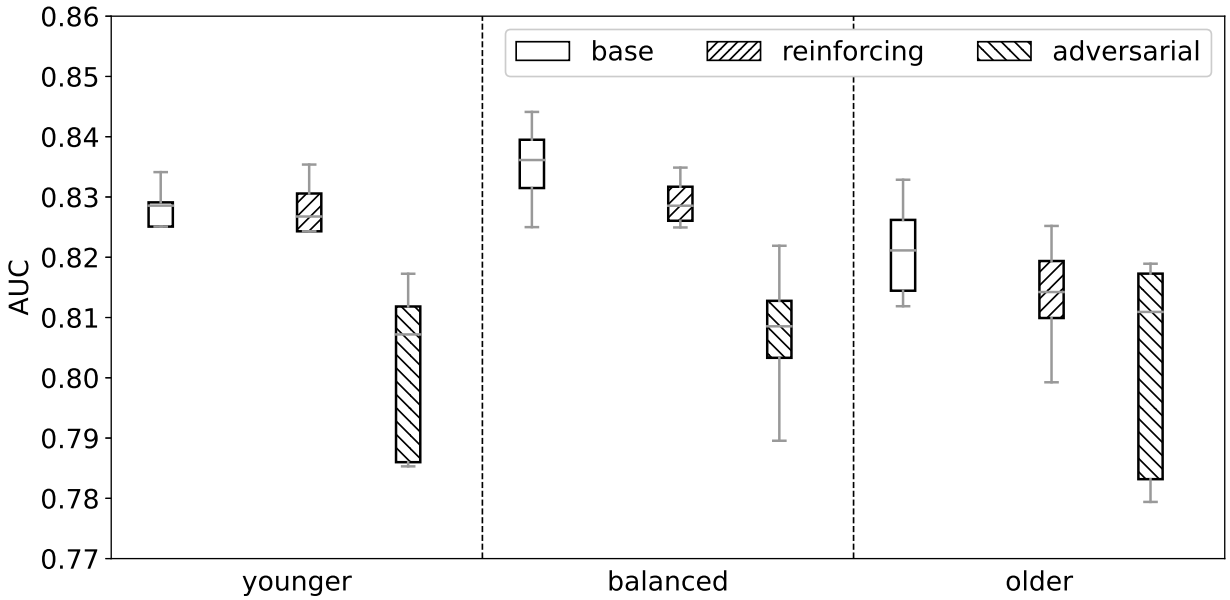


Figure 7: Comparison of model performance across different age distributions using curated **ISIC dataset**. Box plots show the AUC metrics of three model architectures (base, reinforcing, and adversarial) trained and validated on age-biased datasets (YOUNGER, BALANCED, OLDER) and evaluated on a balanced test set.

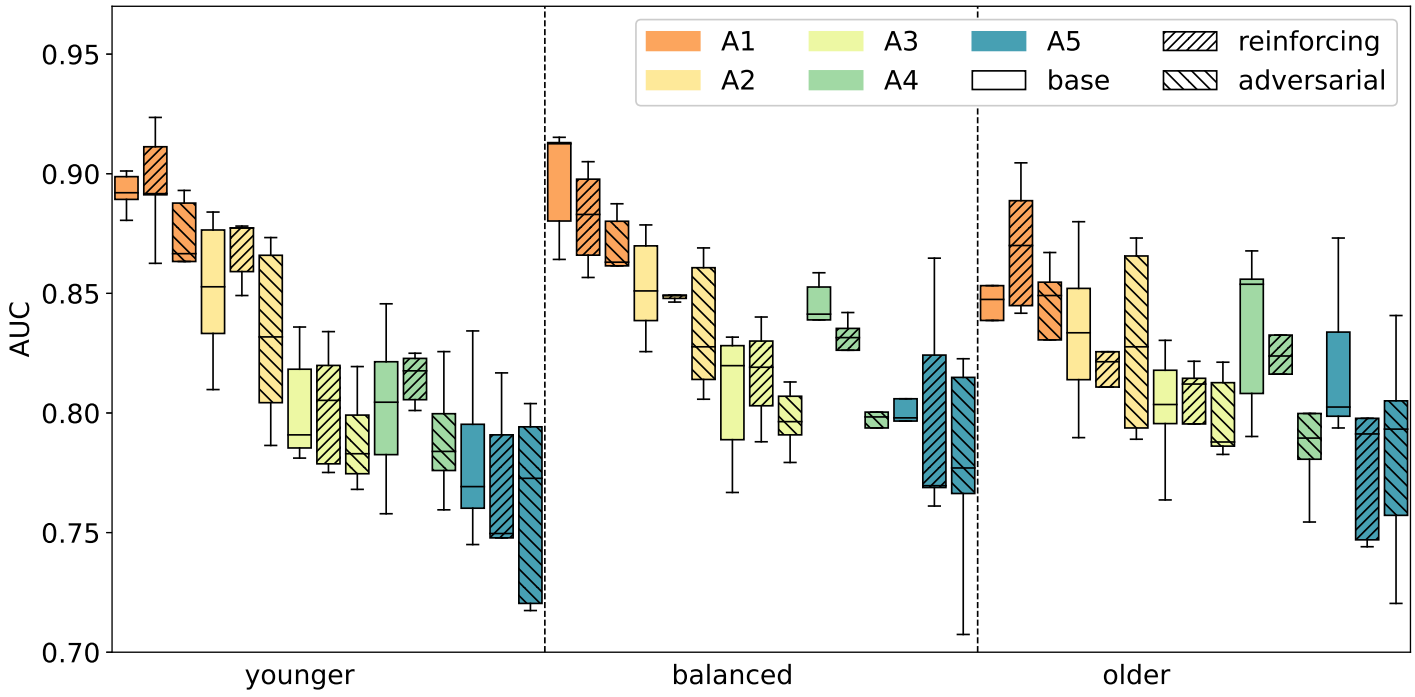


Figure 8: Age-stratified (using five age categories ($A_1 - A_5$)) model evaluation across datasets with varying age distributions based on the curated **ISIC dataset**. The analysis compares AUC scores for three model architectures (base, reinforcing, and adversarial) using age-biased training sets (YOUNGER, BALANCED, OLDER). Each model was evaluated on a balanced test set, showing performance variations across different age distributions. The age brackets are defined as follows, where a represents the patient's age in years: $A_1 = 0 \leq a \leq 50$, $A_2 = 51 \leq a \leq 60$, $A_3 = 61 \leq a \leq 70$, $A_4 = 71 \leq a \leq 80$, and $A_5 = a \geq 81$.

in performance in different seeds compared to the other two strategies, particularly in the YOUNGER and OLDER cases. All three model approaches demonstrate comparable base performance levels with AUC scores falling within a 0.06 range, with both the base and reinforcement models showing a slight advantage compared with the adversarial model.

Declining trend across categories As shown in Figure 8, all three distributions show a decreasing AUC trend, with the youngest age bracket achieving the highest AUCs and the oldest the lowest. In the balanced distribution, the decreasing trend does not hold for the A_4 age bracket in either baseline or reinforcing models. A_4 performance is higher in the balanced case than in the younger-age distribution. In the older age distribution, A_1 bracket models show a slight performance decrease compared to the younger and balanced distributions. This decline is most pronounced for the baseline and adversarial models. We assume the performance drop with increasing age is due to the nature of older skin, which exhibits more heterogeneous visual traits (such as wrinkles, pigment changes, and vascular alterations) that mask lesion cues.

Strong performance for younger age categories Models trained on the balanced dataset show the highest AUC for age category A_1, but experience a small performance drop for age category A_2 compared to models trained on the YOUNGER dataset (see Figure 8). In contrast, the AUC values for the age categories A_3, A_4, and A_5 generally fall below 0.85. We hypothesise that the model excels in the youngest age group because young skin presents the most explicit lesion cues, with fewer wrinkles, pigment variations, or vascular artefacts to obscure the diagnostic signal, and the balanced training set supplies fewer examples of these clean patterns.

Performance improvement of balanced models Figure 8 illustrates that the base and reinforcing models trained on a balanced dataset show improved performance for the A_4 age category compared to the same models and age category trained in the younger dataset. We assume the balanced set provides enough examples that force the encoder to learn features that work across age groups. In the younger-skewed data the model overfits to smooth, youthful textures, so its performance drops on the A_4 group. Adding balanced age samples thus improves AUC for that category.

5.3 Sex-prediction head evaluation

Table 5 reports the auxiliary sex-prediction head's performance in the reinforcing model, including overall AUC and Brier scores for male and female subgroups. Predictive performance was limited across all skewed distributions but reached reasonable accuracy in the balanced case (AUC $= 0.732$), indicating that the encoder successfully learns sex-related features when the data is balanced. This aligns with the observed reduction in sex bias for the reinforcing model under balanced settings. However, in skewed scenarios (e.g., F5M95), the auxiliary head's performance becomes highly asymmetric: it predicts the majority sex with high confidence (a weak Brier score) but fails on the minority sex (a strong Brier score). This inability to learn a robust sex discriminator explains why the reinforcing model cannot effectively regularise the encoder under skewed distributions, leading to persistent bias gaps (see F5M95 in Figure 6).

Table 5: Performance evaluation of the auxiliary sex-prediction head within the reinforcement model. Metrics include overall AUC and Brier scores (β), computed separately for male and female patient subgroups. Bold values denote optimal performance per column (highest AUC, lowest Brier score).

Training dataset	AUC	β Female	Male
F50M50	$\mathbf{0.732 \pm 0.012}$	0.234 ± 0.049	0.238 ± 0.033
F5M95	0.644 ± 0.012	0.959 ± 0.013	$\mathbf{0.001 \pm 0.000}$
F25M75	0.682 ± 0.043	0.730 ± 0.132	0.028 ± 0.021
F75M25	0.691 ± 0.014	0.040 ± 0.012	0.645 ± 0.086
F95M5	0.612 ± 0.012	$\mathbf{0.001 \pm 0.001}$	0.948 ± 0.026

Highest AUC and weak Brier for the balanced training set When we train the reinforcing model with equal numbers of male and female patients, the auxiliary sex-prediction head learns a discriminative representation and achieves the highest AUC among the experiments. Brier scores for the F50M50 case are weak for female and male patients (see Table 5).

Strong Brier for the majority, very weak for the minority When we train on a majority of male patients (and F5M95 and F25M75), the Brier score for males is strong whereas that for females is very weak. When we train with a minority of male patients (F95M5 and F75M25), the opposite occurs (see Table 5).

5.4 Age-prediction head evaluation

We evaluated the auxiliary age prediction head by reporting both the Pearson correlation (Table 6) and the distribution of the mean absolute error (Figure 9). Table 7 provides the overall MAE for the three training-bias configurations.

In the BALANCED training configuration, the auxiliary head shows a moderate Pearson correlation ($\rho = 0.433$), indicating that the model recognises age patterns without relying on them as a dominant shortcut. This supports the hypothesis that balanced sampling forces the encoder to learn features that generalise across age groups. In contrast,

Table 6: Pearson correlation (ρ) between predicted and true ages for the reinforcing multi-task model. Correlation coefficients are reported for three training-bias configurations (younger-skewed, balanced, older-skewed) evaluated on a balanced test set. The table lists the overall ρ as well as the ρ for each age category. "-" indicates too low to be meaningfully reported.

ρ	YOUNGER	BALANCED	OLDER
overall	0.319	0.433	**0.517**
A_1	0.354	0.471	**0.710**
A_2	0.104	0.105	0.078
A_3	-	0.070	-
A_4	-	-	-
A_5	-	0.069	-

the older-skewed training yields a strong correlation ($\rho = 0.710$) for the youngest cohort, suggesting the model relies heavily on shortcut-related age cues when the distribution is imbalanced.

Increasing correlation with older-skewed training The reinforcement model trained on the OLDER dataset exhibits a strong Pearson correlation between predicted and ground-truth ages. In contrast, models trained on the BALANCED and YOUNGER datasets show only moderate and weak correlations, with the BALANCED model achieving a slightly higher correlation than the YOUNGER model (Table 6).

Youngest cohort shows strongest correlation A_1 correlates most strongly with older-skewed data (Table 6).

Middle group shows weakest correlation The middle age groups (A_2–A_4) correlations are uniformly weak and often marked with a dash "-" in (Table 6).

Oldest cohort shows practically no correlation Only the balanced scheme shows a very weak correlation for the oldest cohort (A_5); for the other schemes, the correlation score is marked with a dash "-" in Table 6.

Table 7: Mean Absolute Error (MAE) of the auxiliary age-prediction head for the reinforcing multi-task model under three training-bias configurations.

	YOUNGER	BALANCED	OLDER
MAE	15.335 ± 0.643	$\mathbf{13.409 \pm 1.031}$	14.252 ± 0.586

MAE of age categories We find that the model we trained on a balanced age distribution yields the lowest overall average error (Table 7), whereas the younger-skewed and older-skewed configurations show higher MAE. For the youngest cohort of patients, we see relatively high MAE scores for all three models (Figure 9). The MAE scores of the youngest cohort increase as the number of patients in that cohort in the training set decreases. We observe the same pattern for the oldest patient cohort. However, in skewed training, the oldest cohort scores better when trained with mainly older patients than the youngest cohort trained with mostly younger patients. The oldest age groups (A_4 and A_5) exhibit high MAE when the model is trained mostly on younger patients, while the youngest age groups (A_1 and A_2) show high MAE when trained mainly on older patients. The A_2 age cohort achieves the lowest score when we train with predominantly younger patients. The scores of the A_3 cohort change little in all models. The MAE score of the A_4 cohort decreases as the proportion of A_4 patients relative to the total training population increases. The A_4 age cohort achieves the lowest score when we train the model with mostly older patients.

5.5 Cross-dataset analysis

5.5.1 PAD-UFES-20

Figures 10 and 11 present our model performance evaluation on the PAD-UFES-20 dataset, examining age-stratified and sex-stratified distributions. While Figure 10 breaks down performance across five age categories (A1-A5) for different age-biased training sets, Figure 11 analyses sex-based performance variations across multiple male-female distribution ratios.

External validation shows performance drop During external validation, model performance in both sex- and age-based evaluations showed notably lower metrics compared to internal validation scenarios (Figures 10 and 11).

External validation shows best performance for younger age groups For age-based models, we observe a different pattern in external validation compared to internal validation. The models trained on the YOUNGER, BALANCED, and OLDER cases show similar performance ranges in the three learning strategies for age categories A_2, A_3, A_4, and A_5. Across all configurations, models show notably better performance for the A_1 age category. The adversarial model also shows improved performance for the A_2 age category, achieving results comparable to the performance of the A_1 category (Figure 10).

Sex-based validations show contrasting patterns In internal validation, we observed an X pattern: as the percentage of male patients in the training and validation sets decreased, male patients' performance declined, whereas female patients' performance improved. However, the X pattern is less pronounced than in the internal validation. For adversarial models, we observed a different pattern (see Figure 11): in the internal validation with F75M25 adversarial models, the performance for female patients was

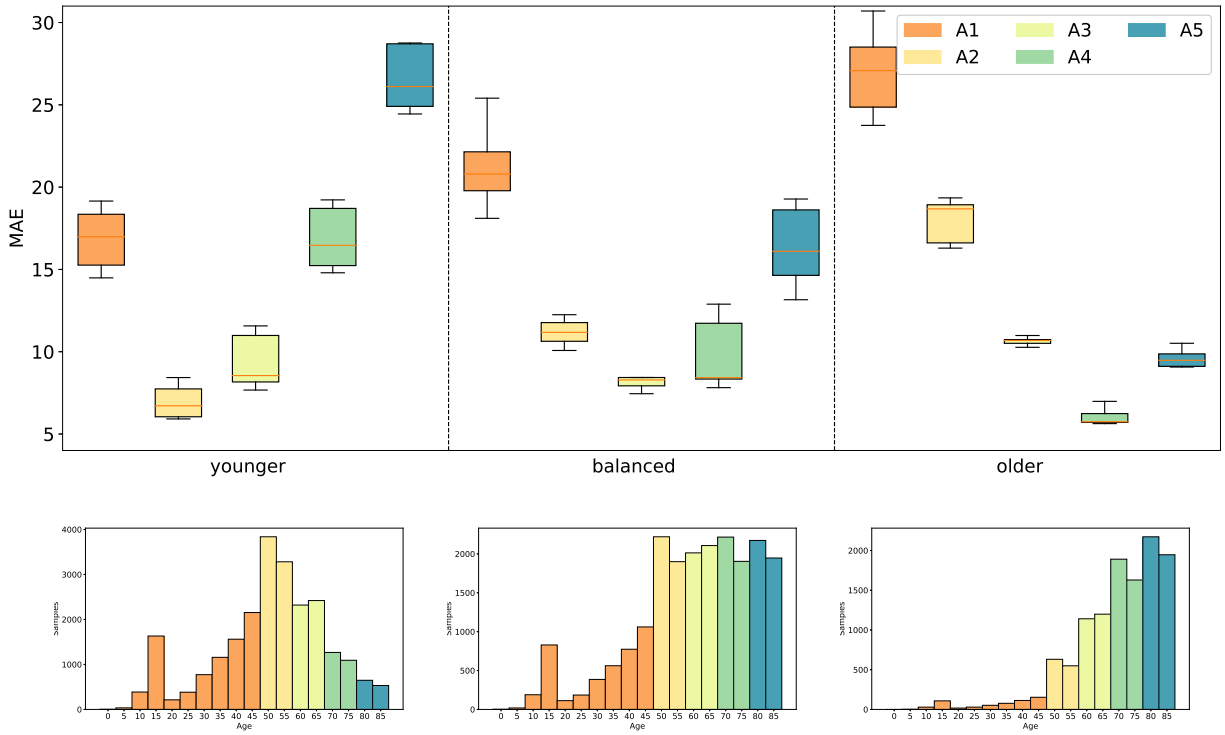


Figure 9: The top figure shows the Mean Absolute Error (MAE) distributions for five age groups (A_1–A_5). Each plot depicts the MAE, calculated from predicted versus reference ages, under one of the three training biases (younger, balanced, older) using the reinforcing model. We evaluated all models on a balanced test set. The bottom three figures show the age distribution for the younger, balanced, and older training datasets (from left to right). The same colour legend that appears in the top panel (mapping A_1–A_5 to their respective colours) is reused for the three lower panels. The age brackets are defined as follows, where a represents the patient's age in years: $A_1 = 0 \leq a \leq 50$, $A_2 = 51 \leq a \leq 60$, $A_3 = 61 \leq a \leq 70$, $A_4 = 71 \leq a \leq 80$, and $A_5 = a \geq 81$.

lower than for male patients, while in the internal validation with F100 adversarial models, the performance for male and female patients was equal. However, in external validation, we observe a notable improvement in performance in female patients.

5.5.2 DERM7PT

External validation shows performance drop During external validation, we observed that the models showed considerably lower AUC metrics than in the internal validation experiments (see Figure 12). Nevertheless, we found that their performance remained higher than the sex-based results obtained on PAD-UFES-20.

Reduced subgroup performance gaps across external datasets We observed that the performance of the male and female sub-groups remained much closer together across the various training scenarios than it did in the ISIC-based experiments, and less pronounced than that seen with PAD-UFES-20. In other words, the sex ratio distribution in the training data had a markedly smaller effect on the performance gaps between subgroups for the DERM7PT validation.

6. Discussion and conclusions

We investigated the effect of demographic bias on skin lesion classification performance using three ResNet-50-based CNN models, with a specific focus on variations in patient sex and age in the training data. Using linear programming to generate datasets with controlled demographic distributions, we evaluated three learning strategies: a single-task model, a reinforcing multi-task model, and an adversarial learning scheme. Additionally, we performed cross-dataset validation to assess the model's generalisation capabilities. Overall, the results highlight that sex-related performance gaps are largely driven by training set imbalances, whereas age-related declines persist even with balanced sampling, and that domain shifts (from dermoscopic to smartphone images) cause substantial drops in external validation accuracy. Bias patterns were not consistent across all datasets.

Sex based analysis In our sex-based analysis, we observed that sex-specific training data produced better results, though single-task models exhibited notable sex bias. The reinforcement approach was partially effective in reducing this bias. The adversarial model eliminated sex bias specifically in cases involving predominantly female distributions (F95M5). An unexpected result emerged regarding sex-related bias in ISIC-based datasets. When the patient cohort was male-dominated, models achieved higher accuracy on male patients. Conversely, in female-dominated cohorts, models did not show a comparable boost for female patients. This asymmetry suggests that models tend to overfit to male-specific cues. In contrast to ISIC-based datasets, PAD-UFES-20 and DERM7PT showed that the adversarial model performed better on female patients than on male patients in female-dominated cohorts.

The auxiliary sex-prediction head heavily relies on data composition: it learns demographic signals under balanced distributions but struggles in skewed scenarios. This shows that the reinforcing model cannot build a robust discriminator when the minority class is underrepresented, leading to the collapse of the regularisation signal. Our Brier score analysis confirms this mechanism: in balanced datasets, the head makes well-calibrated predictions for both sexes, whereas in skewed datasets, it becomes over-confident for the majority and inaccurate for the minority. Ultimately, effective debiasing depends on the auxiliary head's ability to reliably learn the attribute. When data scarcity prevents this, the reinforcing model overfits to the dominant group, and the regularisation mechanism fails.

Our findings on sex-related bias contradict the conclusion of Sies et al. that *despite sex-related imbalances in open access training data, the diagnostic performance of the CNN tested showed no sex-related bias in the classification of skin lesions* (Sies et al. (2022)). We hypothesise that dataset size contributes to these divergent findings. Whereas Sies et al. leveraged over 150,000 dermoscopic images, our smaller dataset may have rendered the CNN more vulnerable to sex-related biases. Larger datasets typically offer greater diversity across demographic groups, facilitating the learning of robust and generalisable features. Conversely, smaller datasets may amplify existing biases or result in overfitting to specific demographic characteristics.

As expected, the base model shows sensitivity for sex bias, likely driven by overfitting and various anatomical confounders in the training data. These include sex-specific variations in skin thickness, hair distribution, lesion location, and underlying vasculature, as well as differences in sun exposure patterns linked to sex-linked behavioural factors. While the reinforcing and adversarial models incorporate regularisation techniques to mitigate such bias, our experiments revealed limited success: bias correction was observed only in the adversarial model for female-only cohorts. This suggests that complex anatomical confounders continue to substantially influence model predictions in mixed-sex populations, resisting standard regularisation approaches.

Age based analysis In age-related experiments, we found comparable baseline performance across all three model approaches, with an evident decline in performance in older age categories. We observed a clear decline trend across age categories: the age group A_1 consistently achieved the highest performance, while subsequent age groups showed progressively lower scores, regardless of the training data distribution used (YOUNGER, BALANCED or OLDER).

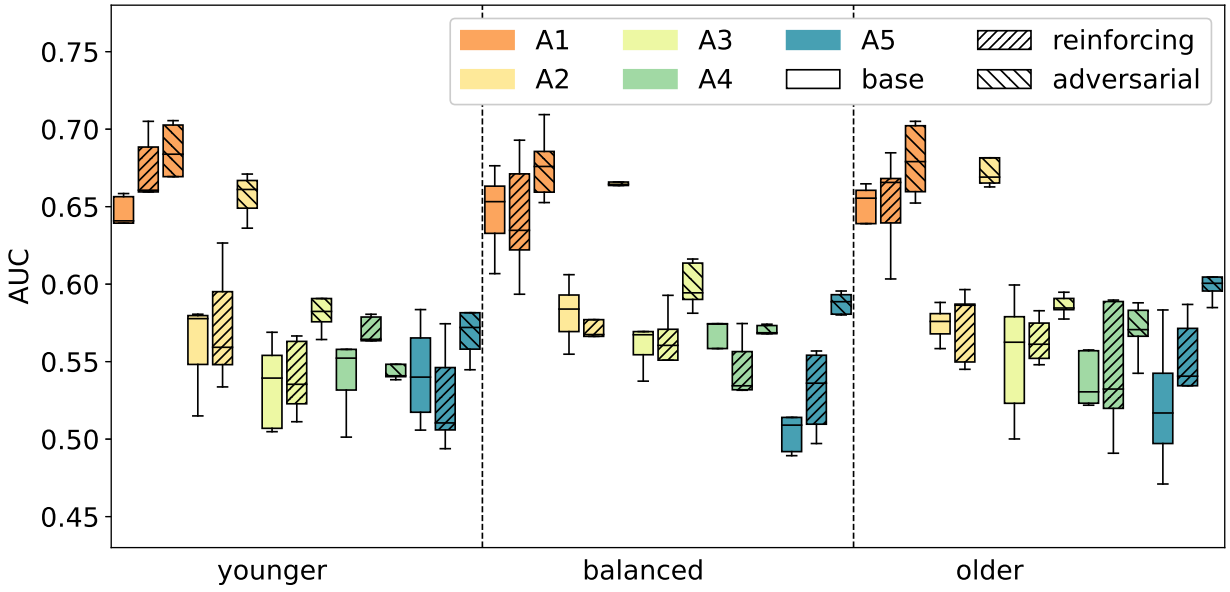


Figure 10: Age-stratified model evaluation showing AUC performance across different age categories (A1-A5) for three model architectures (base, reinforcing, and adversarial) trained and validated on age-biased ISIC datasets (YOUNGER, BALANCED, OLDER) and evaluated on the curated **PAD-UFES-20 dataset**. The analysis demonstrates how each model type performs across different age distributions. The age brackets are defined as follows, where a represents the patient's age in years: $A_1 = 0 \leq a \leq 50$, $A_2 = 51 \leq a \leq 60$, $A_3 = 61 \leq a \leq 70$, $A_4 = 71 \leq a \leq 80$, and $A_5 = a \geq 81$.

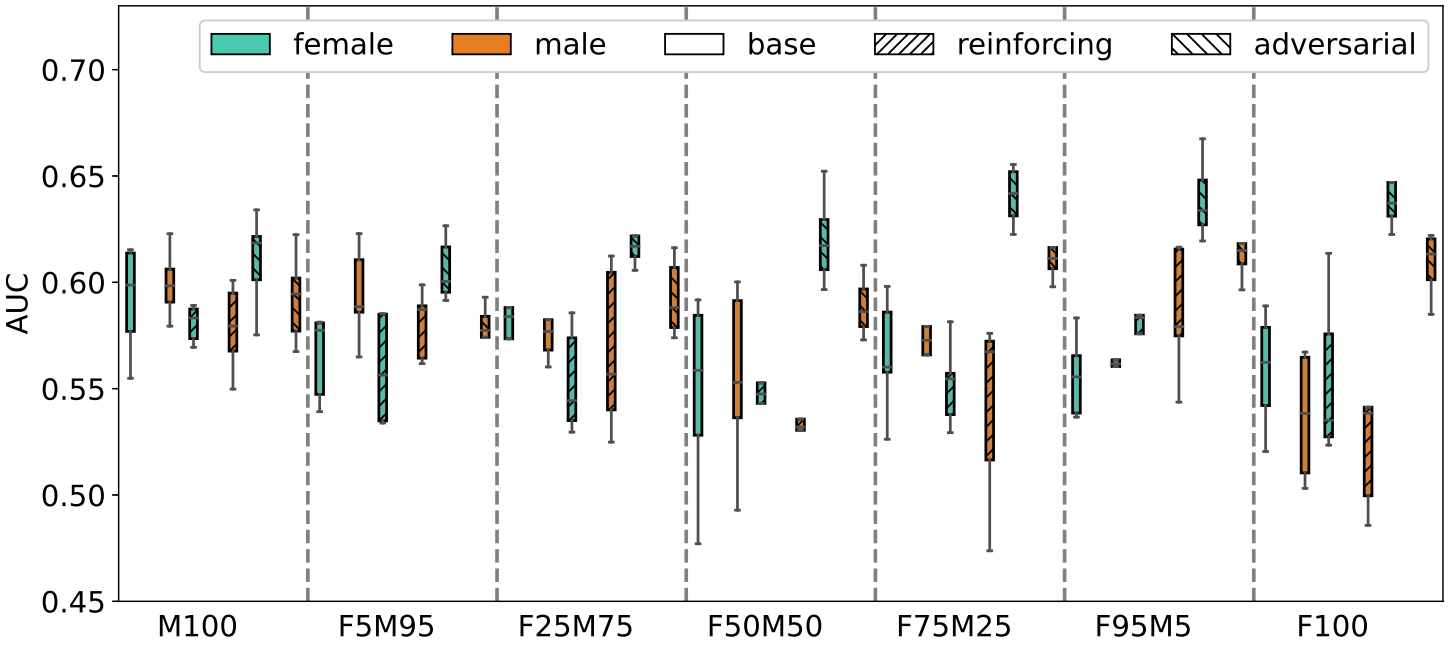


Figure 11: Sex-stratified model evaluation showing AUC performance across male and female categories for three model architectures (base, reinforcing, and adversarial) trained and validated on sex-biased ISIC datasets (M100, F5M95, F25M75, F50M50, F75M25, F95M5 and F100) and evaluated on the curated **PAD-UFES-20 dataset**.

The moderate predictive power of the age head ($0.319 \leq \rho \leq 0.517$) suggests it may not strongly influence bias mitigation. As shown in Figure 8, performance disparities remain evident across all training distributions. While the OLDER dataset shows smaller subgroup gaps than the YOUNGER and BALANCED cases, we cannot definitively

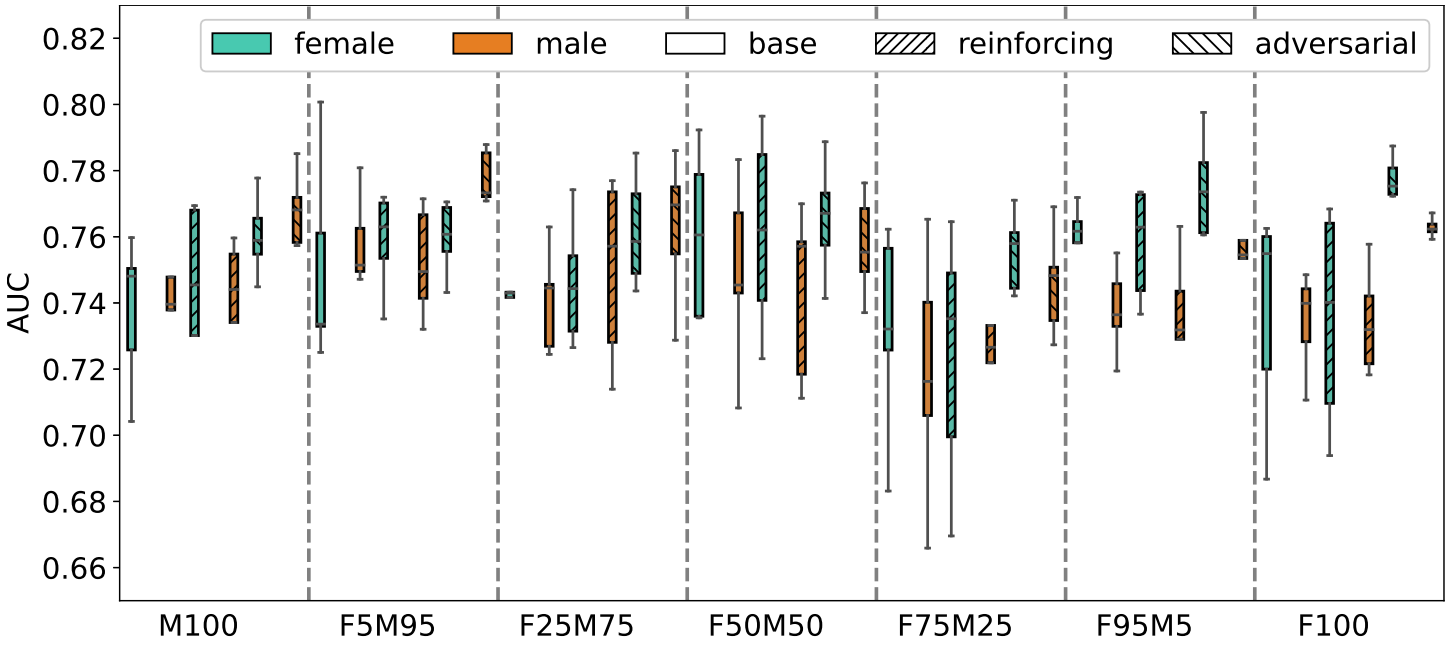


Figure 12: Sex-stratified model evaluation showing AUC performance across male and female categories for three model architectures (base, reinforcing, and adversarial) trained and validated on sex-biased ISIC datasets (M100, F5M95, F25M75, F50M50, F75M25, F95M5 and F100) and evaluated on the curated **DERM7PT dataset**.

attribute this to successful bias mitigation. Instead, it appears to result from the model exploiting age-related shortcuts. This is supported by the strong correlation for the youngest cohort (A_1, $\rho = 0.710$) under older-skewed training, suggesting the model may rely on age-specific cues.

A plausible explanation for the trend of increasing overall correlation from the youngest cohort to the oldest cohort is that the model trained on a cohort dominated by older patients learns a more reliable mapping from image characteristics to chronological age because aging skin shows more significant structural alterations, such as wrinkles, skin texture degradation, increased density of pigmentary spots, and larger pigmentary spots (Flament et al. (2019)).

Based on Pearson correlation and MAE analyses, the auxiliary age-prediction head functions as a demographically sensitive regressor that performs optimally when age distributions are balanced and match the training data. Fairness-focused pipelines should therefore employ balanced age sampling or explicit re-weighting and augmentation to mitigate systematic bias. This is particularly evident in the MAE analysis, which shows substantially higher errors for underrepresented age groups.

Several factors likely contribute to the observed age-related performance differences. Natural changes in skin characteristics with age can influence lesion classification. Even with balanced sampling, intrinsic skin conditions may differ between age groups, potentially affecting model performance.

Within the youngest age group (A_1), malignant cases show a concentration toward the end of the age range, close to the upper boundary of the group. This contrasts with benign lesions, which are well-represented throughout the group, including at the beginning. This internal skew results in an unequal distribution within the category: an overrepresentation of benign samples at the beginning of the group and malignant samples at the end. Such a subgroup distribution can influence model training and complicate the interpretation of the corresponding test results. While we maintained class balance across age distributions through our LP constraints, we recognise that the natural prevalence of malignancy varies by age in clinical practice.

The consistent decline in performance for older age categories initially suggests a need for more balanced age representation in training data. However, since this pattern persists even with balanced training sets, it indicates that factors beyond mere data distribution, such as intrinsic biological or physiological differences, are driving these performance gaps.

The discretisation of continuous variables such as age presents methodological challenges, given that age is inherently continuous. While categorical variables like patient sex have natural groupings, age categorisation necessitates arbitrary cut-off points that may not correspond to biologically or clinically meaningful boundaries.

Adversarial learning is designed to reduce reliance on confounding features by training a discriminator to predict protected attributes, while updating the main model to minimise the discriminator's prediction. However, our results show that this approach succeeded only in certain data com-

positions, failing to mitigate bias in other scenarios. This discrepancy suggests that shortcut learning (the reliance on superficial correlations rather than clinically relevant features) remains a significant concern. For instance, body hair patterns, which dermatologists have noted can significantly interfere with dermatoscopic examinations (Fink et al. (2020)), could be unintentionally used by the model as a proxy for sex or age classification, potentially affecting performance differences between male and female patients. Similarly, confounding factors such as skin colour and image artefacts may influence classification performance across demographic subgroups. Future research should systematically investigate these factors to determine whether adversarial learning can be adapted to address them more robustly across all demographic compositions. Furthermore, we will examine the bias mechanisms by analysing the auxiliary head's behaviour in the adversarial setting.

We demonstrate that skewed distributions in training data cause performance disparities across sex and age groups. Consequently, when using a dataset with a certain level of skewness, one possible mitigation strategy is to rebalance the data through augmentation. However, in many practical scenarios, acquiring additional instances is often infeasible in the short term. Under such constraints, introducing synthetically generated images may provide an alternative way to restore balance (Stanley et al. (2024), Kebaili et al. (2023)). Nonetheless, it remains crucial to understand the root sources of bias, such as demographic representation, acquisition protocols, or annotation practices, so these factors can be explicitly considered during the synthesis process.

Cross-dataset validation and future directions In our cross-dataset validation, we observed performance patterns by demographics: younger age groups generally performed better but showed a notable decline in performance during external validation. In the external validation of models trained on varying female-to-male ratios, we observed improved performance for female patients.

External-validation performance may be affected by numerous factors, including differences in training data, imaging equipment (such as smartphones versus dermoscopes), population demographics, and image-collection protocols. The PAD-UFES-20 set comprises digital camera photographs that differ markedly from the dermoscopic images used for training. Because we evaluated the models without any fine-tuning, we observed a sharp drop in accuracy for the PAD-UFES-20 set. We assume that the lack of adaptation contributed to this decline, although other factors may also play a role. Previous research has provided valuable information in this area. DERM7PT showed lower AUC than internal validation but outperformed PAD-UFES-20, with smaller male/female performance gaps than ISIC-based experiments. This likely reflects modality consistency (dermoscopic-to-dermoscopic transfer reduces domain shift) and potentially less salient sex-related visual cues in DERM7PT images compared to ISIC.

Previous research has provided valuable information in this area. Bevan and Atapour-Abarghouei demonstrated improved generalisation through "unlearning" spurious variations in skin lesion imaging instruments (Bevan and Atapour-Abarghouei (2023)). Daneshjou and colleagues emphasised the need to address skin tone bias in dermatology AI systems before deploying them to diverse populations (Daneshjou et al. (2022)). These studies confirm that performance decline results from differences in image acquisition methods and varying population characteristics between datasets. Additional research is needed to investigate cross-dataset performance across different imaging modalities, such as comparing model robustness between dermoscopic, smartphone, and conventional digital camera images. We recommend that future studies separate modality effects from demographic bias. One way is to assemble paired datasets where the same lesions are photographed with dermoscopes, smartphones, and regular digital cameras. These "multi-view" benchmarks would let us measure performance loss caused by modality changes while keeping lesion identity and patient demographics constant.

Concluding remarks In conclusion, our experiments show that imbalanced training data mainly cause sex-related performance differences in skin-lesion classification. Conversely, age-related performance gaps remain even when the training set is balanced.

Our findings highlight the importance of understanding two key aspects when designing and implementing AI in medical imaging: explicit factors (such as imaging protocols, patient demographic data, data set distributions, and sampling methods) and implicit factors (such as geographic differences, biological variations, and demographic imbalances). Specifically, our research concludes that even with balanced training sets, performance disparities between demographic groups persisted, indicating that the relationship between these factors is complex. Previous studies have shown that bias in medical imaging arises from multiple interconnected factors (Zong et al. (2023)). Disentangling how these factors reinforce or oppose each other, as well as their impact on the performance of medical image applications, remains a challenging area for further research.

We hope that the methodologies and insights developed through our research can serve as building blocks to create more sophisticated and equitable healthcare AI systems that better serve diverse patient populations.

Acknowledgments

This work was supported by the Netherlands Organisation for Scientific Research, grant no. 023.014.010.

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

Conflicts of Interest

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

Data availability

The data used to create the various datasets in this study is publicly available through the ISIC (International Skin Imaging Collaboration) archive, the PAD-UFES-20 dataset, and the DERM7PT dataset. The code necessary for dataset creation and analysis is available via a public GitHub repository. Researchers interested in reproducing or building on this work can access the ISIC archive at `https://www.isic-archive.com`, the PAD-UFES-20 dataset at `https://data.mendeley.com/datasets/zr7vgbcyr2/1`, the DERM7PT dataset at `http://derm.cs.sfu.ca`, and find our code at `https://github.com/raumannsr/demographic-fairness-extended`.

References

Samaneh Abbasi-Sureshjani, Ralf Raumanns, Britt EJ Michels, Gerard Schouten, and Veronika Cheplygina. Risk of training diagnostic algorithms on data with demographic bias. In *MICCAI LABELS workshop, Lecture Notes in Computer Science*, volume 12446, pages 183–192. Springer, 2020.

Ehsan Adeli, Qingyu Zhao, Adolf Pfefferbaum, Edith V Sullivan, Li Fei-Fei, Juan Carlos Niebles, and Kilian M Pohl. Representation learning with statistical independence to mitigate bias. *IEEE Winter Conf Appl Comput Vis*, 2021: 2512–2522, January 2021.

Valentin Amrhein, Sander Greenland, and Blake McShane. Scientists rise up against statistical significance. *Nature*, 567(7748):305–307, March 2019.

Giuseppe Argenziano et al. Interactive atlas of dermoscopy: A tutorial. Book and CD-ROM, 2000.

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM van der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, 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.

Marin Benčević, Marija Habijan, Irena Galić, Danilo Babin, and Aleksandra Pižurica. Understanding skin color bias in deep learning-based skin lesion segmentation. *Comput. Methods Programs Biomed.*, 245:108044, March 2024.

Peter J Bevan and Amir Atapour-Abarghouei. Skin deep unlearning: Artefact and instrument debiasing in the context of melanoma classification. *arXiv preprint arXiv:2109.09818*, April 2023.

Alceu Bissoto, Eduardo Valle, and Sandra Avila. Debiasing skin lesion datasets and models? not so fast, 2020.

Richard A Caruana. Multitask learning: A knowledge-based source of inductive bias. In *Machine Learning Proceedings 1993*, pages 41–48. Elsevier, 1993.

Bill Cassidy, Connah Kendrick, Andrzej Brodzicki, Joanna Jaworek-Korjakowska, and Moi Hoon Yap. Analysis of the ISIC image datasets: Usage, benchmarks and recommendations. *Med. Image Anal.*, 75:102305, January 2022.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic), 2019.

Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern. 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.

Marc Combalia, Noel C. F. Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C. Halpern, Susana Puig, and Josep Malvehy. Bcn20000: Dermoscopic lesions in the wild, 2019.

Roxana Daneshjou, Kailas Vodrahalli, Roberto A Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M Swetter, Elizabeth E Bailey, Olivier Gevaert, Pritam Mukherjee, Michelle Phung, Kiana Yekrang, Bradley Fong, Rachna Sahasrabudhe, Johan A C Allerup, Utako Okata-Karigane, James Zou, and Albert S Chiou. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Sci. Adv.*, 8(32): eabq6147, August 2022.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, February 2017.

Christine Fink, Lorenz Uhlmann, Karsten Vogt, Roland Schneiderbauer, Christian Menzer, Ferdinand Toberer, Timo E Schank, Alexander Enk, and Holger A Haenssle. Physicians' level of hindrance by body hair in dermatoscopy and clinical benefit of an automated hair removal algorithm. *J. Dtsch. Dermatol. Ges.*, 18(1): 27–32, January 2020.

F Flament, D Velleman, S Yamamoto, A Nicolas, K Udodaira, S Yamamoto, C Morimoto, S Belkebla, C Negre, and C Delaunay. Clinical impacts of sun exposures on the faces and hands of japanese women of different ages. *Int. J. Cosmet. Sci.*, 41(5):425–436, October 2019.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, November 2020.

Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. "O'Reilly Media, Inc.", October 2022.

Nele Gerrits, Bart Elen, Toon Van Craenendonck, Danai Triantafyllidou, Ioannis N Petropoulos, Rayaz A Malik, and Patrick De Boever. Publisher correction: Age and sex affect deep learning prediction of cardiometabolic risk factors from retinal images. *Sci. Rep.*, 11(1):1198, January 2021.

Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, Po-Chih Kuo, Matthew P Lungren, Lyle J Palmer, Brandon J Price, Saptarshi Purkayastha, Ayis T Pyrros, Lauren Oakden-Rayner, Chima Okechukwu, Laleh Seyyed-Kalantari, Hari Trivedi, Ryan Wang, Zachary Zaiman, and Haoran Zhang. AI recognition of patient race in medical imaging: a modelling study. *Lancet Digit. Health*, 4(6):e406–e414, June 2022a.

Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, et al. AI recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, 4(6):e406–e414, 2022b.

Ben Glocker, Charles Jones, Mélanie Roschewitz, and Stefan Winzeck. Risk of bias in chest radiography deep learning foundation models. *Radiol. Artif. Intell.*, 5(6):e230060, November 2023.

Matthew Groh, Caleb Harris, L Soenksen, Felix Lau, Rachel Han, Aerin Kim, A Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1820–1828, April 2021.

Matthew Groh, Caleb Harris, Roxana Daneshjou, Omar Badri, and Arash Koochek. 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, November 2022.

David Gutman, Noel C. F. Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. 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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

ISIC2024. ISIC archive. `https://gallery.isic-archive.com`. Accessed: 2024-06-07.

Amelia Jiménez-Sánchez, Dovile Juodelyte, Bethany Chamberlain, and Veronika Cheplygina. Detecting shortcuts in medical images-a case study in chest x-rays. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023.

Charles Jones and Ben Glocker. A primer on causal and statistical dataset biases for fair and robust image analysis. *arXiv [cs.LG]*, 2025.

Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. 7-point checklist and skin lesion classification using multi-task multi-modal neural nets. *IEEE J. Biomed. Health Inform.*, 23(2):538–546, April 2018.

Aghiles Kebaili, Jérôme Lapuyade-Lahorgue, and Su Ruan. Deep learning approaches for data augmentation in medical imaging: A review. *J. Imaging*, 9(4):81, April 2023.

H Kittler, H Pehamberger, K Wolff, and M Binder. Diagnostic accuracy of dermoscopy. *Lancet Oncol.*, 3(3): 159–165, March 2002.

Malte Klingenberg, Didem Stark, Fabian Eitel, Céline Budding, Mohamad Habes, Kerstin Ritter, and Alzheimer's Disease Neuroimaging Initiative. Higher performance for women than men in MRI-based alzheimer's disease detection. *Alzheimers. Res. Ther.*, 15(1):84, April 2023.

Ira Ktena, Olivia Wiles, Isabela Albuquerque, Sylvestre-Alvise Rebuffi, Ryutaro Tanno, Abhijit Guha Roy, Shekoofeh Azizi, Danielle Belgrave, Pushmeet Kohli, Taylan Cemgil, Alan Karthikesalingam, and Sven Gowal. Generative models improve fairness of medical classifiers under distribution shifts. *Nat. Med.*, 30(4):1166–1173, April 2024.

Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. 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.

Mingxia Liu, Jun Zhang, Ehsan Adeli, and Dinggang Shen. Joint classification and regression via deep multi-task multi-channel learning for alzheimer's disease diagnosis. *IEEE Trans. Biomed. Eng.*, 66(5):1195–1206, May 2019.

Meike Nauta, Ricky Walsh, Adam Dubowski, and Christin Seifert. Uncovering and correcting shortcut learning in machine learning models for skin cancer diagnosis. *Diagnostics (Basel)*, 12(1), December 2021.

Andre GC Pacheco, Gustavo R Lima, Amanda S Salomao, Breno Krohling, Igor P Biral, Gabriel G de Angelo, Fábio CR Alves Jr, José GM Esgario, Alana C Simora, Pedro BC Castro, et al. Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in brief*, 32:106221, 2020.

A Paszke, S Gross, F Massa, A Lerer, J P Bradbury, G Chanan, T Killeen, Z Lin, N Gimelshein, L Antiga, and Others. An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.*, 32:8026–8037, 2019.

Eike Petersen, Aasa Feragen, Maria Luise da Costa Zemsch, Anders Henriksen, Oskar Eiler Wiese Christensen, Melanie Ganz, and Alzheimer's Disease Neuroimaging Initiative. 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*, pages 88–98. Springer, 2022.

Ralf Raumanns, Gerard Schouten, Josien P. W. Pluim, and Veronika Cheplygina. Dataset distribution impacts model fairness: Single vs. multi-task learning. In Esther Puyol-Antón, Ghada Zamzmi, Aasa Feragen, Andrew P. King, Veronika Cheplygina, Melanie Ganz-Benjaminsen, Enzo Ferrante, Ben Glocker, Eike Petersen, John S. H. Baxter, Islem Rekik, and Roy Eagleson, editors, *Ethics and Fairness in Medical Imaging*, pages 14–23, Cham, 2025. Springer Nature Switzerland.

V. Rotemberg, N. Kurtansky, B. Betz-Stablein, L. Caffery, E. Chousakos, N. Codella, M. Combalia, S. Dusza, P. Guitera, D. Gutman, A. Halpern, B. Helba, H. Kittler, K. Kose, S. Langer, K. Lioprys, J. Malvehy, S. Musthaq, J. Nanda, O. Reiter, G. Shih, A. Stratigos, P. Tschandl, J. Weber, and H.P. Soyer. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific Data; London*, 8(1):s41597–021, 2021.

Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

Kaspar Rufibach. Use of brier score to assess binary predictions. *J. Clin. Epidemiol.*, 63(8):938–9; author reply 939, August 2010.

A. Saha, J.S. Bosma, J.J. Twilt, B. van Ginneken, A. Bjartell, A.R. Padhani, D. Bonekamp, G. Villeirs, G. Salomon, G. Giannarini, J. Kalpathy-Cramer, J. Barentsz, K.H. Maier-Hein, M. Rusu, O. Rouvière, R. van den Bergh, V. Panebianco, V. Kasivisvanathan, N.A. Obuchowski, D. Yakar, et al. Artificial intelligence and radiologists in prostate cancer detection on mri (pi-cai): an international, paired, non-inferiority, confirmatory study. *Lancet Oncol.*, June 2024.

Pratinav Seth and Abhilash K Pai. 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*, pages 570–577, 2024.

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B A McDermott, Irene Y Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat. Med.*, 27(12):2176–2182, December 2021.

Katharina Sies, Julia K Winkler, Christine Fink, Felicitas Bardehle, Ferdinand Toberer, Timo Buhl, Alexander Enk, Andreas Blum, Wilhelm Stolz, Albert Rosenberger, and Holger A Haenssle. Does sex matter? analysis of sex-related differences in the diagnostic performance of a market-approved convolutional neural network for skin cancer detection. *Eur. J. Cancer*, 164:88–94, March 2022.

Emma A M Stanley, Raissa Souza, Anthony J Winder, Vedant Gulve, Kimberly Amador, Matthias Wilms, and Nils D Forkert. Towards objective and systematic evaluation of bias in artificial intelligence for medical imaging. *J. Am. Med. Inform. Assoc.*, 31(11):2613–2621, November 2024.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, 2018.

Anurag Vaidya, Richard J Chen, Drew F K Williamson, Andrew H Song, Guillaume Jaume, Yuzhe Yang, Thomas Hartvigsen, Emma C Dyer, Ming Y Lu, Jana Lipkova, Muhammad Shaban, Tiffany Y Chen, and Faisal Mahmood. Demographic bias in misdiagnosis by computational pathology models. *Nat. Med.*, 30(4):1174–1190, April 2024.

Martin J Willemink, Wojciech A Koszek, Cailin Hardell, Jie Wu, Dominik Fleischmann, Hugh Harvey, Les R Folio, Ronald M Summers, Daniel L Rubin, and Matthew P Lungren. Preparing medical imaging data for machine learning. *Radiology*, page 192224, 2020.

Yawen Wu, Dewen Zeng, Xiaowei Xu, Yiyu Shi, and Jingtong Hu. FairPrune: Achieving fairness through pruning for dermatological disease diagnosis. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 743–753. Springer Nature Switzerland, 2022.

Wanni Xu, You-Lei Fu, and Dongmei Zhu. ResNet and its application to medical image processing: Research progress and challenges. *Comput. Methods Programs Biomed.*, 240(107660):107660, October 2023.

Jenny Yang, Andrew A S Soltan, David W Eyre, Yang Yang, and David A Clifton. An adversarial training framework for mitigating algorithmic biases in clinical machine learning. *NPJ Digit Med*, 6(1):55, March 2023.

Paul H Yi, Jinchi Wei, Tae Kyung Kim, Jiwon Shin, Haris I Sair, Ferdinand K Hui, Gregory D Hager, and Cheng Ting Lin. Radiology "forensics": determination of age and sex from chest radiographs using deep learning. *Emerg. Radiol.*, 28(5):949–954, October 2021.

Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Trans. Knowl. Data Eng.*, 34(12):5586–5609, December 2022.

Yongshuo Zong, Yongxin Yang, and Timothy Hospedales. MEDFAIR: Benchmarking fairness for medical imaging. *arXiv [cs.LG]*, February 2023.

Appendix A. Sex distribution LP model

A.1 Decision variables

x_1: # malignant instances
x_2: # benign instances
x_3: # male patients (M) with malignant lesions
x_4: # female patients (F) with malignant lesions
x_5: # benign M
x_6: # benign F
x_7: # malignant lesions of M (age < 60)
x_8: # malignant lesions of M (age ≥ 60)
x_9: # malignant lesions of F (age < 60)
x_{10}: # malignant lesions of F (age ≥ 60)
x_{11}: # benign M (age < 60)
x_{12}: # benign M (age ≥ 60)
x_{13}: # benign F (age < 60)
x_{14}: # benign F (age ≥ 60)

A.2 Constraints

$$x_1 - x_2 = 0 \quad (1)$$
$$rx_4 - x_3 = 0 \quad (2)$$
$$sx_8 - x_7 = 0 \quad (3)$$
$$tx_{10} - x_9 = 0 \quad (4)$$
$$x_1 - x_3 - x_4 = 0 \quad (5)$$
$$x_3 - x_7 - x_8 = 0 \quad (6)$$
$$x_4 - x_9 - x_{10} = 0 \quad (7)$$
$$ux_{12} - x_{11} = 0 \quad (8)$$
$$vx_{14} - x_{13} = 0 \quad (9)$$
$$x_2 - x_5 - x_6 = 0 \quad (10)$$
$$x_6 - x_{13} - x_{14} = 0 \quad (11)$$
$$x_5 - x_{11} - x_{12} = 0 \quad (12)$$
$$wx_6 - x_5 = 0 \quad (13)$$

Eq. 1: # of malignant records equals # benign records.

Eq. 2: Ratio r of malignant male (M) to female patients (F).

Eq. 3: Ratio s of malignant M (age < 60) to M (age ≥ 60).

Eq. 4: Ratio t of malignant F (age < 60) to F (age ≥ 60).

Eq. 5: # of malignant lesions equals sum of malignant M and F.

Eq. 6: # M with malignant lesions is equal to that of all ages.

Eq. 7: # F with malignant lesions is equal to that of all ages.

Eq. 8: Ratio u of benign M (age < 60) to M (age ≥ 60).

Eq. 9: Ratio v of benign F (age < 60) to F (age ≥ 60).

Eq. 10: # benign lesions equals sum of benign M and F.

Eq. 11: # F with benign lesions is equal to # all ages.

Eq. 12: # M with benign lesions is equal to # all ages.

Eq. 13: Ratio w of benign M to F.

Appendix B. Age distribution LP model

B.1 Decision variables

The age brackets are defined as follows, where a represents the patient's age in years.
x_1:# $0 \leq a \leq 50 (A_1)$ instances
x_2:# $51 \leq a \leq 60 (A_2)$ instances
x_3:# $61 \leq a \leq 70 (A_3)$ instances
x_4:# $71 \leq a \leq 80 (A_4)$ instances
x_5:# $a \geq 81 (A_5)$ instances

B.2 Objective function

Find a vector x (decision variables)
that maximises $f = x_1$ (objective function)
subject to $a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{i5}x_5 \leq b_i$ (constraints)
for $i = 1, \ldots, 5$
and $x_j \geq 0$ (non-negativity constraints)
for $j = 1, \ldots, 5$

B.3 Constraints

$$(100 - p_1)x_1 - p_1x_2 - p_1x_3 - p_1x_4 - p_1x_5 = 0 \quad (1)$$
$$-p_2x_1 + (100 - p_2)x_2 - p_2x_3 - p_2x_4 - p_2x_5 = 0 \quad (2)$$
$$-p_3x_1 - p_3x_2 + (100 - p_3)x_3 - p_3x_4 - p_3x_5 = 0 \quad (3)$$
$$-p_4x_1 - p_4x_2 - p_4x_3 + (100 - p_4)x_4 - p_4x_5 = 0 \quad (4)$$
$$-p_5x_1 - p_5x_2 - p_5x_3 - p_5x_4 + (100 - p_5)x_5 = 0 \quad (5)$$

Eq. 1 : Distribution for first category with percentage p_1.

Eq. 2 : Distribution for second category with percentage p_2.

Eq. 3 : Distribution for third category with percentage p_3.

Eq. 4 : Distribution for fourth category with percentage p_4.

Eq. 5 : Distribution for fifth category with percentage p_5.

Appendix C. PAD-UFES-20 demographics

Table 8: Age distribution ($A_1 - A_5$) of skin lesions across sex and diagnosis in the curated **PAD-UFES-20** dataset, showing the breakdown between male and female patients. The age brackets are defined as follows, where a represents the patient's age in years: $A_1 = 0 \leq a \leq 50$, $A_2 = 51 \leq a \leq 60$, $A_3 = 61 \leq a \leq 70$, $A_4 = 71 \leq a \leq 80$, and $A_5 = a \geq 81$.

Category	Female benign	Female malignant	Male benign	Male malignant
A_1	34	52	28	56
A_2	52	110	36	98
A_3	31	88	45	119
A_4	59	88	22	114
A_5	22	63	17	45