MELBA Journal Machine Learning for Biomedical Imaging

Understanding-informed Bias Mitigation for Fair CMR Segmentation

Tiarna **Lee** [5], Esther **Puyol-Anton** [5, 3] Bram **Ruijsink** [5, 2] Pier-Giorgio **Masci** [5] Louise **Keehn** [4] Phil **Chowienczyk** [4] Emily **Haseler** [4] Miaojing **Shi** [1] Andrew **King** [5]

1 College of Electronic and Information Engineering, Tongji University
2 Guy's and St Thomas' NHS Foundation Trust
3 HeartFlow Inc, United States
4 British Heart Foundation Centre, Department of Clinical Pharmacology, King's College London
5 School of Biomedical Engineering & Imaging Sciences, King's College London

Abstract

Keywords
Machine Learning, Bias Mitigation

Article informations
https://doi.org/https://doi.org/10.59275/j.melba.2025-6747
Volume 3, Received: 2025-04-21, Published 2025-12-30
Corresponding author: tiarna.lee@kcl.ac.uk

Artificial intelligence (AI) is increasingly being used for medical imaging tasks. However, there can be biases in AI models, particularly when they are trained using imbalanced training datasets. One such example has been the strong ethnicity bias effect in cardiac magnetic resonance (CMR) image segmentation models. Although this phenomenon has been reported in a number of publications, little is known about the effectiveness of bias mitigation algorithms in this domain. We aim to investigate the impact of common bias mitigation methods to address bias between Black and White subjects in AI-based CMR segmentation models. Specifically, we use oversampling, importance reweighing and Group DRO as well as combinations of these techniques to mitigate the ethnicity bias. Second, motivated by recent findings on the root causes of AI-based CMR segmentation bias, we evaluate the same methods using models trained and evaluated on cropped CMR images. We find that bias can be mitigated using oversampling, significantly improving performance for the underrepresented Black subjects whilst not significantly reducing the majority White subjects' performance. Using cropped images increases performance for both ethnicities and reduces the bias, whilst adding oversampling as a bias mitigation technique with cropped images reduces the bias further. When testing the models on an external clinical validation set, we find high segmentation performance and no statistically significant bias.

1. Introduction

Artificial intelligence (AI) is increasingly being used to aid medical diagnosis, prognosis and treatment planning. However, AI models have been shown to exhibit bias by protected attributes in many different applications (Larrazabal et al., 2020; Seyyed-Kalantari et al., 2021a; Klingenberg et al., 2023), including AI-based segmentation of cardiac magnetic resonance (CMR) images (Puyol-Antón et al., 2021, 2022; Lee et al., 2022, 2023). AI bias can have detrimental downstream impacts in medical imaging applications. For example, CMR segmentations are used to derive biomarkers whose values impact patient management, so greater errors in these biomarkers for certain protected groups can lead to inappropriate treatment choices and worse outcomes (Puyol-Antón et al., 2022).

Previous work has aimed to address biases in AI models for medical imaging tasks by using generic bias mitigation methods. For example, Zong et al. (2022) proposed a framework with eleven algorithms which aimed to measure and mitigate biases in medical imaging classification datasets. However, the fairness gains achieved by such methods often cause reduced performance for some protected groups, a phenomenon known as the 'fairness-accuracy trade-off' (Li and Li, 2024). Furthermore, research has suggested that, whilst such generic bias mitigation approaches may

appear to reduce bias when evaluated on internal validation sets, the fairness gains often do not hold when evaluated externally (Schrouff et al., 2022; Yang et al., 2024). We hypothesise that a possible reason for this lack of effectiveness is that these methods take a "blind" approach to addressing bias in AI models, i.e. they apply a generic bias mitigation algorithm that does not take account of the underlying cause of the bias. More recently, a new strand of research has emerged which attempts to understand the bias, with a view to using this knowledge to develop more informed mitigation approaches. For example, Glocker et al. (2023) analysed bias in chest X-ray classifiers, applying test-set resampling, multitask learning, and model inspection to provide insights into the way protected attributes were encoded in the AI model. Similarly, Olesen et al. (2025) used 'slice discovery methods' to reveal the root causes of sex bias, also in chest X-ray classifiers. Such methods potentially enable the development of bias mitigation techniques that are targeted at the underlying cause(s) of the bias, and could lead to improved and more robust mitigation, which will hold under the effects of domain shift to external validation sets.

Recently, an investigation was performed to discover the root cause(s) of AI-based CMR segmentation bias (Lee et al., 2025a). One key finding of this work was that distributional differences in areas outside the heart were significant in the learning of biased representations in AI segmentation models. This finding opens up new avenues of enquiry in bias mitigation, which we explore in this paper.

2. Related Works

2.1 Bias in medical imaging

AI bias has been observed in a wide range of medical imaging modalities and tasks. For example, Seyyed-Kalantari et al. (2021a) found bias in chest X-ray classification models in terms of sex and ethnicity. A CNN-based model showed a favourable bias towards males and older people, with these groups having higher true positive rates. Similarly, Seyyed-Kalantari et al. (2021b) found that younger people, females, patients under 20 years old, Black patients and Hispanic patients had higher rates of under-diagnosis. Larrazabal et al. (2020) also used chest X-ray data, and investigated the effect of training a thoracic disease classification model on datasets that were imbalanced by sex. They found a relationship between training set representation and performance for both sexes, and that training a model with more balanced data did not significantly decrease accuracy for the majority group but significantly improved performance for the under-represented group.

Bias in brain magnetic resonance imaging (MRI) classification has also been reported. A key early work here was Petersen et al. (2022), who reported bias in a CNN-based model for Alzheimer's disease (AD) detection. Similarly, Klingenberg et al. (2023) found higher accuracy in MRI-based AD detection in females than males, despite balancing the training dataset by age and sex. Wang et al. (2023) also assessed bias in brain MRI disease classification, investigating the impact of training choices on bias.

In dermatology imaging , Abbasi-Sureshjani et al. (2020) found bias by age and sex when training lesion classification models. Daneshjou et al. (2022) found disparities in accuracy between lighter and darker skin tones, with the model achieving higher accuracy on the lighter skin tones. Fine-tuning the model on a diverse dataset closed this gap in performance and improved overall model performance so that it was equal to, or exceeded, clinician accuracy.

Puyol-Antón et al. (2021) was the first paper to report bias in AI-based segmentation models. This work found that an nnU-Net model trained on ethnicity-imbalanced CMR data produced biased results by ethnicity. The segmentation performance favoured White subjects, who were in the majority in the training set. Subsequent work has investigated the downstream clinical impact of this bias (Puyol-Antón et al., 2022) and analysed it in more controlled experiments by both ethnicity and sex (Lee et al., 2022, 2023). Segmentation bias has since also been reported in brain MRI (Ioannou et al., 2022), dermatology images (Benčević et al., 2024) and orthopaedic radiographs (Siddiqui et al., 2024).

2.2 Bias mitigation methods

Bias mitigation methods aim to reduce the bias observed between protected groups in an AI model. They can be applied as pre-, in- or post-processing methods (Mehrabi et al., 2019). Pre-processing methods aim to transform the data before training to remove bias. These methods include targeted data collection and importance reweighing (which can also be performed as an in-processing method). Efforts to collect data from more diverse populations include a dataset of dermatology data from four African countries (Gottfrois et al., 2024), a leprosy skin imaging dataset from Brazil (Barbieri et al., 2022), brain MRI images from women with fibromyalgia in Mexico (Balducci et al., 2022) and brain tumour segmentation data from Nigeria (Adewole et al., 2024).

In-processing methods can be used to change the objective function or apply constraints to the model during training to reduce bias. An example of such an approach is Group Distributionally Robust Optimisation, or Group DRO (Sagawa et al., 2020), which alters the loss function to optimise performance for the worst performing group in a dataset. Group DRO was one of the approaches evaluated for mitigating bias in chest X-ray classifiers in the

comparative analysis by Zhang et al. (2022). Another example of a modified loss function is Pareto minimax optimisation (Martinez et al., 2020), which aims to minimise an importance-weighted maximum 'risk' across protected groups. Adversarial learning has also been suggested as an in-processing method for bias mitigation. In Madras et al. (2018), a model was adversarially trained to learn fair representations using an encoder-decoder network. Similarly, Zhang et al. (2018) used adversarial debiasing with the added constraint of satisfying fairness definitions such as demographic parity, equalised odds or equalised opportunity.

Post-processing methods aim to modify the predictions made by the model based on a subject's protected attributes. Lohia et al. (2019) proposed an algorithm which assesses an individual sample's prediction, establishes whether it is biased and changes the prediction to the privileged group's label if the sample experiences bias. Reject option classification, proposed in Kamiran et al. (2012), considers the confidence of predictions. For a binary classifier, prediction probabilities close to 0 or 1 represent confident predictions, whereas probabilities close to 0.5 represent more uncertain predictions. Samples are not assigned a label if their probabilities lie within a certain uncertainty range as they are considered more prone to bias and are relabelled depending on the group they belong to.

Bias mitigation in AI-based segmentation has received relatively little attention compared to classification tasks. Relevant works here include FairSeg (Tian et al., 2023), which published a fairness dataset for medical image segmentation and proposed a fair error-bound scaling approach to reweight the loss function. Siddiqui et al. (2024) evaluated stratified batch sampling, a balanced dataset model and a protected group-specific model for orthopaedic image segmentation. Finally, in CMR segmentation, Puyol-Antón et al. (2021) internally evaluated protected group-specific models, oversampling of minority protected groups and a multi-task learning approach which learnt both a segmentation model and a protected attribute classifier.

3. Contributions

In the context of AI-based segmentation of cine CMR, the main contributions of this novel work are:

1. We perform the most extensive and comprehensive investigation to date of multiple bias mitigation methods in CMR segmentation, as well as combinations of these methods.

2. We perform one of the first investigations into using knowledge of the root cause of bias for mitigation. Specifically, we train AI segmentation models using CMR images that are cropped to remove features outside of the heart. We evaluate a baseline model and state-of-the-art bias mitigation techniques in this setting.

3. We evaluate the efficacy of all approaches under both internal and external validation settings.

A preliminary version of this work has been published in Lee et al. (2025b). This paper extends that work by (i) incorporating a wider range of metrics and more in-depth analysis and discussion, (ii) inclusion of a new clinically-applicable 'cascaded' cropping based mitigation approach and (iii) including external validation of all approaches on a clinical dataset.

4. Materials and Methods

4.1 Data

To train and internally validate all models we used CMR images from the UK Biobank (Petersen et al., 2016). The dataset consists of end diastolic (ED) and end systolic (ES) cine short-axis images from 5,778 subjects. Manual segmentation of the left ventricular blood pool (LVBP), left ventricular myocardium (LVM), and right ventricular blood pool (RVBP) was performed for the ED and ES images of each subject. The LV endocardial and epicardial borders and the RV endocardial border were outlined using cvi42 (version 5.1.1, Circle Cardiovascular Imaging Inc., Calgary, Alberta, Canada). The same guidelines were provided to a panel of ten experts with one expert annotating each image. Each expert was provided with a random selection of images for annotation which included subjects of different sexes and ethnicities. They were not provided with demographic information about the subjects.

Previous work (Lee et al., 2022, 2023) has shown that bias is greater when the imbalance between ethnicities in the training set is greater. Therefore, we curated a dataset where biases would be significant to allow for better evaluation of mitigation methods. Our main training set comprises 15 Black subjects and 4,221 White subjects, all randomly sampled from the full dataset. The remaining subjects were used as the internal validation test set. We also investigate the effect of using different proportions of Black and White subjects in the training set in the Supplementary Material. The demographic information of the subjects in the training and test sets can be seen in Table 1 and Table 2. Note that training was performed using only Black and White subjects but testing was performed using all other available subjects including Mixed, Asian, Chinese and Other. All ethnicity information refers to self-reported ethnicity based on the categories provided in the UK Biobank dataset.

In addition, for external validation a dataset of cine short-axis CMR images from St. Thomas' Hospital, London was used. All subjects were scanned using a 1.5T MRI

Health measure	Overall	White	Black
n subjects	4236	4221	15
Age (years)	64.6 (7.7)	64.6 (7.7)	57.0 (4.9)*
Weight (kg)	77.0 (15.0)	77.0 (15.0)	74.8 (11.9)*
Height (cm)	169.5 (9.2)	169.5 (9.2)	168.8 (6.9)*
Body Mass Index	26.7 (4.3)	26.7 (4.3)	26.2 (3.7)

Table 1: Characteristics of subjects used in the training dataset. Mean (standard deviation) values are presented for each characteristic. Statistically significant differences between subject groups and the overall average are indicated with an asterisk * ($p < 0.05$) and were determined using a two-tailed Student's t-test.

scanner (Aera-Magneton, Siemens Healthcare, Erlangen, Germany) between November 2019 and November 2021. The patients scanned had suspected hypertension. Ethnicity was recorded as Black if both parents self-identified as African descendants or White if both parents self-identified as European descendants. The demographic information for the subjects used can be seen in Table 3. Ground truth segmentations were performed manually by clinical experts using cvi42 (Circle Cardiovascular Imaging Inc., Calgary, Alberta, Canada). Further details of the imaging protocol can be found in Georgiopoulos et al. (2024).

Health measure	Overall	White	Black
n subjects	84	30	54
Age (years)	46.0 (12.8)	40.5 (14.1)	49.1 (11.0)
Weight (kg)	94.6 (19.7)	92.6 (19.3)	95.1 (20.2)
Height (cm)	174.3 (8.96)	176.9 (7.33)	172.9 (9.48)
Body Mass Index	31.2 (6.23)	29.6 (6.13)	31.8 (6.23)

Table 3: Characteristics of subjects used in the external validation test set. Mean (standard deviation) values are presented for each characteristic. Statistically significant differences between subject groups and the overall average are indicated with an asterisk * ($p < 0.05$) and were determined using a two-tailed Student's t-test.

4.2 Baseline Model

As a baseline model, we trained a 2D nnU-Net v1 model (Isensee et al., 2020) using the UK Biobank training set to segment the LVM, LVBP and RVBP.

4.3 Oversampling

We also trained a 2D nnU-Net model using the same data as the baseline but applied oversampling during batch selection (Kamiran and Calders, 2012). Oversampling refers to the process of increasing the sampling of a minority group in a dataset. Here, we oversample the Black subjects in the training set so that they were equal to the number of White subjects in each batch used during training. This was performed using random sampling with replacement so each subject could in principle be selected more than once in a training batch.

4.4 Reweighing

An nnU-Net model was also trained using a reweighing mitigation strategy (Kamiran and Calders, 2012). Reweighing refers to the process of increasing the importance of underrepresented groups to the model. We implemented this strategy by adding a weighting term to the combined Cross Entropy (CE)-Dice loss function of the nnU-Net. Each group was weighted inversely proportionally to the group size, as shown in Eq. (1). The weights were then normalised so that they summed to 1, as shown in Eq. (2).

$$\text{Weights per group:} \quad w_g = \frac{n_G}{n_g + \epsilon}, \quad \text{where } \epsilon = 10^{-6}. \tag{1}$$

$$\text{Normalized weights:} \quad \hat{w}_g = \frac{w_g}{\sum_{j=1}^{N_G} w_j}, \quad g = 1, 2, \ldots, N_G. \tag{2}$$

where n_g is the number of samples in protected group g, n_G is the number of samples in all groups and N_G is the number of groups.

4.5 Group Distributionally Robust Optimisation

The final mitigation approach was Group DRO, which was first proposed in Sagawa et al. (2020). The method aims to optimise the performance of the worst-performing group in a dataset. The Group DRO loss function can be formalised as:

$$L_{DRO} = max_{g \in G} \frac{1}{n_g} \sum\nolimits_{i \in g} L\left(y_i, \hat{y}_i\right) \tag{2}$$

where L is the loss function computed between predicted labels $\hat{y}_i$ and ground truth labels y_i.

In this method, CE loss was used instead of CE-Dice loss which was used for the oversampling and reweighing experiments. The reason for this is that Group DRO uses losses from individual samples to calculate the average loss for the groups. However, Dice loss is calculated using global statistics of the true positives, false positives and false negatives for a group or batch. It is non-additive as the numerator and denominator will change if the calculation is performed on a per-sample basis rather than for a group or batch. For Dice loss, the average group loss and global loss are different, which causes instability in training.

4.6 Training Using Combinations of Mitigation Methods

We also combined the mitigation methods into pairs to test whether combinations of methods would improve performance. This results in three additional methods: oversampling + Group DRO, reweighing + Group DRO, and

Health measure	Overall	White	Mixed	Asian	Black	Chinese	Other
n subjects	1542	469	170	387	223	111	182
Age (years) *	61.7 (7.9)	64.8 (7.7)	59.8 (7.2)	61.0 (8.2)	59.1 (7.1)	59.7 (6.4)	62.1 (7.5)
Height (cm) *	167.4 (9.2)	169.8 (9.4)	166.4 (8.5)	166.7 (8.6)	168.7 (9.4)	162.4 (7.1)	165.4 (9.7)
Weight (kg) *	75.1 (15.1)	77.7 (15.1)	75.1 (15.3)	72.3 (12.8)	82.0 (16.0)	63.5 (9.9)	73.3 (15.5)
Body Mass Index *	26.7 (4.4)	26.9 (4.2)	27.1 (5.2)	25.9 (3.7)	28.8 (5.1)	24.0 (3.1)	26.6 (4.4)

Table 2: Characteristics of subjects used in the internal validation test dataset. Mean (standard deviation) values are presented for each characteristic. One-way ANOVA was used to test for differences across ethnic groups. Statistically significant differences between groups are indicated with an asterisk. * ($p < 0.05$).

oversampling + reweighing. These methods were applied in the same way as above. CE loss was used for experiments with Group DRO and CE-Dice loss was used for oversampling + reweighing.

4.7 Training Using Cropped Images

Following the findings of Lee et al. (2025a), which found that areas outside the heart were a contributing factor to CMR segmentation performance bias, we also performed experiments using all of the above techniques for a nnU-Net model trained using cropped CMR images. The images were cropped around the heart using a bounding box defined based on a segmentation mask. All images were cropped to the same size, i.e. the size of the largest heart in the training set plus a buffer of 5 pixels in both the x and y directions.

When using cropped images, we evaluated two different approaches. First, we used the ground truth segmentations to calculate the cropping region at both training and inference time. Note that such a technique could not be used when the model is deployed since ground truth segmentations would not be available at inference time. Therefore, we evaluate this method to establish an upper bound on the performance of the cropping-based mitigation approach.

Second, we developed a 'cascaded' cropping-based approach, in which the cropping region was estimated using an initial nnU-Net-based segmentation. This first nnU-Net model was trained using the full images and its output was used to estimate the cropping region. The resulting cropped image was then used as input to a second nnU-Net model which was trained using cropped images. The croppings used on the training data in this second nnU-Net were based the ground truth segmentations, which are available at training time. This approach is illustrated in Fig. 1.

4.8 Evaluation

For each of the methods described above, performance was measured by finding the overall Dice similarity coefficient (DSC) and Hausdorff distance (HD) for subjects in the internal and external validation test sets. Performance was quantified using the median and inter-quartile range (IQR) of the DSC/HD. Unless otherwise stated, we report the values of these metrics averaged over the LVBP, LVM and RVBP.

Fairness metrics are also reported. The fairness gap (FG), represents the difference in median DSC between the protected groups (i.e. ethnicities) and is given by $FG = D_{White} - D_{Black}$ where D is the median DSC. The skewed error ratio (SER), as defined in Puyol-Antón et al. (2021), is given by $SER = \frac{max_g(1-D_g)}{min_g(1-D_g)}$ where D_g is the median DSC for protected group g. This measures the ratio of the errors between the median DSCs for the worst-performing and best-performing protected groups. For both fairness metrics a low value represents less bias and a more fair model will have a FG closer to 0 and a SER closer to 1.

4.9 Code availability

The code to reproduce this work is available at `https://github.com/tiarnaleeKCL/nnUNet-bias-mitigation`.

5. Results

5.1 Internal validation

The internal validation results of the baseline and bias mitigation methods can be seen in Table 4 and Fig. 2. Statistical tests were performed using Mann-Whitney U tests on the overall DSC scores. Oversampling was the only method to increase performance for the Black subjects such that there was no significant difference between the median DSC scores of the Black and White subjects. The method increased the median DSC for Black subjects by 0.045. Fairness performance metrics (SER and FG) also decreased, showing more equitable performance. Using oversampling also caused performance to increase for the other ethnicities compared to the baseline (see Fig. 2). This is perhaps surprising as performance on these other ethnicities was not optimised during training. Further results on the effect of different levels of oversampling, both without and without cascaded cropping, can be seen in Fig. S2. Plots of performance (median DSC) against FG for these two approaches can be seen in Fig. S3. The effect of using

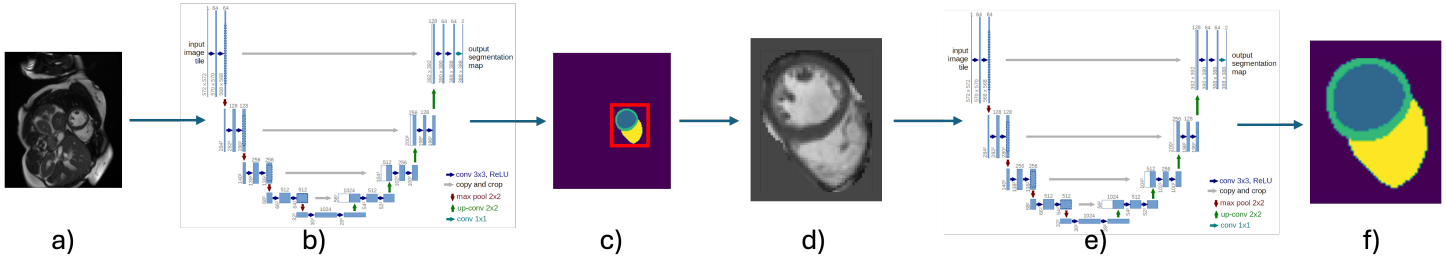

Figure 1: 'Cascaded' approach to bias mitigation based on using cropped images. Images (a) are first segmented using a full-image nnU-Net (b) to produce a segmentation (c). This segmentation is then used to crop the images (d) and used in another nnU-Net model trained using ground truth segmentation-based cropped images (e) to produce the final segmentations (f).

different proportions of Black and White subjects in the training set are reported in Fig. S1. Additional experiments using Asian and White subjects can be seen in Table S9 .

The other mitigation methods did not significantly improve performance for the Black subjects. Reweighing resulted in worse performance for the Black subjects, with the median DSC decreasing and HD increasing. Group DRO resulted in increased performance for the Black subjects but performance remained significantly lower than for the White subjects. Both oversampling and Group DRO slightly decreased median DSC for White subjects (although not statistically significantly), but reweighing significantly decreased median DSC for White subjects.

	Baseline *		Oversampling p = 0.22		Reweighing *		Group DRO *	
	White	Black	White	Black	White	Black	White	Black
Median DSC	**0.896**	0.846	0.894	**0.891**	0.891	0.831	0.893	0.865
IQR DSC	0.046	0.064	0.045	0.049	0.049	0.069	0.047	0.051
Median HD (mm)	6.725	9.364	6.835	7.164	7.174	10.037	6.802	8.286
IQR HD (mm)	3.707	4.725	3.713	3.861	4.207	4.902	3.581	4.157
SER	1.486		1.032		1.544		1.260	
Fairness gap	0.050		0.003		0.059		0.028	

Table 4: DSC and HD values for each of the bias mitigation methods tested on the internal validation set. The p-values were computed between White and Black subjects based on a two-sided Mann Whitney U test on the DSC scores. * p< 0.05. The best median DSC score for Black and White subjects is shown in bold.

Combining the mitigation methods did not produce significantly less biased results, as shown in Table 5. Mann-Whitney U tests were performed to test the significance of differences between the overall DSC scores. All three combinations decreased performance for White subjects but increased performance for Black subjects. The best combination was oversampling and Group DRO which reduced the performance for the White subjects the least and improved performance for the Black subjects the most, leading to the lowest FG.

As shown in Table 6, the baseline model trained using cropped images improved performance for both White and Black subjects and reduced bias. Oversampling significantly improved performance for the Black subjects compared to the baseline, resulting in performance that was higher than for White subjects. Group DRO improved the DSC for Black subjects compared to the baseline but this increase was not significant. However, the HD for both groups of subjects decreased. Reweighing made performance slightly worse for both ethnicities, with median DSC decreasing and SER and FG measures increasing.

	Baseline *		Oversampling + Group DRO *		Group DRO + Reweighing *		Oversampling + Reweighing *	
	White	Black	White	Black	White	Black	White	Black
Median DSC	**0.896**	0.846	0.889	**0.881**	0.893	0.863	0.889	0.871
IQR DSC	0.046	0.064	0.048	0.054	0.047	0.055	0.051	0.057
Median HD (mm)	6.725	9.364	6.985	7.740	6.831	8.417	7.364	8.654
IQR HD (mm)	3.707	4.725	3.666	4.079	3.520	4.001	4.320	4.558
SER	1.486		1.08		1.285		1.161	
Fairness gap	0.050		0.009		0.031		0.017	

Table 5: DSC and HD values for the combined bias mitigation methods tested on the internal validation set using original sized images. The p-values were computed between White and Black subjects based on a two-sided Mann Whitney U test on the DSC scores. * p< 0.05. The best median DSC score for Black and White subjects is shown in bold.

	Baseline *		Oversampling *		Reweighing *		Group DRO *	
	White	Black	White	Black	White	Black	White	Black
Median DSC	0.917	0.888	0.906	**0.914**	0.915	0.877	**0.917**	0.903
IQR DSC	0.039	0.047	0.031	0.044	0.036	0.052	0.035	0.046
Median HD (mm)	5.713	7.895	6.394	6.479	5.987	8.945	5.700	7.170
IQR HD (mm)	2.669	3.960	3.359	3.211	3.073	4.097	2.805	3.584
SER	1.340		1.088		1.448		1.161	
Fairness gap	0.028		-0.008		0.038		0.013	

Table 6: DSC and HD values for each of the bias mitigation methods tested on the internal validation set using cropped images. The p-values were computed between White and Black subjects based on a two-sided Mann Whitney U test on the DSC scores. The best median DSC score for Black and White subjects is shown in bold.

Table 7 and Figs. 2e to 2f show the results for the cascaded cropping on the internal validation dataset. Sur-

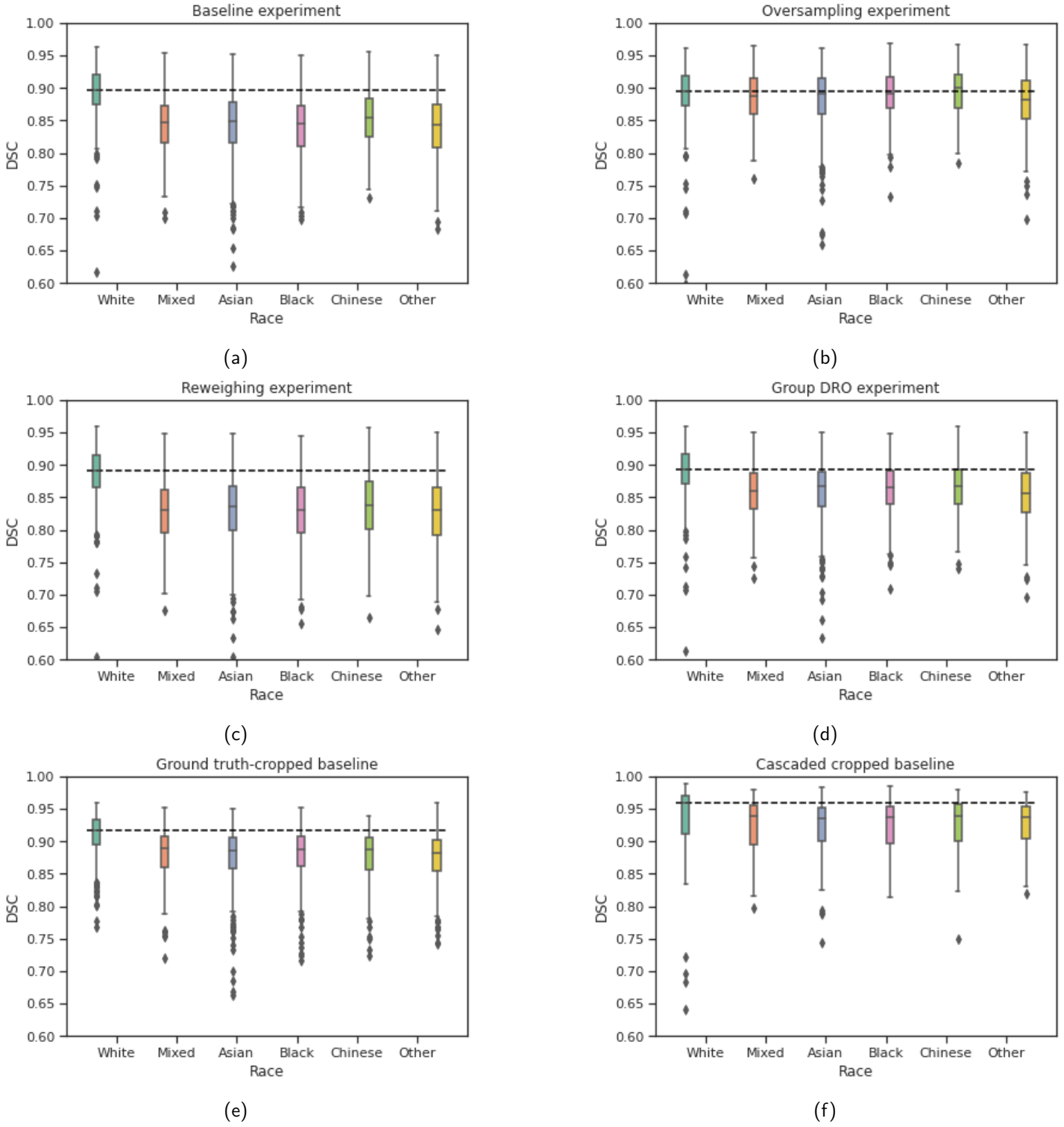

Figure 2: Overall DSC for bias mitigation methods on uncropped images. The dashed line indicates median DSC for White test subjects.

prisingly, using the cascaded cropping approach significantly improved the DSC score compared to using the ground-truth cropped results shown in Table 6 for all bias mitigation methods. We discuss a possible reason for this in Section 6.1. Using oversampling on the cascaded cropping approach did not improve DSC score but decreased the HD, SER and FG compared to the baseline. Reweighing also improved the SER and FG.

5.2 External validation

Table 8 shows the external validation results. Tables S1 to S3 show the results broken down by cardiac region. As oversampling was found to be the best bias mitigation on

	Baseline *		Oversampling *		Reweighing *		Group DRO *	
	White	Black	White	Black	White	Black	White	Black
Median DSC	**0.959**	0.937	0.958	0.941	0.951	0.935	0.955	**0.943**
IQR DSC	0.060	0.057	0.064	0.058	0.067	0.070	0.068	0.058
Median HD (mm)	3.757	6.564	3.731	4.929	5.176	6.426	3.841	4.623
IQR HD (mm)	2.268	5.012	2.129	2.717	3.262	4.172	2.039	3.327
SER	1.517		1.380		1.332		1.276	
Fairness gap	0.021		0.016		0.016		0.012	

Table 7: DSC and HD values for each of the bias mitigation methods using the cascaded cropping approach on the internal validation data. The p-values were computed between White and Black subjects based on a two-sided Mann Whitney U test on the DSC scores. * $p < 0.05$. The best median DSC score for Black and White subjects is shown in bold.

internal validation, only the results for this model and the baseline model are shown. Note that, as the LVM is not routinely segmented for the ES frame in clinical workflows at our institution, results for the external validation dataset are shown for the ED frames only.

	Baseline $p = 0.093$		Oversampling $p = 0.16$	
	White	Black	White	Black
Median DSC	**0.886**	**0.879**	0.885	0.872
IQR DSC	0.030	0.043	0.037	0.038
Median HD (mm)	5.728	5.913	6.152	5.919
IQR HD (mm)	1.522	1.966	1.238	1.840
SER	1.064		1.106	
Fairness gap	0.007		0.012	

Table 8: DSC and HD values for the baseline and oversampling methods tested on the external validation set. The p-values were computed between White and Black subjects based on a two-sided Mann Whitney U test on the DSC scores. * $p < 0.05$. The best median DSC score for Black and White subjects is shown in bold.

Overall, the model had high performance on the external validation set. The performance was comparable to that of the bias mitigation methods tested on the internal validation set, with the baseline performance for the Black subjects being higher than for the Black subjects in the internal validation set. The fairness gap and SER were also smaller than for the internal validation set. However, for the external validation set, using oversampling decreased performance for both groups, with the FG, SER and median HD increasing.

The approach based on cropping using ground truth segmentations produced a significantly better baseline performance in the external validation set for both groups compared to not using cropping, as shown in Table 9. Using oversampling slightly decreased the performance for both groups and resulted in a significantly different performance for Black and White subjects.

The results from using the cascaded cropping based approach can be seen in Table 9. Mann-Whitney U tests were used to compare the overall DSC scores. Furthermore, the errors in the sizes of the bounding boxes using the cascaded approach compared to those calculated using the ground-truth segmentations can be seen in Table S4. Tables S5 to S7 show the results broken down by cardiac region for the cascaded approach. Using this approach, performance was lower than the upper bound of using ground truth segmentations for cropping for both groups but performance remained higher than for the baseline model using uncropped images seen in Table 8. Performance increased for both groups when using oversampling and the FG decreased. Table S8 shows the results for the oversampling cascaded model broken down by sex and age.

6. Discussion

6.1 Internal validation

This work has performed a comprehensive examination of bias mitigation methods for AI segmentation models used for cine CMR images. We have shown that bias in CMR segmentation models can be mitigated by using such methods. In particular, oversampling minority subjects reduces bias so that there is no significant difference between the performance of the Black and White subjects. Although oversampling did not add any extra information to the dataset, the method allowed the network to train on Black subjects more frequently than if oversampling was not used, allowing for better balance between protected groups. It could be anticipated that training using a small number of (oversampled) Black subjects would increase the risk of overfitting to those subjects, having a detrimental effect on generalisation, but this effect was not seen in our experiments as test performance remained high. Previous work in Lee et al. (2025a) has showed that ethnicity can be classified from cine CMR images, indicating that there are distinct features in the images of different ethnicities that are recognisable to AI models. Using oversampling will allow the network to see more of these distinct features to learn better representations for the under-represented group.

Reweighing did not improve segmentation performance, instead decreasing performance for both protected groups. This may be due to increased importance being given to a small group of subjects, decreasing focus on the larger group. As described in Lee et al. (2022) and Lee et al. (2023), when White subjects comprised 75% of the training set, their segmentation performance was still lower than for the Black subjects who comprised 25% of the training set. This suggests that segmentation of the White subjects'

	Cropped baseline ground-truth model p = 0.069		Cropped oversampling ground-truth model *		Cropped baseline cascaded model p = 0.12		Cropped oversampling cascaded model p = 0.68	
	White	Black	White	Black	White	Black	White	Black
Median DSC	**0.931**	0.916	0.918	0.906	0.917	0.903	0.928	**0.920**
IQR DSC	0.043	0.037	0.040	0.033	0.045	0.038	0.039	0.026
Median HD (mm)	4.087	4.090	4.793	4.594	4.772	4.954	4.658	3.875
IQR HD (mm)	1.630	1.178	1.321	1.529	3.124	1.919	2.69	1. 432
SER	1.207		1.141		1.168		1.118	
Fairness gap	0.014		0.012		0.014		0.008	

Table 9: DSC and HD values for the models using cropping on the external validation dataset. The p-values were computed between White and Black subjects based on a two-sided Mann Whitney U test on the DSC scores. * $p < 0.05$. The best median DSC score for Black and White subjects is shown in bold.

images may be a more difficult task as there may be more outliers and variation in the hearts than in the Black subjects. Reweighing these White subjects so that their importance is lower in the loss function could decrease accuracy.

The combination of Group DRO and oversampling produced a model which improved performance for Black subjects but was still significantly lower than for White subjects. Therefore, using a combined method was better than using Group DRO alone but worse than oversampling alone.

Using cropped images increased performance and reduced bias compared to using uncropped images. Using oversampling with cropped images reduced bias further, but interestingly not to the same extent as using uncropped images. The cascaded cropping approach resulted in a performance that was higher than using ground truth segmentations for cropping. This may be due to the intermediate predictions on the uncropped images (c in Fig. 1) being less accurate and the segmented area being larger than the ground truth segmentations, resulting in a bounding box which was larger and contained more information. This suggests that performance could be optimised further by increasing the size of the pixel buffer used in the cropping process.

6.2 External validation

Using the bias mitigation models on the external validation set produced comparable DSC scores to internal validation results. No significant bias was observed in the baseline model for this dataset, as shown in Table 8. This is surprising as the dataset is out-of-distribution and some distributional shift is expected, which could lead to worse performance. However, unlike the models tested on the internal validation set, oversampling did not decrease the FG or improve the performance of the models under external validation. This is consistent with previous work that has reported limited effectiveness of generic bias mitigation algorithms under complex domain shifts (Schrouff et al., 2022; Yang et al., 2024).

Cropping the images increased performance overall, both using the ground truth-based cropping and the cascaded approach. Importantly, using the cascaded approach, performance was higher than the baseline model using uncropped images for both groups under external validation. Interestingly, using oversampling in combination with the cascaded approach improved performance for both groups further. These results indicate that, not only does cropping improve the performance of models for both protected groups (i.e. ethnicities), but also that ground truth segmentations are not needed to crop these images. This finding can be vital for the clinical translation of bias mitigation algorithms in CMR analysis, as better segmentation performance will allow for better assessment of clinical biomarkers and better treatment planning, prognosis and diagnosis. Provided that there is both a network trained on uncropped images and a network trained on cropped images, the method will be deployable and scalable. In the future, the first model used to localise the cardiac region of interest (b in Fig. 1) could be replaced with a similar network trained for bounding box detection such as that proposed in He et al. (2017).

The best method overall (cascaded cropping used with oversampling) was based on an understanding of the root cause of bias in AI-based CMR analysis as reported in Lee et al. (2025a). Specifically, in Lee et al. (2025a) it was reported that the main source of the distributional shift between ethnicities (and hence the bias in segmentation performance) was outside the heart. The cascaded cropping approach is a simple but elegant approach to removing the source of the bias in a robust way but maintaining (and even improving) segmentation performance. This important result highlights that, when using a bias understanding-informed mitigation approach in the context of CMR segmentation, there is no fairness-accuracy trade-

off. The fairness-accuracy trade-off is sometimes used as a reason not to implement bias mitigation algorithms in clinical practice, due to the medical principle of non-maleficence. Our work suggests that this trade-off is not inevitable, but that it can be avoided by a careful analysis of bias in individual scenarios rather than application of a generic mitigation algorithm.

6.3 Limitations and future work

This work has some limitations. For example, only two ethnicities (Black/White and Asian/White) were used during training in our experiments. Future work could investigate the effect of mitigating biases using multiple ethnicities, in other protected attributes such as age and socioeconomic status, or in intersectional groups. There was also a relatively small number of subjects in the external dataset when compared to the number of subjects in the training and internal validation sets from the UK Biobank, so externally validating on a larger, more diverse dataset would be beneficial. In future work we also plan to investigate more advanced baseline bias mitigation methods than the ones used in this work. However, it should be noted that the literature on bias mitigation in segmentation (e.g. Puyol-Antón et al. (2021); Benčević et al. (2024); Siddiqui et al. (2024)) is limited compared to the more extensive literature in classification problems. (R3.1) Finally, the different loss functions used for Group DRO compared to reweighing and oversampling may have resulted in different regularisation for the models so future work could investigate the use of a single loss function.

7. Conclusions

This paper has reported the most comprehensive investigation of bias mitigation in AI-based CMR segmentation to date. We have shown that the fairness-accuracy trade-off can be avoided by using a bias understanding-informed approach to mitigation, rather than using a generic mitigation algorithm. Therefore, this represents an important finding that should motivate further investigations into such bias understanding-informed approaches to mitigation in other applications.

Acknowledgments

Engineering & Physical Sciences Research Council Doctoral Training Partnership (EPSRC DTP) grant EP/T517963/1. This research has been conducted using the UK Biobank Resource under Application Number 17806.

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 and human subjects

Conflicts of Interest

E.P-A. is an employee of Heartflow Inc. All work presented is independent of her role at Heartfow. P-G.M. worked as a consultant for Perspectum Diagnostics Ltd until 2023. The remaining authors declare no conflict of interest.

Data availability

The UK Biobank dataset is publicly available for approved research projects. Requests to access the dataset should be directed to `https://www.ukbiobank.ac.uk/`. The clinical dataset cannot be publicly shared due to limitations of the ethical approval.

References

Samaneh Abbasi-Sureshjani, Ralf Raumanns, Britt E.J. Michels, Gerard Schouten, and Veronika Cheplygina. *Risk of Training Diagnostic Algorithms on Data with Demographic Bias*, volume 12446 LNCS. Springer International Publishing, 2020. ISBN 9783030611651. . URL `http://dx.doi.org/10.1007/978-3-030-61166-8_20`.

M. Adewole, J.D. Rudie, A. Gbadamosi, D. Zhang, C. Raymond, J. Ajigbotoshso, O. Toyobo, K. Aguh, O. Omidiji, Akinola R., M.A. Suwaid, A. Emegoakor, N. Ojo, C. Kalaiwo, Babatunde G., A. Ogunleye, Y. Gbadamosi, K. Iorpagher, Onuwaje M., Betiku B., R. Saluja, B. Menze, U. Baid, S. Bakas, F. Dako, Fatade A., and U.C. Anazodo. Expanding the Brain Tumor Segmentation (BraTS) data to include African Populations (BraTS-Africa) (version 1) [Dataset]. *The Cancer Imaging Archive*, 2024. URL `https://www.cancerimagingarchive.net/collection/brats-africa/`.

Thania Balducci, Jalil Rasgado-Toledo, Alely Valencia, Marie-José Van Tol, André Aleman, and Eduardo A Garza-Villarreal. A behavioral and brain imaging dataset with focus on emotion regulation of women with fibromyalgia. *Scientific Data volume*, 9(581), 2022. . URL `www.nature.com/scientificdata`.

Raquel R. Barbieri, Yixi Xu, Lucy Setian, Paulo Thiago Souza-Santos, Anusua Trivedi, Jim Cristofono, Ricardo Bhering, Kevin White, Anna M. Sales, Geralyn Miller, José Augusto C. Nery, Michael Sharman, Richard

Bumann, Shun Zhang, Mohamad Goldust, Euzenir N. Sarno, Fareed Mirza, Arielle Cavaliero, Sander Timmer, Elena Bonfiglioli, Cairns Smith, David Scollard, Alexander A. Navarini, Ann Aerts, Juan Lavista Ferres, and Milton O. Moraes. Reimagining leprosy elimination with AI analysis of a combination of skin lesion images with demographic and clinical data. *Lancet regional health. Americas*, 9, 5 2022. ISSN 2667-193X. . URL https://pubmed.ncbi.nlm.nih.gov/36776278/.

Marin Benčević, Marija Habijan, Irena Galić, Danilo Babin, and Aleksandra Pižurica. Understanding skin color bias in deep learning-based skin lesion segmentation. *Computer Methods and Programs in Biomedicine*, 245:108044, 3 2024. ISSN 0169-2607. .

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. *Science Advances*, 8(32):6147, 8 2022. ISSN 23752548. . URL https://www.science.org/doi/10.1126/sciadv.abq6147.

Georgios Georgiopoulos, Luca Faconti, Aqeel T. Mohamed, Stefano Figliozzi, Clint Asher, Louise Keehn, Ryan McNally, Khaled Alfakih, Samuel Vennin, Amedeo Chiribiri, Pablo Lamata, Philip Chowienczyk, and Pier Giorgio Masci. Ethnicity differences in geometric remodelling and myocardial composition in hypertension unveiled by cardiovascular magnetic resonance. *European Heart Journal - Cardiovascular Imaging*, 25(7):901–911, 6 2024. ISSN 2047-2404. . URL https://dx.doi.org/10.1093/ehjci/jeae097.

Ben Glocker, Charles Jones, Mélanie Bernhardt, and Stefan Winzeck. Algorithmic encoding of protected characteristics in chest X-ray disease detection models. *EBioMedicine*, 89, 3 2023. ISSN 2352-3964. . URL https://pubmed.ncbi.nlm.nih.gov/36791660/.

Philippe Gottfrois, Fabian Gröger, Faly Herizo Andriambololoniaina, Ludovic Amruthalingam, Alvaro Gonzalez-Jimenez, Christophe Hsu, Agnes Kessy, Simone Lionetti, Daudi Mavura, Wingston Ng'ambi, Dingase Faith Ngongonda, Marc Pouly, Mendrika Fifaliana Rakotoarisaona, Fahafahantsoa Rapelanoro Rabenja, Ibrahima Traoré, and Alexander A. Navarini. PASSION for Dermatology: Bridging the Diversity Gap with Pigmented Skin Images from Sub-Saharan Africa. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 703–712, 2024. ISSN 1611-3349. . URL https://link.springer.com/chapter/10.1007/978-3-031-72384-1_66.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):386–397, 3 2017. ISSN 19393539. . URL https://arxiv.org/abs/1703.06870v3.

Stefanos Ioannou, Hana Chockler, Alexander Hammers, and Andrew P. King. A Study of Demographic Bias in CNN-Based Brain MR Segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13596 LNCS:13–22, 2022. ISSN 1611-3349. . URL https://link.springer.com/chapter/10.1007/978-3-031-17899-3_2.

Fabian Isensee, Paul F. Jaeger, Simon A.A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods 2020 18:2*, 18(2):203–211, 12 2020. ISSN 1548-7105. .

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 12 2012. ISSN 02193116. . URL https://link.springer.com/article/10.1007/s10115-011-0463-8.

Faisal Kamiran, Asim Karim, and Xiangliang Zhang. Decision theory for discrimination-aware classification. *Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 924–929, 2012. ISSN 15504786. .

Malte Klingenberg, Didem Stark, Fabian Eitel, Céline Budding, Mohamad Habes, and Kerstin Ritter. Higher performance for women than men in MRI-based Alzheimer's disease detection. *Alzheimer's Research and Therapy*, 15(1):1–13, 12 2023. ISSN 17589193. . URL https://alzres.biomedcentral.com/articles/10.1186/s13195-023-01225-6http://creativecommons.org/publicdomain/zero/1.0/.

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 of the United States of America*, 117(23):12592–12594, 2020. ISSN 10916490. .

Tiarna Lee, Esther Puyol-Antón, Bram Ruijsink, Miaojing Shi, and Andrew P. King. A Systematic Study of Race and Sex Bias in CNN-Based Cardiac MR Segmentation. In *Lecture Notes in Computer Science (including subseries*

Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 13593 LNCS, pages 233–244. Springer Science and Business Media Deutschland GmbH, 2022. ISBN 9783031234422. .

Tiarna Lee, Esther Puyol-Antón, Bram Ruijsink, Keana Aitcheson, Miaojing Shi, and Andrew P. King. An Investigation into the Impact of Deep Learning Model Choice on Sex and Race Bias in Cardiac MR Segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 14242 LNCS:215–224, 2023. ISSN 16113349. . URL https://link.springer.com/chapter/10.1007/978-3-031-45249-9_21.

Tiarna Lee, Esther Puyol-Antón, Bram Ruijsink, Sebastien Roujol, Theodore Barfoot, Shaheim Ogbomo-Harmitt, Miaojing Shi, and Andrew King. An investigation into the causes of race bias in AI-based cine CMR segmentation. *European Heart Journal - Digital Health*, 2 2025a. ISSN 2634-3916. . URL https://dx.doi.org/10.1093/ehjdh/ztaf008.

Tiarna Lee, Esther Puyol-Antón, Bram Ruijsink, Miaojing Shi, and Andrew P. King. Does a Rising Tide Lift All Boats? Bias Mitigation for AI-based CMR Segmentation. *arXiv*, 3 2025b. URL https://arxiv.org/abs/2503.17089v1.

Jingyang Li and Guoqiang Li. The Triangular Trade-off between Robustness, Accuracy and Fairness in Deep Neural Networks: A Survey. *ACM Computing Surveys*, 2024. ISSN 0360-0300. .

Pranay K. Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R. Varshney, and Ruchir Puri. Bias Mitigation Post-processing for Individual and Group Fairness. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2019-May:2847–2851, 5 2019. ISSN 15206149. .

David Madras, Elliot Creager, Toniann Pitassi, and Richards Zemel. Learning adversarially fair and transferable representations. *35th International Conference on Machine Learning, ICML 2018*, 8:5423–5434, 2018.

Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Minimax Pareto Fairness: A Multi Objective Perspective. *Proceedings of machine learning research*, 119:6755, 2020. ISSN 2640-3498. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC7912461/.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6), 8 2019. ISSN 15577341. . URL https://arxiv.org/abs/1908.09635v3.

Vincent Olesen, Nina Weng, Aasa Feragen, and Eike Petersen. Slicing Through Bias: Explaining Performance Gaps in Medical Image Analysis Using Slice Discovery Methods. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 15198 LNCS:3–13, 2025. ISSN 1611-3349. . URL https://link.springer.com/chapter/10.1007/978-3-031-72787-0_1.

Eike Petersen, Aasa Feragen, Maria Luise da Costa Zemsch, Anders Henriksen, Oskar Eiler Wiese Christensen, and Melanie Ganz. Feature Robustness and Sex Differences in Medical Imaging: A Case Study in MRI-Based Alzheimer's Disease Detection. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13431 LNCS:88–98, 2022. ISSN 16113349. . URL https://link.springer.com/chapter/10.1007/978-3-031-16431-6_9.

Steffen E. Petersen, Paul M. Matthews, Jane M. Francis, Matthew D. Robson, Filip Zemrak, Redha Boubertakh, Alistair A. Young, Sarah Hudson, Peter Weale, Steve Garratt, Rory Collins, Stefan Piechnik, and Stefan Neubauer. UK Biobank's cardiovascular magnetic resonance protocol. *Journal of Cardiovascular Magnetic Resonance*, 18 (1):1–7, 2 2016. ISSN 1532429X. .

Esther Puyol-Antón, Bram Ruijsink, Stefan K. Piechnik, Stefan Neubauer, Steffen E. Petersen, Reza Razavi, and Andrew P. King. Fairness in Cardiac MR Image Analysis: An Investigation of Bias Due to Data Imbalance in Deep Learning Based Segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, volume 12903 LNCS, pages 413–423. Springer International Publishing, 2021. ISBN 9783030871987. .

Esther Puyol-Antón, Bram Ruijsink, Jorge Mariscal Harana, Stefan K. Piechnik, Stefan Neubauer, Steffen E. Petersen, Reza Razavi, Phil Chowienczyk, and Andrew P. King. Fairness in Cardiac Magnetic Resonance Imaging: Assessing Sex and Racial Bias in Deep Learning-Based Segmentation. *Frontiers in Cardiovascular Medicine*, 0: 664, 4 2022. ISSN 2297-055X. .

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. DISTRIBUTIONALLY ROBUST NEURAL NETWORKS FOR GROUP SHIFTS: ON THE IMPORTANCE OF REGULARIZATION FOR WORST-CASE GENERALIZATION. In *8th International Conference on Learning Representations, ICLR 2020*, 2020.

Jessica Schrouff, Natalie Harris, Google Research, Oluwasanmi Koyejo, Ibrahim Alabdulmohsin, Eva Schnider, Krista Opsahl-Ong, Alex Brown, Subhrajit Roy, Diana Mincu, Christina Chen, Awa Dieng, Yuan Liu, Vivek Natarajan, Alan Karthikesalingam, Katherine Heller, Silvia Chiappa, Deepmind Alexander D', and Amour Google Research. Diagnosing failures of fairness transfer across distribution shift in real-world medical settings. *Advances in Neural Information Processing Systems*, 35:19304–19318, 12 2022.

Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y. Chen, and Marzyeh Ghassemi. CheXclusion: Fairness gaps in deep chest X-ray classifiers. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 26:232–243, 2021a. ISSN 23356936. .

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. *Nature Medicine*, 27(12):2176–2182, 2021b. ISSN 1078-8956. .

Ismaeel A Siddiqui, Nickolas Littlefield, Luke A Carlson, Matthew Gong, Avani Chhabra, Zoe Menezes, George M Mastorakos, Sakshi Mehul Thakar, Mehrnaz Abedian, Ines Lohse, Kurt R Weiss, Johannes F Plate, Hamidreza Moradi, Soheyla Amirian, and Ahmad P Tafti. Fair AI-powered orthopedic image segmentation: addressing bias and promoting equitable healthcare. *Scientific Reports —*, 14:16105, 2024. . URL https://doi.org/10.1038/s41598-024-66873-6.

Yu Tian, Min Shi, Yan Luo, Ava Kouhana, Tobias Elze, and Mengyu Wang. FairSeg: A Large-Scale Medical Image Segmentation Dataset for Fairness Learning Using Segment Anything Model with Fair Error-Bound Scaling. *arXiv*, 11 2023. URL https://arxiv.org/abs/2311.02189v5.

Rongguang Wang, Pratik Chaudhari, and Christos Davatzikos. Bias in machine learning models can be significantly mitigated by careful training: Evidence from neuroimaging studies. *Proceedings of the National Academy of Sciences of the United States of America*, 120(6):e2211613120, 2 2023. ISSN 10916490. . URL https://www.pnas.org/doi/abs/10.1073/pnas.2211613120.

Yuzhe Yang, Haoran Zhang, Judy W. Gichoya, Dina Katabi, and Marzyeh Ghassemi. The limits of fair medical imaging AI in real-world generalization. *Nature Medicine 2024 30:10*, 30(10):2838–2848, 6 2024. ISSN 1546-170X. . URL https://www.nature.com/articles/s41591-024-03113-4.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating Unwanted Biases with Adversarial Learning. *AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018. ISSN 23318422. .

Haoran Zhang, Natalie Dullerud, Karsten Roth, Lauren Oakden-Rayner, Stephen Robert Pfohl, and Marzyeh Ghassemi. Improving the Fairness of Chest X-ray Classifiers. *arXiv*, page 2022, 3 2022. . URL https://arxiv.org/abs/2203.12609v1.

Yongshuo Zong, Yongxin Yang, and Timothy Hospedales. MEDFAIR: Benchmarking Fairness for Medical Imaging. *arXiv*, 10 2022. . URL https://arxiv.org/abs/2210.01725v1.

Supplementary Material

Metrics for regions of the heart using the external validation set

	Baseline p = 0.11		Oversampling p = 0.46	
	White	Black	White	Black
Median DSC	**0.940**	0.931	0.936	**0.937**
IQR DSC	0.041	0.035	0.038	0.032
Median HD (mm)	3.929	3.416	3.705	3.518
IQR HD (mm)	1.622	1.498	1.296	2.186
SER	1.146		1.007	
Fairness gap	0.009		-0.0004	

Table S1: Metrics for LVBP segmentation based on inference performed on external validation set using different models trained on UK Biobank data. The p-values were computed between White and Black subjects based on a two-sided Mann Whitney U test on the DSC scores. * p< 0.05. The best median DSC score for Black and White subjects is shown in bold.

	Baseline p = 0.84		Oversampling p = 0.57	
	White	Black	White	Black
Median DSC	**0.817**	**0.803**	0.808	0.802
IQR DSC	0.046	0.075	0.053	0.06
Median HD (mm)	5.026	5.141	5.096	5.318
IQR HD (mm)	1.583	1.941	1.995	2.042
SER	1.077		1.034	
Fairness gap	0.014		0.006	

Table S2: Metrics for LVM segmentation based on inference performed on the external validation set using different models trained on UK Biobank data. The p-values were computed between White and Black subjects based on a two-sided Mann Whitney U test on the DSC scores. * p< 0.05. The best median DSC score for Black and White subjects is shown in bold.

	Baseline p = 0.43		Oversampling p = 0.27	
	White	Black	White	Black
Median DSC	0.883	**0.882**	**0.885**	0.868
IQR DSC	0.057	0.048	0.070	0.071
Median HD (mm)	8.902	7.783	8.869	10.181
IQR HD (mm)	3.208	3.096	3.962	4.401
SER	1.004		1.148	
Fairness gap	0.0004		0.017	

Table S3: Metrics for RVBP segmentation based on inference performed on the external validation set using different models trained on UK Biobank data. The p-values were computed between White and Black subjects based on a two-sided Mann Whitney U test on the DSC scores. * p< 0.05. The best median DSC score for Black and White subjects is shown in bold.

Difference in the size of the bounding box between images

Experiment	Error x (%)	Error y (%)
Baseline	3.70	6.57
Oversampling	0.0	7.30
Reweighing	1.48	8.76
Group DRO	0.0	0.730

Table S4: Difference in the size of the bounding box between images cropped using the ground truth and cascaded approach for the internal set.

Metrics for regions of the heart for the external validation set using the cascaded cropping method

	Cropped baseline p = 0.088		Cropped oversampling p = 0.76		Cropped baseline cascaded model *		Cropped oversampling cascaded model p = 0.72	
	White	Black	White	Black	White	Black	White	Black
Median DSC	**0.962**	**0.949**	0.958	0.943	0.959	0.943	0.959	0.946
IQR DSC	0.030	0.031	0.024	0.027	0.031	0.029	0.036	0.024
Median HD (mm)	2.887	2.488	3.043	2.870	3.132	2.839	2.842	2.493
IQR HD (mm)	0.984	0.854	1.113	1.697	1.119	1.049	1.352	2.646
SER	1.309		1.364		1.398		1.316	
Fairness gap	0.012		0.015		0.016		0.013	

Table S5: Metrics for LVBP segmentation based on inference performed on the external validation set using the cascaded approach to crop the images. The p-values were computed between White and Black subjects based on a two-sided Mann Whitney U test on the DSC scores. * p< 0.05. The best median DSC score for Black and White subjects is shown in bold.

	Cropped baseline p = 0.39		Cropped oversampling p = 0.63		Cropped baseline cascaded model p = 0.69		Cropped oversampling cascaded model p = 0.86	
	White	Black	White	Black	White	Black	White	Black
Median DSC	**0.881**	0.860	0.852	0.849	0.856	0.855	0.879	**0.863**
IQR DSC 7	0.103	0.093	0.096	0.083	0.093	0.106	0.071	0.066
Median HD (mm)	3.512	3.310	4.071	3.937	4.051	4.008	3.422	3.259
IQR HD (mm)	1.738	1.251	2.030	2.492	2.532	2.098	2.051	3.120
SER	1.175		1.019		1.009		1.130	
Fairness gap	0.021		0.003		0.001		0.016	

Table S6: Metrics for LVM segmentation based on inference performed on external validation set using the cascaded approach to crop the images. The p-values were computed between White and Black subjects based on a two-sided Mann Whitney U test on the DSC scores. * p$<$ 0.05. The best median DSC score for Black and White subjects is shown in bold.

	Cropped baseline p = 0.75		Cropped oversampling p = 0.71		Cropped baseline cascaded model p =0.84		Cropped oversampling cascaded model *	
	White	Black	White	Black	White	Black	White	Black
Median DSC	**0.938**	0.925	0.918	0.907	0.912	0.905	0.929	**0.932**
IQR DSC	0.024	0.022	0.034	0.026	0.045	0.040	0.049	0.031
Median HD (mm)	5.932	5.787	6.540	6.662	7.477	7.147	4.871	6.426
IQR HD (mm)	2.580	3.557	2.991	2.753	4.459	5.708	2.018	4.328
SER	1.197		1.136		1.131		1.036	
Fairness gap	0.012		0.011		0.011		-0.003	

Table S7: Metrics for RVBP segmentation based on inference performed on external validation set using the cascaded approach to crop the images. The p-values were computed between White and Black subjects based on a two-sided Mann Whitney U test on the DSC scores. * p$<$ 0.05. The best median DSC score for Black and White subjects is shown in bold.

DSC and HD values by sex and age for the cascaded approach

	Cropped baseline cascaded model p = 0.12		Cropped baseline cascaded model p = 0.68	
	Female	Male	Age $<$ 50	Age $>$ 50
n	28	56	52	32
Median DSC	0.913	0.927	0.922	0.922
IQR DSC	0.032	0.036	0.032	0.031
Median HD (mm)	4.069	4.143	3.869	4.208
IQR HD (mm)	1.927	1.402	1.558	1.738
SER	1.193		1.004	
Fairness gap	0.0141		-0.0003	

Table S8: DSC and HD values by sex and age for the cascaded approach using oversampling on the external validation set. The p-values were computed based on a two-sided Mann Whitney U test on the DSC scores. * p$<$ 0.05. FG was calculated using $DSC_{male} - DSC_{female}$ and $DSC_{Age<50} - DSC_{Age>50}$

Metrics for the bias mitigation methods using White and Asian subjects

	Baseline *		Oversampling p = 0.17		Reweighing *		Group DRO *		Cropped baseline cascaded model *		Cropped oversampling cascaded model *	
	White	Asian	White	Asian	White	Asian	White	Asian	White	Asian	White	Asian
Median DSC	0.901	0.851	0.8990	0.9039	0.8971	0.8349	0.8971	0.8704	**0.9584**	**0.9362**	0.9576	0.8949
IQR DSC	0.0452	0.0635	0.0445	0.0678	0.0501	0.0686	0.0471	0.0584	0.0559	0.0502	0.0581	0.0583
Median HD (mm)	6.761	9.585	6.783	7.204	7.051	10.057	6.688	8.628	3.957	7.254	3.862	8.093
IQR HD (mm)	3.690	4.935	3.715	3.662	4.140	5.076	3.744	4.310	2.470	5.115	2.321	5.491
SER	1.498		1.051		1.604		1.260		1.533		2.477	
Fairness gap	0.0494		-0.0049		0.0621		0.0267		0.0222		0.0627	

Table S9: DSC and HD values for each of the bias mitigation methods tested on the internal validation set. The p-values were computed between White and Asian subjects based on a two-sided Mann Whitney U test on the DSC scores. * $p < 0.05$. The best median DSC score for Asian and White subjects is shown in bold.

The effect of changing the number of Black subjects in the training dataset

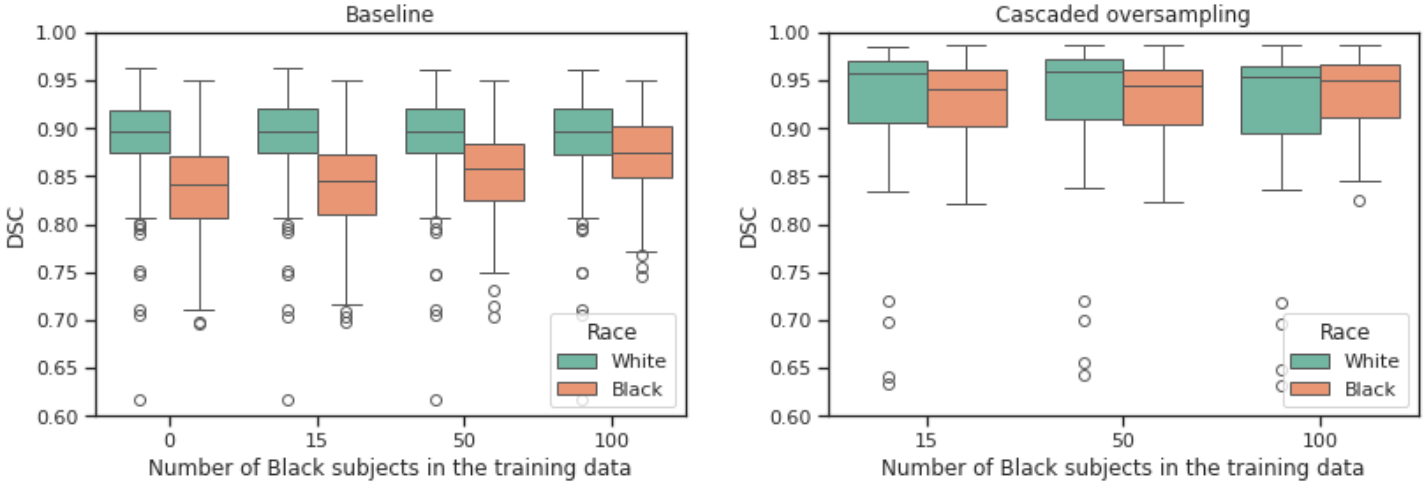

Figure S1: The effect of changing the number of Black subjects in the training dataset.

The effect of changing the level of oversampling in the training dataset

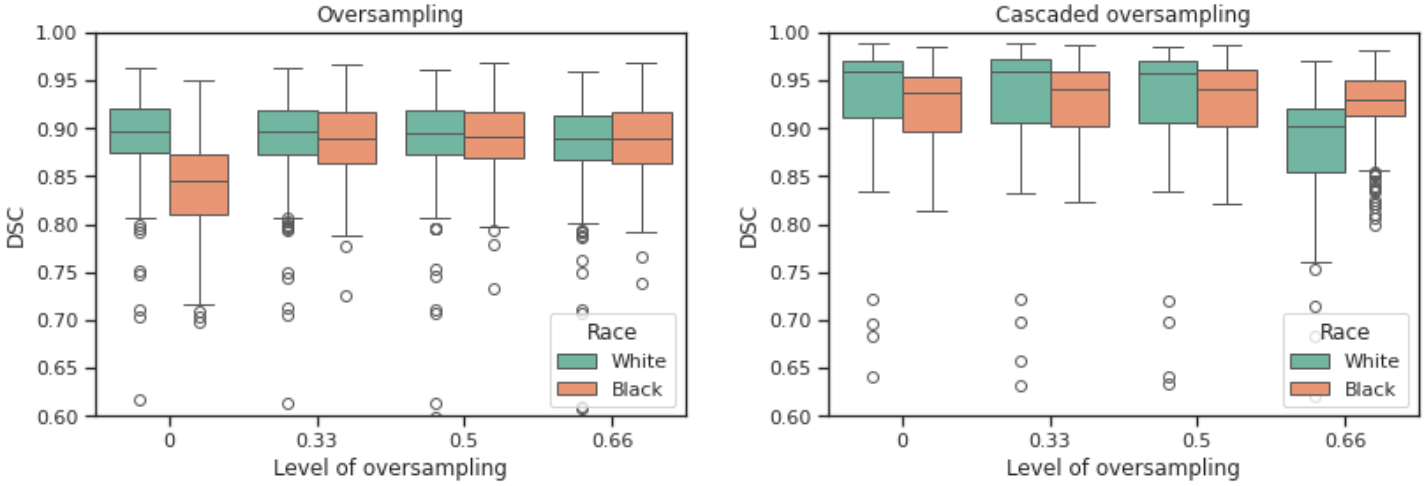

Figure S2: The effect of changing level of oversampling of Black subjects in the training dataset.

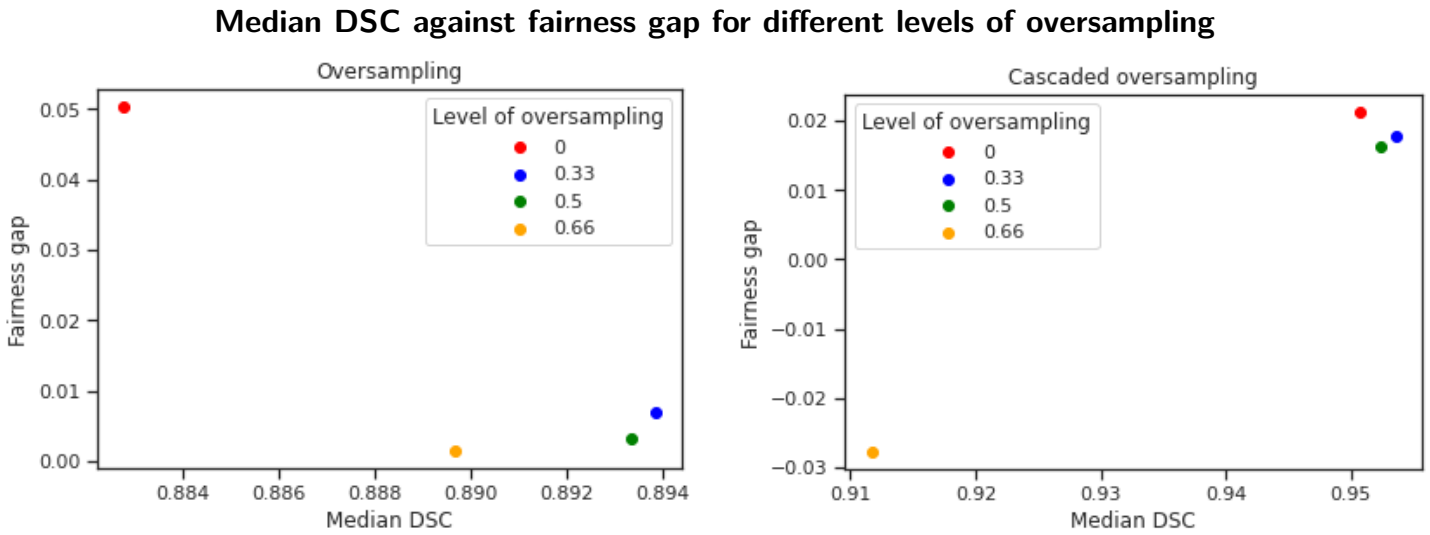

Figure S3: Median DSC against fairness gap for different levels of oversampling in experiments using oversampling and cascaded cropping + oversampling. The fairness gap is calculate by subtracting the median DSC for Black subjects from the median DSC for White subjects.