Divergence-Aware Training with Automatic Subgroup Mitigation for Breast Tumor Segmentation

Eleonora Poeta[1 [0009−0001−7289−0036]], Luisa Vargas[2 [0009−0006−5741−8485]], Daniele Falcetta[2 [0009−0009−7199−5424]], Vincenzo Marciano'[2 [0009−0000−2238−8040]], Eliana Pastor[1 [0000−0002−3664−4137]], Tania Cerquitelli[1 [0000−0002−9039−6226]], Elena Baralis[1 [0000−0001−9231−467X]], and Maria A. Zuluaga[2 [0000−0002−1147−766X]]

[1] Politecnico di Torino, Turin, Italy {eleonora.poeta, eliana.pastor, tania.cerquitelli, elena.baralis}@polito.it
[2] EURECOM, Biot, France {luisa.vargas, daniele.falcetta, vincenzo.marciano, maria.zuluaga}@eurecom.fr

Abstract. Deep learning models for breast tumor segmentation in DCE-MRI may exhibit disparities in performance across demographic and clinical subgroups, raising concerns about fairness and clinical trustworthiness. In this work, we propose a subgroup-aware in-processing mitigation strategy that integrates divergence-based regularization directly into the training loop. By leveraging interpretable metadata (e.g., age, menopausal status, breast density), we identify subgroups where the model underperforms and assign higher loss weights to these samples in proportion to their divergence from average performance. Our method enables the model to focus training on underrepresented or harder-to-segment subpopulations, without requiring external data or post-processing correction. We evaluate our approach on the MAMA-MIA 2025 challenge dataset, demonstrating improvements in both overall segmentation quality and fairness score. Our results highlight the potential of in-processing mitigation as an effective and practical pathway toward equitable medical image segmentation.

Keywords: Fairness · Breast Tumor Segmentation · Bias Mitigation · Subgroup Analysis

1 Introduction

Breast tumor delineation using dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) plays a crucial role in the diagnosis, treatment planning, and monitoring of breast cancer [15, 32]. Advances in deep learning techniques for medical image segmentation have led to state-of-the-art models that achieve remarkable accuracy in the task [22]. However, these models may be susceptible to algorithmic unfairness, meaning they perform unevenly across patient subgroups based on protected attributes such as ethnicity, age, or socioeconomic

background [8, 23, 28]. These disparities undermine the model's fairness and the overall trust in AI systems.

In-processing strategies have been proposed to improve fairness dynamically during model training, without relying on external data or post-hoc adjustments [26, 31]. Existing techniques typically focus on mitigating disparities related to a single sensitive attribute, such as sex [2, 19], age, or clinical site [27]. However, focusing solely on one attribute at a time may overlook disparities arising at the intersection of multiple factors. Moreover, these methods often require a priori knowledge of which subgroups to monitor or protect.

In this work, we propose FairMedSeg, an in-processing mitigation strategy tailored for medical image segmentation. FairMedSeg automatically identifies subgroups that exhibit *divergent* performance, defined as statistically significant deviations from the population-level average [17], and reduces such disparities during training. We achieve this by extending the divergence-based reweighting framework proposed initially by Koudounas et al. [10] for speech processing, and adapting it to the specific challenges of 3D medical image segmentation. We validate our method on the MAMA-MIA 2025 challenge dataset [5], which provides annotated DCE-MRI breast scans enriched with clinical metadata, demonstrating that FairMedSeg improves segmentation accuracy while reducing fairness gaps across clinically meaningful subgroups.

Our approach introduces several key innovations for in-processing fairness mitigation in this domain: (i) an automatic and intersectional subgroup discovery pipeline tailored to clinical metadata; (ii) a fully dynamic, performance-driven sample reweighting mechanism integrated directly into the training loop; (iii) a test-time-agnostic design that eliminates the need for metadata at inference, since fairness is addressed during training; (iv) the removal of any dependence on predefined sensitive attributes, enabling fairness-aware learning without prior subgroup specification.

2 Related Work

Fairness-enhancing strategies can be broadly categorized into *pre-processing*, *post-processing*, and *in-processing* methods [4, 12].

Pre-processing techniques modify the training data to reduce biases before model learning. Strategies include dataset rebalancing [11], generative data augmentation [18], and harmonization techniques to eliminate confounding variables [21]. Causal perturbation methods create counterfactual samples by altering sensitive attributes while maintaining clinical validity [20]. Although model-agnostic and easy to implement, these methods may distort natural data distributions or struggle to generalize across domains.

Post-processing methods operate on model outputs or internal predictions after training is complete, without altering the model architecture or input data. Common techniques include subgroup-specific threshold adjustment [6] and post-hoc confidence calibration [14]. While effective in some cases, these methods

typically rely on access to metadata at inference time, an assumption that may not hold in real-world clinical deployments.

In-processing strategies incorporate fairness constraints into the training procedure. Adversarial approaches [31] learn representations that are predictive of the target task while invariant to sensitive attributes, but can be unstable, compromising task performance, particularly in high-dimensional data. Group Distributionally Robust Optimization [25] and fairness-aware regularization [26] penalize disparities across predefined subgroups, yet require accurate group labels and can struggle to scale when subgroup definitions are noisy or evolving. Moreover, they are primarily developed for classification tasks and may not translate directly to structured prediction settings such as medical image segmentation.

Divergence-based regularization [10] is a recent in-processing mitigation technique that dynamically adjusts training based on disparities among subgroups. It identifies subgroups whose performance deviates from the overall model accuracy and assigns higher weights to the samples from those groups. This encourages the model to focus on learning from the divergent subgroups. So far, it has been used in the context of speech processing. In this work, we adapt the divergence-aware framework to 3D medical image segmentation, extending the method to automatically discover vulnerable subgroups based on clinical metadata. Unlike classical reweighting approaches based on group frequency statistics [9] or meta-learned strategies [29], our method leverages subgroup-level performance divergence as a dynamic signal to guide training. In contrast to focal-loss-based techniques [1, 30], which prioritize sample-level difficulty, our approach explicitly addresses performance disparities across clinically meaningful subgroups defined by metadata.

3 Method

3.1 Problem Formulation

Let $D = \{(x_i, y_i, m_i)\}_{i=1}^{N}$ denote a dataset of N annotated DCE-MRIs, where x_i is the input image, y_i the corresponding ground truth segmentation mask, and m_i a vector of clinical metadata attributes (e.g., age, menopausal status, breast density), i.e., $m_i = (m_{i,1}, \ldots, m_{i,k})$, with k the number of attributes. A segmentation model M_θ, is trained to predict $\hat{y}_i = M_\theta(x_i)$ by minimizing a segmentation loss function $\mathcal{L}_{\text{seg}}$ over the training data.

We define a *subgroup* S as a subset of samples sharing specific metadata characteristics, i.e., $S = \{(x_i, y_i, m_i) \in D \mid m_{i,j_1} = m^*_{j_1} \wedge m_{i,j_2} = m^*_{j_2} \wedge \ldots \wedge m_{i,j_p} = m^*_{j_p}\}$ with $j_1, j_2, \ldots, j_p$ the indices of the features within m_i. For instance, a subgroup can be defined by $\{$*age = 41–50*, *menopausal status = pre*$\}$. Let $f(S, M_\theta)$ represent the model's performance on subgroup S (e.g., mean Dice score), and $f(D, M_\theta)$ its performance over the full dataset. The divergence of subgroup S is [16]:

$$\Delta_f(S, M_\theta) = f(S, M_\theta) - f(D, M_\theta), \tag{1}$$

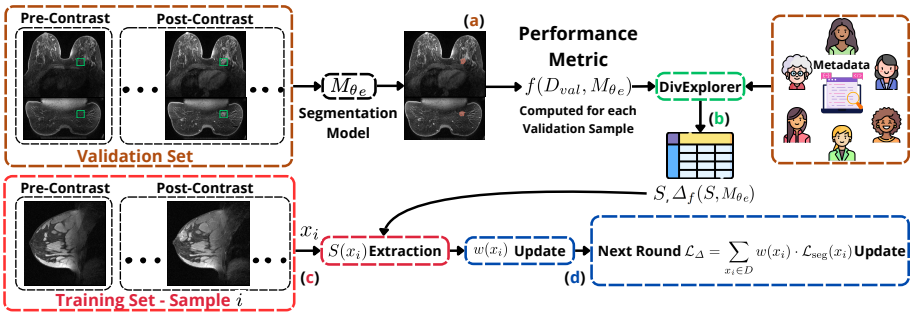

Fig. 1. Overview of FairMedSeg weight update scheme. After each training epoch e, performance is evaluated on a validation set (a), and metadata-defined subgroups are discovered using `DivExplorer` (b). Sample weights are updated based on subgroup divergence, prioritizing underperforming subgroups (c). These weights are used to compute a fairness-aware loss for the next epoch, iteratively reducing subgroup disparities throughout training (d).

In our setting, where we measure model performance using the average Dice score, negative values indicate performance below the global average. Our goal is to reduce divergence while preserving overall segmentation accuracy by defining a training strategy that is sensitive to subgroup disparities and can proactively mitigate them.

3.2 FairMedSeg Mitigation Strategy

Figure 1 provides an overview of the training process for FairMedSeg. During each training epoch e, the resulting model M_{θ_e} is evaluated using a validation set, which allows us to assess the model's performance. By exploiting the metadata available at training (e.g., age, menopausal status, breast density), we automatically identify subgroups and estimate their divergence. The estimated performance gaps for each subgroup are then used to update sample weights and adjust the loss function in the subsequent epoch. By reducing fairness gaps at training, the final model M_θ is test-time-agnostic by eliminating the need for metadata at inference.

Subgroup Discovery. We define subgroups as conjunctions of interpretable clinical metadata attributes (e.g., menopausal status, age group, breast density). To systematically construct these subgroups, we leverage the `DivExplorer` framework [16], which enumerates all combinations of metadata attributes that satisfy a minimum support threshold, denoted as *minsup*. This threshold ensures that each subgroup is sufficiently represented in the dataset, enabling reliable estimation of performance metrics.

Divergence-Based Sample Weighting. We assign training sample weights based on subgroup-level performance divergences identified during step (b). Each training sample x_i may belong to one or more metadata-defined subgroups,

denoted as $S(x_i) = \{S \in \mathcal{S} \mid x_i \in S\}$. For example, a 45-year-old patient in pre-menopause belongs not only to the marginal subgroups $\{age = 45\}$ and $\{menopausal\ status = pre\}$ but also to the intersectional subgroup $\{age = 45,\ menopausal\ status = pre\}$. After each training epoch, we evaluate the model of current epoch e (M_{θ_e}) on the validation set and compute the divergence score $\Delta_f(S, M_{\theta_e})$ for each subgroup $S \in \mathcal{S}$ (Eq. 1).

To update training priorities, we assign each training sample a weight equal to the maximum absolute divergence across the subgroups to which it belongs:

$$w(x_i) = \max_{S \in S(x_i)} |\Delta_f(S, M_{\theta_e})|. \tag{2}$$

The updated weights are then used in the next training epoch via $\mathcal{L}_{\text{total}}$ (Eq. 4). This weighting mechanism increases the influence of samples from subgroups where the model exhibits the most divergent behavior, encouraging the model to focus on those more challenging or underperforming samples.

Loss Formulation. The loss function is composed of two terms: segmentation loss $\mathcal{L}_{\text{seg}}$ and divergence-aware loss $\mathcal{L}_{\Delta}$. Specifically, they are defined as:

$$\mathcal{L}_{\Delta} = \sum_{x_i \in D} w(x_i) \cdot \mathcal{L}_{\text{seg}}(x_i), \tag{3}$$

where $\mathcal{L}_{\text{seg}}(x_i)$ denotes the standard segmentation loss (e.g., Dice Cross-entropy, Dice Focal) for the training sample x_i. The final loss function combines the original segmentation loss with the divergence-aware component:

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{seg}} + (1 - \alpha) \cdot \mathcal{L}_{\Delta}, \tag{4}$$

where $\alpha \in [0, 1]$ balances the trade-off between overall segmentation performance and subgroup fairness mitigation.

4 Experimental Setup

We evaluate our method using three widely used backbone 3D medical image segmentation architectures implemented within the MONAI framework [3]. These are: UNet [24], a baseline convolutional encoder–decoder architecture; SegResNet [13], a residual convolutional network optimized for volumetric medical data; and SwinUNETR [7], a hybrid transformer-based model that combines a Swin Transformer encoder with a UNet-like decoder. For each backbone architecture considered, we compare the base model (trained with a standard segmentation loss) to its fairness-enhanced counterpart using FairMedSeg.

4.1 Dataset and Data Preparation

We evaluate our method on the MAMA-MIA dataset [5], which comprises 1,506 T1-weighted DCE-MRI cases from female patients diagnosed with breast cancer,

divided into training and validation sets. Each case comprises one pre-contrast and up to five post-contrast dynamic phases, acquired across multiple clinical centers under varying imaging protocols. Tumor regions were manually segmented by a panel of 16 expert radiologists, ensuring high-quality ground truth annotations of the primary lesions. In addition to imaging data, the dataset includes rich clinical metadata. We focus on three key attributes, patient age, menopausal status, and breast density, to define clinically meaningful subgroups for fairness evaluation, as specified by the challenge guidelines.

We preprocess each DCE-MRI case by first reorienting the images to RAS (Right-Anterior-Superior) and resampling them to an isotropic resolution of 1.0 mm, using bilinear interpolation for the images and nearest-neighbor interpolation for the labels. We concatenate the pre-contrast and two post-contrast phases into a single 3-channel 3D volume. To ensure consistent spatial dimensions across the dataset, we standardize all inputs to a fixed shape of $320 \times 320 \times 128$ by applying cropping or padding as needed. To focus learning on the relevant anatomy, we apply a foreground cropping step based on non-zero voxel intensity. We apply a set of data augmentation strategies during training. Namely, random flipping along each spatial axis (x, y, z) with a probability of 0.5, intensity normalization restricted to non-zero voxels, and random intensity scaling and shifting with factors and offsets up to $\pm 10\%$.

For post-processing the model output, we refine each predicted segmentation by leveraging patient-specific anatomical priors. Specifically, we use metadata-provided bounding box coordinates that delineate the breast region for each case. These coordinates are used to apply a spatial mask that restricts predictions to the anatomically plausible area, effectively removing any segmented regions outside the breast tissue.

4.2 Evaluation Metrics

We follow the evaluation protocol from the MAMA-MIA challenge, relying on five metrics: the Dice Similarity Coefficient (DSC), the Normalized 95th Percentile Hausdorff Distance (NormHD), the Performance Score (PS), the Fairness Score (FS), and the Total Score (TS).

The DSC quantifies the spatial overlap between predicted and ground truth segmentation masks. The NormHD assesses boundary-level accuracy by computing the 95th percentile Hausdorff Distance, normalized by the image resolution and clipped to the range $[0, 1]$. The PS summarizes segmentation quality as the average of the DSC and the complement of the NormHD:

$$\mathrm{PS} = 0.5 \cdot (\mathrm{DSC} + (1 - \mathrm{NormHD})) \,.$$

The FS assesses fairness across variables. For each variable $v \in \mathcal{V}$, the challenge computes a disparity measure D_v based on subgroup-level performance variations, defining FS as:

$$\mathrm{FS} = \frac{1}{|\mathcal{V}|} \sum_{v=1}^{|\mathcal{V}|} (1 - D_v),$$

Table 1. FairMedSeg-enhanced and base models' performance on MAMA-MIA's validation and test sets. The highlighted row refers to the model submitted to the *MAMA-MIA* challenge (private test set). **Bold** denotes the best scores on the validation set.

Model	Method	DSC	NormHD	Perf. Score	Fairness Score	Total Score
UNet	Base (val)	0.7251	0.1380	0.7935	0.8334	0.8135
	FairMedSeg (val)	0.7691	0.0960	0.8364	**0.8580**	0.8472
	FairMedSeg (test)	0.66	0.16	0.75	0.86	0.8
SegResNet	Base (val)	0.7453	0.1226	0.8113	0.7414	0.7763
	FairMedSeg (val)	**0.7865**	**0.0806**	**0.8529**	0.8445	**0.8475**
SwinUNETR	Base (val)	0.6421	0.1970	0.7225	0.6480	0.6853
	FairMedSeg (val)	0.7478	0.1236	0.8121	0.8379	0.8250

with $|\mathcal{V}| = 3$, the three clinically relevant variables: *age*, *menopausal status*, and *breast density*. Finally, PS and FS are aggregated using equal weighting into the Total Score, i.e., $\text{TS} = (1 - \alpha) \cdot \text{PS} + \alpha \cdot \text{FS}$, with $\alpha = 0.5$.

5 Experimental results

Quantitative Results. Table 1 summarizes the results using the best-performing configuration for each model using the validation set within MAMA-MIA. We also report the result of the best model on the official MAMA-MIA private test set, as evaluated by the challenge organizers. In particular, we found experimentally that a regularization weight of $\alpha = 0.2$ and a $minsup = 0.001$ value for `DivExplorer` provided the most consistent improvements across architectures.

On the validation set, we observe substantial gains in TS (+14.0). On the official test set (Table 1, highlighted row) the UNet with FairMedSeg achieves a FS of 0.86. Overall, these results demonstrate that FairMedSeg achieves strong generalization across different patient groups while ensuring fairness is preserved or enhanced. These results confirm that fairness-aware training can reduce disparities among subgroups without compromising overall accuracy.

Subgroup Analysis with Heatmaps. To better understand subgroup-specific disparities, we visualize the average DSC across combinations of clinical metadata (Fig. 2) for the UNet model using FairMedSeg evaluated on the MAMA-MIA test set. The left plot shows DSC stratified by age group and breast density. The right one shows DSC by age group and menopausal status.

We observe substantial variability in segmentation performance across intersectional subgroups. For instance, patients aged 41-50 with breast density category `a`, exhibit the lowest DSC (0.158), while older patients (71+) generally achieve higher DSCs, especially in high-density categories. Similarly, the right heatmap highlights substantial differences in the case of post-menopausal patients aged 0–40, which achieves only 0.146 DSC, compared to 0.643 for their premenopausal counterparts. These findings confirm that subgroup disparities

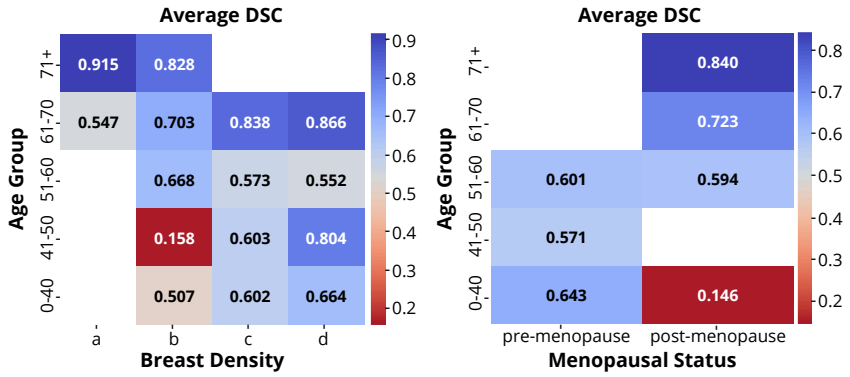

Fig. 2. Subgroup-level DSC Heatmaps. *Left:* Average DSC by age group and breast density (a–d). *Right:* Average DSC by age group and menopausal status.

are not only present but also intersectional in nature, motivating the need for intersectionality-aware training strategies such as FairMedSeg.

6 Discussion and Conclusions

We introduced FairMedSeg, a novel in-processing fairness mitigation strategy tailored for medical image segmentation. Our results demonstrate that FairMedSeg can significantly reduce subgroup disparities in breast tumor segmentation without compromising overall accuracy. By leveraging metadata-defined subgroups and dynamically adjusting training priorities based on divergence, our method effectively directs learning toward underperforming cohorts.

Automatic subgroup identification. A key strength of FairMedSeg lies in its ability to identify and mitigate fairness gaps without requiring prior subgroup definitions or manual intervention. This is particularly advantageous in clinical scenarios where vulnerable populations may be unknown or evolve over time. Moreover, the divergence-aware regularization can be flexibly integrated into a wide range of segmentation backbones, as demonstrated in our evaluation across convolutional and transformer-based architectures.

Intersectional subgroup disparities. The subgroup-level heatmap analysis (Fig. 2) reveals that performance disparities are not uniformly distributed but often emerge at specific intersections of clinical metadata. For example, patients aged 41-50 with breast density `a`, and postmenopausal patients under age 40, achieve the lowest DSC values. These gaps are often difficult to detect or mitigate using fairness strategies that treat attributes independently. FairMedSeg's capacity to autonomously discover and prioritize such subgroups underscores its practical value in real-world medical AI systems. While Figure 2 illustrates persistent disparities in certain intersectional subgroups, our method reduces these gaps compared to the baseline. However, achieving full equity across all

combinations remains a challenge due to limited sample sizes in rare subgroups.

Limitations and Future Work. We have identified some pending limitations. First, the subgroup divergence computation depends on validation performance, which may be unstable for low-sample subgroups. Moreover, our current sample weighting relies on the maximum subgroup divergence per sample, which may oversimplify intersectional subgroup dynamics. Incorporating richer aggregation mechanisms represents a promising direction for future refinement.. Second, we focused on three metadata variables (age, menopausal status, and breast density) due to the MAMA-MIA challenge rules. Extending our approach to additional metadata, e.g., ethnicity, hormonal receptor status, or treatment history, is a promising future direction. Finally, although we fixed the regularization weight ($\alpha = 0.2$), automatic tuning could further enhance generalization across datasets.

Acknowledgments. This work is partly supported by the ANR-BMBF TRAIN (ANR-22-FAI1-0003-02) and by the FAIR - Future Artificial Intelligence Research (Piano Nazionale Di Ripresa e Resilienza (PNRR) – Missione 4 Componente 2, Investimento 1.3 – D.D. 1555 11/10/2022, PE00000013) and the spoke "FutureHPC & BigData" of the ICSC - Centro Nazionale di Ricerca in High-Performance Computing, Big Data and Quantum Computing, both funded by the European Union - NextGenerationEU.

Disclosure of Interests. This manuscript reflects only the authors' views and opinions; neither the European Union nor the European Commission can be considered responsible for them. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abraham, N., Khan, N.M.: Adaptive focal loss for sparse object detection. arXiv preprint arXiv:1810.07842 (2021)
2. Alsulaimawi, A.: Enforcing fairness in deep medical imaging via adversarial learning. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 2557–2560. IEEE (2021)
3. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., et al.: Monai: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022)
4. Chen, R.J., Wang, J.J., Williamson, D.F., Chen, T.Y., Lipkova, J., Lu, M.Y., Sahai, S., Mahmood, F.: Algorithmic fairness in artificial intelligence for medicine and healthcare. Nature biomedical engineering **7**(6), 719–742 (2023)
5. Garrucho, L., Kushibar, K., Reidel, C.A., Joshi, S., Osuala, R., Tsirikoglou, A., Bobowicz, M., Del Riego, J., Catanese, A., Gwoździewicz, K., et al.: A large-scale multicenter breast cancer dce-mri benchmark dataset with expert segmentations. Scientific data **12**(1), 453 (2025)
6. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29** (2016)
7. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI brainlesion workshop. pp. 272–284. Springer (2021)

8. Huti, M., Lee, T., Sawyer, E., King, A.P.: An investigation into race bias in random forest models based on breast dce-mri derived radiomics features. In: Workshop on Clinical Image-Based Procedures. pp. 225–234. Springer (2023)
9. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems **33**(1), 1–33 (2012)
10. Koudounas, A., Pastor, E., de Alfaro, L., Baralis, E.: Mitigating subgroup disparities in speech models: A divergence-aware dual strategy. IEEE Transactions on Audio, Speech and Language Processing (2025)
11. Larrazabal, A.J., Nieto, N., Peterson, V., Milone, D.H., Ferrante, E.: Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proceedings of the National Academy of Sciences **117**(23), 12592–12594 (2020)
12. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM computing surveys (CSUR) **54**(6), 1–35 (2021)
13. Myronenko, A.: 3d mri brain tumor segmentation using autoencoder regularization. In: International MICCAI brainlesion workshop. pp. 311–320. Springer (2018)
14. Noriega-Campero, A., Bakker, M.A., Garcia-Bulle, B., Pentland, A.: Active fairness in algorithmic decision making. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. pp. 77–83 (2019)
15. Park, G.E., Kim, S.H., Nam, Y., Kang, J., Park, M., Kang, B.J.: 3d breast cancer segmentation in dce-mri using deep learning with weak annotation. Journal of Magnetic Resonance Imaging **59**(6), 2252–2262 (2024)
16. Pastor, E., De Alfaro, L., Baralis, E.: Looking for trouble: Analyzing classifier behavior via pattern divergence. In: Proceedings of the 2021 International Conference on Management of Data. pp. 1400–1412 (2021)
17. Pastor, E., Gavgavian, A., Baralis, E., de Alfaro, L.: How divergent is your data? Proceedings of the VLDB Endowment **14**(12), 2835–2838 (2021)
18. Paxton, K., Aslansefat, K., Thakker, D., Papadopoulos, Y.: Evaluating fairness and mitigating bias in machine learning: A novel technique using tensor data and bayesian regression. arXiv preprint arXiv:2506.11627 (2025)
19. Pérez-García, F., Sparks, R., Ouyang, J., et al.: Fairness in deep cardiac mri segmentation: assessing sex and racial bias in deep learning-based segmentation. Medical Image Analysis **80**, 102358 (2022)
20. Pfohl, S.R., Duan, T., Ding, D.Y., Shah, N.H.: Counterfactual reasoning for fair clinical risk prediction. In: Machine Learning for Healthcare Conference. pp. 325–358. PMLR (2019)
21. Puyol-Antón, E., Ruijsink, B., Piechnik, S.K., Neubauer, S., Petersen, S.E., Razavi, R., King, A.P.: 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: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24. pp. 413–423. Springer (2021)
22. Rayed, M.E., Islam, S.S., Niha, S.I., Jim, J.R., Kabir, M.M., Mridha, M.: Deep learning for medical image segmentation: State-of-the-art advancements and challenges. Informatics in medicine unlocked **47**, 101504 (2024)
23. Ricci Lara, M.A., Echeveste, R., Ferrante, E.: Addressing fairness in artificial intelligence for medical imaging. nature communications **13**(1), 4581 (2022)
24. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)

25. Sagawa*, S., Koh*, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks. In: International Conference on Learning Representations (2020), https://openreview.net/forum?id=ryxGuJrFvS
26. Sarhan, M.H., Navab, N., Eslami, A., Albarqouni, S.: Fairness by learning orthogonal disentangled representations. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16. pp. 746–761. Springer (2020)
27. Seyyed-Kalantari, L., Zhang, H., McDermott, M., Chen, I., Ghassemi, M.: Chexclusion: Fairness gaps in deep chest x-ray classifiers. Scientific Reports **11**(1), 10452 (2021)
28. Xu, Z., Li, J., Yao, Q., Li, H., Zhao, M., Zhou, S.K.: Addressing fairness issues in deep learning-based medical image analysis: a systematic review. npj Digital Medicine **7**(1), 286 (2024)
29. Yan, H., Zou, J., et al.: Forml: Fairness-optimized robust meta-learning. arXiv preprint arXiv:2202.01719 (2022)
30. Yeung, M., et al.: Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. arXiv preprint arXiv:2102.04525 (2022)
31. Zhao, B., Xiao, X., Gan, G., Zhang, B., Xia, S.T.: Maintaining discrimination and fairness in class incremental learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13208–13217 (2020)
32. Zhou, L., Zhang, Y., Zhang, J., Qian, X., Gong, C., Sun, K., Ding, Z., Wang, X., Li, Z., Liu, Z., et al.: Prototype learning guided hybrid network for breast tumor segmentation in dce-mri. IEEE Transactions on Medical Imaging (2024)