Evaluating Subgroup Fairness in Deep Learning Models for Breast Cancer Tumor Segmentation

Edson Holanda[12], **Lilian Bertonl**[1]

[1]ICT – Universidade Federal de São Paulo (Unifesp)
São José dos Campos – SP – Brazil

[2]Universidade Federal da Paraíba – Campina Grande - PB – Brazil

{edson.holanda,lberton}@unifesp.br

Abstract. *Machine learning is widely utilized across diverse fields, including healthcare. In this context, issues such as privacy and model fairness have been gaining increasing importance. It is not sufficient for a model to achieve good overall performance: it must be effective for all subgroups comprising the target population. This study investigates breast cancer tumor segmentation models using DCE-MRI from a fairness perspective. The experiments were conducted using the MAMA-MIA dataset and the nnU-Net framework, which represents the state-of-the-art in several medical image segmentation tasks. Model performances were compared through a fairness lens. In the investigated models, performance discrepancies were found across all evaluation metrics used.*

1. Introduction

Machine Learning (ML) algorithms, particularly those based on Deep Learning [LeCun et al. 2015], have been successfully applied across diverse fields of knowledge, with notable prominence in healthcare [Esteva et al. 2021, Faiyazuddin et al. 2025, Abdelwanis et al. 2025]. These algorithms have demonstrated a remarkable capability to assist in complex tasks, such as disease classification, anatomical segmentation, lesion delineation, and image reconstruction [Ricci Lara et al. 2022], thereby contributing to the development of clinical decision support tools and the achievement of more accurate diagnoses.

Decision support systems, particularly in medical contexts, are capable of handling high levels of complexity that often exceed human analytical capacity. However, these systems are not immune to biases. Consequently, researchers must remain attentive not only to the overall accuracy of the system but also to how it responds to data from specific subgroups within the dataset [Barocas and Selbst 2016]. In this context, relevant ethical questions have emerged, among which algorithmic fairness stands out [Pessach and Shmueli 2020]. There is a common tendency to attribute greater objectivity and impartiality to these systems compared to human decision-making [Maybury 1990]; however, recent studies demonstrate that this premise does not always hold true, revealing that algorithms can reproduce and even amplify biases present in the training data [Kleinberg et al. 2017, Rabonato and Berton 2025]. Certain demographic groups, defined by sensitive attributes such as sex, race, or ethnicity, can be systematically disadvantaged by these systems, resulting in significantly lower performance outcomes

compared to other groups [Buolamwini and Gebru 2018], which highlights the presence of algorithmic biases with potential discriminatory impacts. Pessach and Shmueli [Pessach and Shmueli 2020] identify research on fairness in Artificial Intelligence (AI) systems as an emerging and promising domain.

Among the various applications of ML models, tumor segmentation stands out as highly critical. Among women, breast cancer is the most incident type of cancer in Brazil, when excluding non-melanoma skin tumors. The National Cancer Institute (INCA) estimates that for the 2026-2028 triennium, Brazil will face 78,610 new cases annually [Instituto Nacional de Câncer (INCA) 2026]. Furthermore, cancer mortality rates in Brazil have been steadily approaching those of cardiovascular diseases. When a patient is diagnosed early, the likelihood of a cure is high. A study published in 2019 [Makdissi et al. 2019] illustrates the evolution of survival rates. The data indicate that patient survival associated with tumor sizes T1 and T2 (early stages) rose from 89.7% (2000-2004) to 95% (2010-2012).

In Brazil, public screening policies for patient identification recommend performing a mammogram every 2 years for women aged 50 to 69 [Costa et al. 2019]. Mammography offers the advantage of considerably lower costs compared to magnetic resonance imaging (MRI). Luczynska et al. [Luczynska et al. 2022], however, note that MRI is a widely used examination in breast cancer diagnosis, both for identifying the cancer type and evaluating its recurrence. Bitencourt and Chojniak [Bitencourt and Chojniak 2021] also emphasize that MRI is indicated to resolve inconsistencies resulting from inconclusive mammography and ultrasonography exams. The rising number of breast cancer cases, the importance of early diagnosis—which in turn can increase patient survival rates—and the high complexity of the 3D medical image interpretation task [van Montfort et al. 2021] render ML models an important tool to assist healthcare professionals.

In this scenario, studies on algorithmic fairness and equity have been gaining substantial relevance. Verma and Rubin [Verma and Rubin 2018] show that more than 20 definitions of fairness exist in the literature. Algorithmic fairness can be analyzed under two distinct aspects: group fairness and individual fairness ([Dwork et al. 2012]). From an individual fairness perspective, a ML model is considered fair if, and only if, similar individuals obtain similar outcomes. In the present work, the focus is placed on group fairness. Regarding group fairness, when evaluating two distinct groups, the subgroup achieving higher accuracy is referred to as the privileged group, whereas the one with lower accuracy is termed the underprivileged group. It can be asserted that group fairness is achieved when both privileged and underprivileged groups receive equitable treatment from the model.

Measuring fairness in ML models constitutes a challenge in itself, since there is no universally accepted definition of what characterizes a fair model [Verma et al. 2021]. Several metrics have been proposed in the literature to quantify and evaluate a model's level of equity, each grounded in different conceptions of justice and tailored to distinct application contexts [Mehrabi et al. 2021]. Given this plurality, the selection of the most appropriate metrics must be guided by the specific application domain and the vulnerable groups involved [Barocas and Selbst 2016].

Despite these advancements, an important research gap remains: while fairness interventions are heavily explored in classification tasks, extending these concepts to complex, dense regression spaces like 3D semantic image segmentation is an active and under-researched challenge. Traditional group-fairness metrics do not directly translate to pixel-level predictions, and there is a critical shortage of comprehensive multi-axis subgroup evaluations in diverse clinical backgrounds. Furthermore, in deeply heterogeneous populations such as Brazil's, disparities in scanner hardware acquisition, distinct patient biotypes, and ethnic demographics often compound hidden algorithmic biases, potentially leading to disparate health outcomes across sub-populations.

To bridge this gap, this work systematically investigates breast cancer tumor segmentation models developed using the state-of-the-art nnU-Net framework on 3D DCE-MRI scans from the MAMA-MIA dataset [Garrucho et al. 2025] through a rigorous multi-axis group fairness lens. We analyze the behavior of these models across distinct demographic, clinical, and technical strata to expose hidden performance disparities. Specifically, our main contributions are as follows:

- **Multi-Axis Fairness Benchmarking:** We provide a comprehensive evaluation of semantic segmentation fairness across three distinct and critical context dimensions: scanner manufacturer (c_1) to capture technical acquisition bias, patient ethnicity (c_2) to capture demographic bias, and BMI/weight groups (c_3) to evaluate morphological and anatomical variation bias.
- **Quantification of the Subgroup Performance Gap (SPG):** We establish empirical evidence that state-of-the-art segmentation frameworks, despite showing high overall cumulative performance, suffer from significant degradation and performance gaps when evaluated on minority or clinically underrepresented subgroups.
- **State-of-the-Art Baseline Optimization:** We present an optimized baseline model using the nnU-Net version 2 framework that outperforms previously reported results on the MAMA-MIA dataset, setting a new rigorous performance standard against which future fair-learning mitigation strategies can be benchmarked.

2. Related Work

The intersection of AI and clinical workflow monitoring demands models that are both operationally adaptive and ethnically equitable [Chen et al. 2023]. This section contextually positions our study by reviewing the recent literature across algorithmic fairness in medical computer vision.

Recent years have witnessed growing attention toward identifying and mitigating biases embedded within medical ML systems [Pessach and Shmueli 2020, Jiang and Nachum 2020]. Biases frequently manifest in clinical models due to training data imbalances, disparities in historical data-generating workflows, biological heterogeneity across subgroups (e.g., age, sex, race), or distinct hardware specifications among device manufacturers [Chen et al. 2023, Mehta et al. 2024]. For instance, landmark investigations by Seyyed-Kalantari et al. documented significant underdiagnosis rates and performance drops for underrepresented racial groups when deep classification networks were applied to chest X-rays. Similarly, Larrazabal et al. demonstrated how gender-imbalanced training datasets severely compromise the diagnosis of thoracic diseases in the underrepresented sex.

While fairness interventions are heavily researched in classification domains, extending these concepts to complex, dense regression spaces like semantic image segmentation is an active challenge [Tian et al. 2024]. Traditional group-fairness metrics such as Demographic Parity or Equalized Odds do not directly translate to multi-dimensional, pixel-level predictions [Tian et al. 2024]. To bridge this gap, Tian et al. introduced *Harvard-FairSeg*, the first large-scale medical image segmentation dataset explicitly curated for fairness benchmarking, alongside a Fair Error-Bound Scaling (FEBS) optimization framework to enforce metric equity during boundary parsing [Tian et al. 2024]. More recently, Oh et al. approached distribution-aware fairness from a control-theoretic standpoint, introducing a Distribution-aware Mixture of Experts (dMoE) network to handle anatomical and hardware acquisition shifts across diverse demographic subgroups [Oh et al. 2025]. Furthermore, Mehta et al. established that bias in deep medical vision models extends beyond raw performance drop, heavily skewing the network's predictive confidence and uncertainty estimates across sensitive subpopulations [Mehta et al. 2024].

3. Proposed Method

The proposed methodology consists of the following steps: model development (utilizing nnU-Net), inference and prediction using the trained models, metric calculation, and performance analysis.

3.1. Notation

The MAMA-MIA dataset was analyzed separately under different demographic and clinical contexts. Let the set of contexts be denoted by: $C = \{ c_i \mid i = 1, 2, \ldots, n \}$. Each context c_i, in turn, is composed of a set of subgroups g_j^i as represented as: $G_i = \{ g_j^i \mid j = 1, 2, \ldots, m \}$, where g_j^i represents subgroup j within context i. When the context is clear, we will use simply g_j to denote subgroup j. All analyzed contexts and their respective subgroups are summarized in Table 1.

Table 1. Analyzed contexts and subgroups within the MAMA-MIA dataset

Context	Subgroups	No. of Scans
Manufacturer (c_1)	GE (g_1^1)	706
	Philips (g_2^1)	129
	Siemens (g_3^1)	287
Ethnicity (c_2)	Caucasian (g_1^2)	860
	African-American (g_2^2)	160
	Asian (g_3^2)	73
BMI / Weight (c_3)	Normal (g_1^3)	505
	Mixed Obesity (g_2^3)	54
	Class 3 Obesity (g_3^3)	35

Within a given context i,, each subgroup g_j^i is associated with a subset pair (Tr_j, Ts_j), where Tr_j represents the training data for that subgroup and Ts_j represents its corresponding test data. It is worth noting that the nnU-Net framework automatically handles the internal split of Tr_j into training and validation sets. We denote by $m_{g_j^i}$ a model trained exclusively on data Tr_j^i, simplifying the notation to m_{g_j} when the context i is clear. Furthermore, $met(m_{g_h}, g_j)$ represents the evaluation metric computed on subgroup g_j when predictions are generated using model m_{g_h}.

For each context, multiple models were developed. First, a reference model was trained using all available images, disregarding the underlying subgroup distributions. Subsequently, for each context, a model was trained strictly on the majority subgroup (the one with the largest sample size) and then evaluated to generate predictions for the test sets of each minority subgroup. These outcomes were subsequently analyzed.

3.2. Dataset

The MAMA-MIA dataset, introduced by Garrucho et al. [Garrucho et al. 2025], comprises 1,506 cases of patients diagnosed with breast cancer. For each case, five three-dimensional dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) scans were provided, consisting of one pre-contrast and four post-contrast images. Furthermore, the dataset offers tumor segmentations performed by experts alongside clinical patient data, establishing a comprehensive resource for the development and evaluation of ML models applied to breast cancer diagnosis. In addition to the annotated images, MAMA-MIA provides the pre-trained nnU-Net model weights, as well as a clinical dataset comprising 21 clinical variables, six demographic variables — including ethnicity and weight — and 22 imaging parameters, such as acquisition time and spatial resolution. The dataset is hosted on Synapse[1].

3.3. The Framework

The nnU-Net framework was developed based on the U-Net architecture [Ronneberger et al. 2015], a Convolutional Neural Network (CNN) widely recognized in the field of biomedical image segmentation. U-Net stood out by achieving state-of-the-art results in several medical segmentation challenges [Ronneberger et al. 2015, Isensee et al. 2021], and it is considered the benchmark architecture for medical image segmentation tasks.

The nnU-Net features AutoML capabilities, being able to adaptively self-configure based on the dataset properties and the available hardware setups [Isensee et al. 2021], thereby eliminating the trial-and-error process typically required in the development of domain-specific solutions — a task acknowledged as non-trivial and computationally expensive. This self-configuration capability makes nnU-Net particularly well-suited for scenarios with a scarcity of annotated data, as is the case with DCE-MRI scans for breast cancer segmentation. Furthermore, the performance of nnU-Net outpaces most other deep learning pipelines across 19 international public competitions, being considered the state-of-the-art for 49 medical segmentation tasks [Isensee et al. 2021].

In the present study, a 3D U-Net architecture dynamically configured by nnU-Net was employed. The network consists of six encoder stages and five decoder stages, interconnected via skip connections, which constitute the core characteristic of the U-Net architecture [Ronneberger et al. 2015]. The number of feature channels scales progressively from 32 to 320 along the encoder pathway. Each stage is composed of two convolutional layers with $3 \times 3 \times 3$ kernels, followed by Instance Normalization and LeakyReLU activation functions. Downsampling is executed via strided convolutions, including anisotropic strides to account for voxel spacing anomalies. The network was trained utilizing 3D patches with a target size of $96 \times 160 \times 160$ voxels sampled from

[1]https://doi.org/10.7303/syn60868042

the input volumes. All configurations generated by nnU-Net are compiled within the `nnUNetPlans.json` file, allowing for specific adjustments based on experimental needs [Isensee et al. 2021].

3.4. Hardware and Software

Experiments were conducted utilizing the Santos Dumont cluster[2], a high-performance supercomputer managed by the National Laboratory for Scientific Computing (LNCC). Job scheduling is based on resource demand: larger requests incur longer waits. Because training time was not studied, resources were minimized to the baseline needed for execution. Consequently, a single computational node was configured with the following specifications: two Intel Xeon Cascade Lake Gold 6252 processors (24 cores per processor), 24 GB of RAM, and one NVIDIA Volta V100 GPU. Regarding software infrastructure, the environment integrated Slurm 20.11.8 (Linux), Python 3.9.18, nnU-Net version 2.5.2 [Isensee et al. 2021], and the `seg-metrics` library version 1.2.8 [Jia et al. 2024].

3.5. Preprocessing

Two primary preprocessing operations were automatically executed by nnU-Net in its default 3D pipeline configuration [Isensee et al. 2021]: z-score intensity normalization of the scans, computed from the mean and standard deviation of all acquisition phases (pre- and post-contrast), and voxel spacing standardization to a uniform isotropic resolution of $1 \times 1 \times 1$ mm^3. The nnU-Net incorporates a data augmentation strategy during training, wherein the post-contrast phase images are augmented by a subtraction image, derived from the voxel-wise difference between the pre-contrast volume and the first post-contrast acquisition. The data were partitioned into 80% for training/validation and 20% for independent testing.

3.6. Training

During the first stage of the experiment, model training was carried out using the default configuration of nnU-Net v2 [Isensee et al. 2021]: execution over 1,000 epochs, optimized via Stochastic Gradient Descent (SGD) with an initial learning rate of 0.01, a momentum of 0.99, and a weight decay factor of 3×10^{-5}. The loss function was defined as a combination of the Dice Similarity Coefficient (DSC) loss and cross-entropy loss. All default hyperparameters are stored and can be customized within the `nnUNetTrainer.py` class module.

As a reference baseline scenario, model m_0 was trained using the complete compiled dataset. The obtained performance was evaluated against the results reported by the MAMA-MIA authors in their public repository [Garrucho et al. 2025] (see Table 2).

3.7. Prediction and Inference

For each context c_i, following the training of the primary model m_{g_1}, inference was performed across all remaining subgroups g using the independent test set Ts_l. Figure 1 illustrates the cross-sectional slices of a segmentation generated by a model trained on the full dataset for a patient belonging to the African-American ethnic subgroup. Figure 2 displays the 3D rendering of the same model-generated segmentation volume. This prediction pipeline provided the inputs required to compute the target evaluation metrics for each developed model.

[2]https://sdumont.lncc.br

Figure 1. Examples of segmentation slices

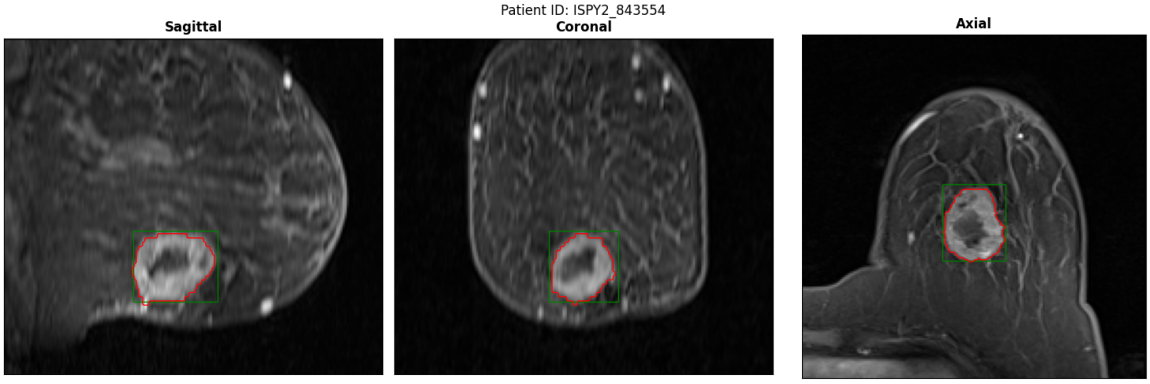
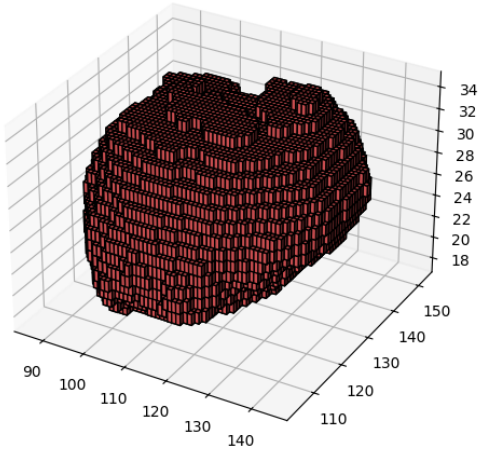


Figure 2. 3D rendering of the segmentation volume



3.8. Evaluation Metrics

The *DSC*, *IoU*, *HD95*, and *MSD* metrics were utilized to evaluate the models. The calculation of all metrics was performed using the Python-implemented `seg-metrics` library[3] [Jia et al. 2024]. These four metrics were chosen because they are among the most widely adopted for evaluating the performance of image segmentation models, and they were also utilized by the team that developed the MAMA-MIA dataset. Consequently, this allows for a direct comparison with the benchmark model for breast cancer image segmentation. We also utilize the Subgroup Performance Gap (SPG) [Parikh et al. 2026] to quantify the discrepancy in model performance between the privileged and underprivileged groups ($\Delta = g_c - g_b$) across each of the four selected metrics.

4. Results and Discussion

To verify that the framework was correctly installed and configured, an initial training run was executed utilizing all available data in the MAMA-MIA dataset. Table 2 presents the comparison of the *DSC*, *IoU*, *HD95*, and *MSD* metrics. The results obtained at this baseline stage, where the model was trained and tested on the complete heterogeneous dataset,

[3] `https://pypi.org/project/seg-metrics/`

constitute the reference baseline scenario (m_0) for this work, against which all subsequent experiments are compared. In the subsequent contexts, training and testing were executed on isolated subgroups to evaluate the impact of continuous learning strategies on model performance. In the following tables, upward arrows (↑) indicate metrics for which higher values are better, whereas downward arrows (↓) indicate metrics for which lower values are better.

Table 2. Preliminary results utilizing the full dataset

Context	DSC ↑	IoU ↑	HD95 ↓	MSD ↓
Garrucho2025 (Mean)	0.762	0.654	37.41	11.08
m_0	0.802	0.702	32.44	8.15

The results achieved by our developed model were slightly superior to those reported by the original MAMA-MIA project [Garrucho et al. 2025], as detailed in Table 2. The model achieved a *DSC* of 0.802, compared with 0.762 reported in the original study and an *IoU* of 0.702 compared to 0.654. Superior performance was also observed in the boundary distance metrics, with *HD95* and *MSD* values reaching 32.44 and 8.15, respectively, compared to 37.41 and 11.08 in the original study. The use of nnU-Net v2, likely contributed to the higher performance observed, as this iteration introduced key structural enhancements and optimization pipeline improvements over its predecessor [Isensee et al. 2021].

To enable model analysis from a fairness perspective, an experimental protocol was designed wherein the data were stratified according to different demographic and clinical contexts (see Table 1). The objective of this approach was to evaluate the impact on performance metrics when a model trained with data from a specific group is applied to a distinct group, thereby simulating real-world scenarios of data variability. Metadata on manufacturer, ethnicity, and weight were omitted in training.

4.1. Context c_1: Manufacturer

In the first context of the experiment, cases were stratified based on the manufacturer of the equipment used for image acquisition, resulting in three groups: GE equipment (g_1), Philips equipment (g_2) and Siemens equipment (g_3)(see Table 1). The model m_{g_1} was trained on the GE group and evaluated separately on the Philips and Siemens groups. The results of the classic metrics for this scenario are presented in Table 3 and compared against the baseline scenario.

The results demonstrate that when the model trained on the GE group is evaluated on the Siemens group, all metrics exhibit lower performance compared to the baseline scenario. When evaluated on the Philips group, a degradation is observed in the *DSC* and *IoU* metrics, whereas an improvement is found in the *HD95* and *MSD* metrics. This pattern suggests that differences in acquisition protocols across distinct manufacturers introduce variability into the image features, negatively impacting the model's generalization capability.

4.2. Context c_2: Ethnicity

In the second scenario analyzed, cases were stratified into three ethnic groups: Caucasian (g_1), African-American (g_2), and Asian (g_3) (see Table 1). The model was trained on Cau-

Table 3. Baseline case and subgroups by manufacturer

Context: c_1	**DSC** ↑	**IoU** ↑	**HD95** ↓	**MSD** ↓
$met(m_{g1}, g_2)$	0.798	0.696	24.45	5.30
$met(m_{g1}, g_3)$	0.782	0.675	46.79	11.29
$SPG(g2, g_3)$	0.016	0.021	-22.34	-5.99

casian group and evaluated separately on African-American and Asian groups. Table 4 presents the results of the classic metrics for the analyzed groups.

The results demonstrate that when the model trained on the Caucasian group is evaluated on the African-American group, all metrics show lower performance compared to the baseline scenario. When evaluated on the Asian group, a degradation is observed in the *DSC* and *IoU* metrics, along with an improvement in the *HD95* and *MSD* metrics. It is worth noting, however, that the Asian group is considerably smaller in terms of sample size — 73 cases compared to 160 in the African-American group —, which may influence the stability and reliability of the results obtained for this specific group.

Table 4. Baseline case and subgroups by ethnicity

Context: c_2	**Dice** ↑	**IoU** ↑	**HD95** ↓	**MSD** ↓
$met(m_{g1}, g_2)$	0.788	0.688	35.10	9.91
$met(m_{g1}, g_3)$	0.784	0.676	28.76	5.90
$SPG(g2, g_3)$	0.004	0.012	-6.34	-4.01

4.3. Context c_3: Weight / BMI

In the final context, cases were stratified into three groups according to Body Mass Index (BMI) classification: patients with normal weight, patients with overweight or mild obesity (denoted as the mixed obesity group), and patients with class 3 obesity (see Table 1). The model was trained on normal-weight group and evaluated separately on the mixed obesity group and the class 3 obesity groups. Table 5 presents the results of the classic metrics for the analyzed groups.

The results demonstrate that when the model trained on the normal-weight group is evaluated on the remaining groups, all metrics exhibit lower performance compared to the baseline scenario. This degradation is particularly pronounced in the *HD95* and *MSD* metrics, suggesting that morphological variability associated with obesity poses severe challenges to model generalization when dealing with cohorts that display distinct anatomical characteristics from those available in the training set.

Table 5. Baseline case and subgroups by weight/BMI

Context: c_3	**DSC** ↑	**IoU** ↑	**HD95** ↓	**MSD** ↓
$met(m_{g_1}, g_2)$	0.727	0.617	71.18	21.29
$met(m_{g_1}, g_3)$	0.719	0.602	71.97	24.19
$SPG(g2, g_3)$	0.004	0.012	-6.34	-4.01

5. Conclusion

The concept of fairness cannot be considered of minor interest by researchers in the field. There is a critical need to achieve a balance within models between metrics that eval-

uate the entire dataset distribution and those that account for the individual subgroups composing that population.

In the conducted experiments, it was observed that the outcomes across the distinct population subgroups are not always equitable, as a significant disparity persists between the performance of the privileged and underprivileged groups. It is crucial that these aspects are carefully taken into account during model evaluation. Future work may expand this fairness evaluation to other imaging datasets, as well as test different ML architectures and architectural variants.

Acknowledgments

The authors would like to thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for supporting the development of this study.

Declaration of Generative AI and AI-Assisted Technologies

The authors used generative AI technologies to polish the phrasing, enhance grammatical readability, and translate draft text into structured English prose. After using these services, the authors thoroughly reviewed, edited, and verified all generated content, and take full responsibility for the final scientific accuracy.

References

Abdelwanis, M., Moawad, K., Mohammed, S., Hummieda, A., Syed, S., Maalouf, M., and Jelinek, H. F. (2025). Sequential classification approach for enhancing the assessment of cardiac autonomic neuropathy. *Computers in Biology and Medicine*, 190:109999.

Barocas, S. and Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review*, 104(3):671–732.

Bitencourt, A. G. V. and Chojniak, R. (2021). Indications for breast magnetic resonance imaging in Brazil: Past, present, and future. *Radiologia Brasileira*, 54:206–206.

Buolamwini, J. and Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR.

Chen, R. J., Wang, J. J., Williamson, D. F. K., Chen, T. Y., Lipkova, J., Lu, M. Y., Sahai, S., and Mahmood, F. (2023). Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature Biomedical Engineering*, 7(6):719–742.

Costa, L. D. L. N., Sardinha, A. H. d. L., Verzaro, P. M., Lisbôa, L. L. C., and Batista, R. F. L. (2019). Mortalidade por Câncer de Mama e Condições de Desenvolvimento Humano no Brasil. *Revista Brasileira de Cancerologia*, 65(1):e–12050.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pages 214–226. Association for Computing Machinery.

Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., Liu, Y., Topol, E., Dean, J., and Socher, R. (2021). Deep learning-enabled medical computer vision. *npj Digital Medicine*, 4(1):1–9.

Faiyazuddin, Md., Rahman, S. J. Q., Anand, G., Siddiqui, R. K., Mehta, R., Khatib, M. N., Gaidhane, S., Zahiruddin, Q. S., Hussain, A., and Sah, R. (2025). The Impact of Artificial Intelligence on Healthcare: A Comprehensive Review of Advancements in Diagnostics, Treatment, and Operational Efficiency. *Health Science Reports*, 8(1):e70312.

Garrucho, L., Kushibar, K., Reidel, C.-A., Joshi, S., Osuala, R., Tsirikoglou, A., Bobowicz, M., Riego, J. d., Catanese, A., Gwoździewicz, K., Cosaka, M.-L., Abo-Elhoda, P. M., Tantawy, S. W., Sakrana, S. S., Shawky-Abdelfatah, N. O., Salem, A. M. A., Kozana, A., Divjak, E., Ivanac, G., Nikiforaki, K., Klontzas, M. E., García-Dosdá, R., Gulsun-Akpinar, M., Lafcı, O., Mann, R., Martín-Isla, C., Prior, F., Marias, K., Starmans, M. P. A., Strand, F., Díaz, O., Igual, L., and Lekadir, K. (2025). A large-scale multicenter breast cancer dce-mri benchmark dataset with expert segmentations. *Scientific Data*, 12(1):453.

Instituto Nacional de Câncer (INCA), B. (2026). *Estimativa 2026: incidência de câncer no Brasil*. INCA.

Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., and Maier-Hein, K. H. (2021). nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211.

Jia, J., Staring, M., and Stoel, B. C. (2024). Seg-metrics: A python package to compute segmentation metrics. *arXiv preprint arXiv:2403.07884*.

Jiang, H. and Nachum, O. (2020). Identifying and Correcting Label Bias in Machine Learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pages 702–712. PMLR.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In Papadimitriou, C. H., editor, *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 43:1–43:23.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Luczynska, E., Piegza, T., Szpor, J., Heinze, S., Popiela, T., Kargol, J., and Rudnicki, W. (2022). Contrast-Enhanced Mammography (CEM) Capability to Distinguish Molecular Breast Cancer Subtypes. *Biomedicines*, 10(10):2384.

Makdissi, F. B., Leite, F. P. M., Peres, S. V., e Silva, D. R. M., de Oliveira, M. M., Lopez, R. V. M., Sanches, S. M., Gondim, G. R. M., Iyeyasu, H., Calsavara, V. F., and Curado, M. P. (2019). BREAST CANCER SURVIVAL IN A BRAZILIAN CANCER CENTER: A COHORT STUDY OF 5,095 PATIENTS. *Mastology*, 29(1):37–46.

Maybury, M. (1990). The mind matters: Artificial intelligence and its societal implications. *IEEE Technology and Society Magazine*, 9(2):7–15.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):115:1–115:35.

Mehta, R., Shui, C., and Arbel, T. (2024). Evaluating the Fairness of Deep Learning Uncertainty Estimates in Medical Image Analysis. In *Medical Imaging with Deep Learning*, pages 1453–1492. PMLR.

Oh, Y., Jin, P., Park, S., Kim, S., Yoon, S., Kim, K., Kim, J. S., Li, X., and Li, Q. (2025). Distribution-aware Fairness Learning in Medical Image Segmentation From A Control-Theoretic Perspective. https://arxiv.org/abs/2502.00619v2.

Parikh, A., Frank, S., Das, S., and Feragen, A. (2026). Towards fairness under label bias in image segmentation: Impact, measurement and mitigation. *arXiv preprint arXiv:2605.06891*.

Pessach, D. and Shmueli, E. (2020). Algorithmic fairness. *arXiv preprint arXiv:2001.09784*.

Rabonato, R. T. and Berton, L. (2025). A systematic review of fairness in machine learning. *AI and Ethics*, 5(3):1943–1954.

Ricci Lara, M. A., Echeveste, R., and Ferrante, E. (2022). Addressing fairness in artificial intelligence for medical imaging. *Nature Communications*, 13(1):4581.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham. Springer International Publishing.

Tian, Y., Shi, M., Luo, Y., Kouhana, A., Elze, T., and Wang, M. (2024). FairSeg: A Large-Scale Medical Image Segmentation Dataset for Fairness Learning Using Segment Anything Model with Fair Error-Bound Scaling. *International Conference on Learning Representations*, 2024:44056–44070.

van Montfort, D., Kok, E., Vincken, K., van der Schaaf, M., van der Gijp, A., Ravesloot, C., and Rutgers, D. (2021). Expertise development in volumetric image interpretation of radiology residents: What do longitudinal scroll data reveal? *Advances in Health Sciences Education*, 26(2):437–466.

Verma, S. and Rubin, J. (2018). Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, pages 1–7, New York, NY, USA. Association for Computing Machinery.

Verma, V. K., Liang, K. J., Mehta, N., Rai, P., and Carin, L. (2021). Efficient Feature Transformations for Discriminative and Generative Continual Learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13860–13870.