

INVESTIGATING LABEL BIAS AND REPRESENTATIONAL SOURCES OF AGE-RELATED DISPARITIES IN MEDICAL SEGMENTATION

Aditya Parikh Sneha Das Aasa Feragen

DTU Compute, Technical University of Denmark

ABSTRACT

Algorithmic bias in medical imaging can perpetuate health disparities, yet its causes remain poorly understood in segmentation tasks. While fairness has been extensively studied in classification, segmentation remains underexplored despite its clinical importance. In breast cancer segmentation, models exhibit significant performance disparities against younger patients, commonly attributed to physiological differences in breast density. We audit the MAMA-MIA dataset, establishing a quantitative baseline of age-related bias in its automated labels, and reveal a critical Biased Ruler effect where systematically flawed labels for validation misrepresent a model’s actual bias. However, whether this bias originates from lower-quality annotations (label bias) or from fundamentally more challenging image characteristics remains unclear. Through controlled experiments, we systematically refute hypotheses that the bias stems from label quality sensitivity or quantitative case difficulty imbalance. Balancing training data by difficulty fails to mitigate the disparity, revealing that younger patient cases are intrinsically harder to learn. We provide direct evidence that systemic bias is learned and amplified when training on biased, machine-generated labels, a critical finding for automated annotation pipelines. This work introduces a systematic framework for diagnosing algorithmic bias in medical segmentation and demonstrates that achieving fairness requires addressing qualitative distributional differences rather than merely balancing case counts.¹

Index Terms— algorithmic fairness, label bias, segmentation, breast MRI

1 Introduction

The integration of deep learning in medical imaging has shown potential for automating critical tasks like tumor segmentation, yet it carries substantial risk of inheriting and amplifying biases present in clinical data [1, 2]. Algorithmic bias in healthcare is concerning and can perpetuate existing demographic disparities [3]. In breast cancer diagnostics, a well-documented challenge is that segmentation performance

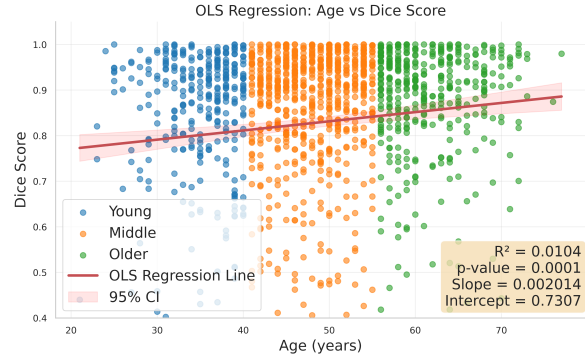


Fig. 1. Inherent age-related bias in the Silver-Standard automated labels. An OLS regression reveals a positive correlation between patient age and segmentation performance (Dice Score), establishing a quantitative baseline of disparity.

is often lower for younger patients [4, 5]. This is commonly attributed to their higher breast density [6], which can obscure tumor margins and complicate segmentation for *both* radiologists and automated systems. Meanwhile, fairness – the principle that a model should not systematically disadvantage certain patient subgroups – remains underexplored in segmentation, a task with direct implications for treatment planning and clinical decision-making.

While the correlation between age and performance is established, its underlying reasons remain unclear. While higher segmentation difficulty due to higher breast density is a previously cited possible reason, there are other potential contributing factors. Performance disparities often come from **representational bias**, arising from differences in case distribution or intrinsic imaging characteristics across age groups [7, 4]. This includes not only class imbalance and prevalence of challenging cases but also their qualitative nature. Factors like non-mass enhancement, irregular tumor morphology, or variable presentation patterns that may be more prevalent in younger women make them systematically harder for the model to learn [8, 9]. Another potential reason is **label bias** – where the ground truth annotations for younger patients are systematically less accurate due to the inherent difficulty of manual segmentation. If the annotations have lower quality in younger subjects, then our ability to measure

¹Code for the experimental framework and evaluation is made available at [https://\(tobereleaseduponacceptance\)](https://(tobereleaseduponacceptance))

performance is also lower in the very same subjects.

Differentiating between the causes and sources of bias is critical for developing effective fairness interventions. To date, label bias has remained an unrecognised and unquantified source of bias in image segmentation. As segmentation benchmarks and even datasets used in real systems are often created using semi- or fully-automatic tools (e.g., ISIC Challenge datasets [10], FreeSurfer for neuroimaging [11], nnU-Net-based annotations [12, 13, 14]), we have every reason to assume that segmentation labels can have systematic biases. It is known from classification tasks that this gives a *biased ruler* effect, where we are unable to effectively measure and mitigate bias. In this paper, we therefore let machine-generated labels from a pre-existing system serve as a methodological probe to measure a model’s *apparent performance* compared to its *real* performance. This allows us to quantify the impact of using a flawed, real-world benchmark.

In this paper, we thus present a systematic analysis designed to disentangle the potential sources of bias in image segmentation through a series of controlled experiments on the MAMA-MIA dataset [14], a large-scale dataset of breast cancer MRI images. Our contributions are: (i) A first comprehensive **fairness audit of the MAMA-MIA dataset**, establishing a quantitative baseline of the age-related bias present in its automated labels; (ii) To the best of our knowledge, the **first study of label bias and its effect on bias audit in image segmentation**; (iii) A framework of controlled experiments designed to **isolate the effects of label bias from representational bias**; (iv) Direct quantitative evidence of bias amplification, demonstrating how **systemic bias is learned and propagated** through machine learning pipelines.

2 Methodology

2.1 Dataset

The MAMA-MIA dataset [14] is a large, publicly available multi-center breast cancer benchmark of dynamic contrast-enhanced magnetic resonance images (DCE-MRI). It comprises 1,506 unique patient cases, each including volumetric imaging data, rich demographic metadata, and a pair of segmentation masks: An expert-annotated mask and a mask automatically generated by a standard nnU-Net framework [12] trained on an external dataset, including resampling to a target isotropic voxel spacing of $[1.0 \times 1.0 \times 1.0\text{mm}]$ and Z-score intensity normalization. For our experiments, we utilize the second T1-weighted, post-contrast phase as 3D volumetric input for all models. The patient age was stratified into three distinct age cohorts: *Young* (≤ 40 , $n = 349$), *Middle* ($40-55$, $n = 754$), and *Older* (≥ 55 , $n = 400$), based on established clinical relevance in breast cancer diagnostics [15].

Evaluation Benchmarks: We use both types of masks as validation labels: The term **Gold-Standard Labels** is used to refer to expert-annotated labels, created by 16 expert radiolo-

| Tier | Criteria | Difficulty |
|---------------------------------------------|----------------------------------------------------------------------|------------|
| <i>Tier 1</i> (Unambiguously Good) | Both experts rate “Good” AND (Dice ≥ 0.80 & HD95 ≤ 10) | Easy |
| <i>Tier 1.5</i> (Expert-Metric Mismatch) | Both experts rate “Good” BUT (Dice < 0.80 OR HD95 > 10) | |
| <i>Tier 2</i> (Clinically Acceptable) | Any case with “Acceptable” ratings OR expert disagreement | Hard |
| <i>Tier 3</i> (Unambiguously Poor) | Both experts rate “Poor” | |

Table 1: Definition of Quality Tiers and Difficulty Categories

gists. As these are likely the least affected by image quality biases, we regard these as ground truth, used as a definitive benchmark for measuring a model’s *true performance*. **Silver-Standard Labels** are the automated nnU-Net masks. *Note* that these resemble the semi- or fully-automatic annotations often found in real-world segmentation datasets, and are therefore a realistic assumption of what segmentation labels frequently look like. The silver-standard labels include dual-expert qualitative ratings (Good, Acceptable, Poor, Missed) assessing their visual quality.

Case Difficulty Stratification: To analyze label quality and case difficulty, we stratify cases into four tiers combining dual-expert ratings with silver-standard metrics (Dice, HD95) as summarized in Table 1. This approach provides a proxy that captures both clinical utility and geometric precision.

2.2 Experimental Framework

Training Protocol: All models are trained using the nnU-Net [12] with `3d_fullres` (3D Full Resolution) configuration for 1000 epochs with Adam optimizer and nnU-Net’s standard data augmentation (random rotations, scaling, elastic deformations, and gamma transformations). All experiments use 5-fold age-stratified cross-validation with a fixed seed for generalizable and reproducible findings.

Evaluation Metrics: We validate the segmentation using Dice Score and 95th percentile Hausdorff Distance (HD95) for boundary accuracy. Demographic disparities are quantified via two complementary fairness metrics [16, 17]: **Demographic Parity Difference (DPD)** measuring absolute performance gaps: $DPD = |P(\hat{y} = 1|A = a) - P(\hat{y} = 1|A = b)|$, while the **Disparate Impact Ratio (DIR)** captures relative disparities: $DIR = \frac{\min(P(\hat{y}=1|A=a), P(\hat{y}=1|A=b))}{\max(P(\hat{y}=1|A=a), P(\hat{y}=1|A=b))}$. Here, $\hat{y} = 1$ represents the beneficial outcome (high-quality segmentation; here, Dice Score > 0.8), A denotes the sensitive attribute (age), and a, b are distinct subgroups. DPD ranges from 0 (perfect parity) to 1 (maximum disparity), while DIR ranges from 0 to 1 (perfect fairness). Following the “four-fifths rule” from [18], a DIR below 0.8 is commonly considered evidence of adverse impact. In our study, we specifically compute $DPD(Y|O)$ and $DIR(Y|O)$, where Y and O denote the *Young* and *Older* subgroups, respectively, as they represent the most extreme demographic contrast.

| Age Group / Metric | Experiment 1 | | Experiment 2 | | Experiment 3 | Experiment 4 |
|-----------------------|------------------------|-------------------------|------------------------|------------------------|------------------------|---------------------------|
| | Observed Perf. | True Perf. [†] | M-SWAP-YOUNG | M-SWAP-OLDER | M-DIFF-BAL | M-BIASED-INPUT |
| Young (n=349) | 0.6941 ± 0.2489 | 0.7304 ± 0.2333 | 0.7320 ± 0.2381 | 0.7298 ± 0.2407 | 0.7317 ± 0.2314 | 0.6797 ± 0.2463 |
| Middle (n=349) | 0.7104 ± 0.2399 | 0.7333 ± 0.2253 | 0.7379 ± 0.2193 | 0.7316 ± 0.2321 | 0.7308 ± 0.2220 | 0.7132 ± 0.2282 |
| Older (n=349) | 0.7500 ± 0.2056 | 0.7703 ± 0.1899 | 0.7739 ± 0.1922 | 0.7869 ± 0.1755 | 0.7678 ± 0.1999 | 0.7458 ± 0.1991 |
| Average | 0.7182 ± 0.2334 | 0.7446 ± 0.2178 | 0.7479 ± 0.2182 | 0.7505 ± 0.2183 | 0.7435 ± 0.2188 | 0.7129 ± 0.2270 |
| Fairness Gap § | 0.0559 | 0.0399 | 0.0419 | 0.0571 | 0.0361 | 0.0661 |
| ANOVA p-value | 0.0049** | 0.0260* | 0.0227* | 0.0149* | 0.0481* | 0.0006** |
| DPD (Y O) | 0.1060 | 0.0802 | 0.0716 | 0.0777 | 0.0762 | 0.1146[‡] |
| DIR (Y O) | 0.8150 | 0.8710 | 0.8853 | 0.8755 | 0.8761 | 0.7895[‡] |

** $p < 0.01$, * $p < 0.05$ indicates statistically significant group differences.

[†] Highlighted column indicates the M-BASELINE (True Perf.) model, used as the reference for comparisons.

[‡] Indicates a potential adverse impact, defined here by common heuristics: a DIR < 0.80 (the four-fifths rule).

Table 2: Experimental results present the mean performance (\pm standard deviation), the Fairness Gap, statistical significance of group differences (ANOVA p-value), and formal fairness metrics for each setting. M-BASELINE (Gold-Standard) serves as the primary reference for comparing the effects of different interventions. All evaluations are against the Gold-Standard ground truth, except where explicitly mentioned.

Additionally, we report the **fairness gap** (§) as the absolute difference in mean performance between the highest- and lowest-performing demographic subgroups [19]. Statistical significance of group differences was tested using OLS regression [20] (Performance \sim Age) and ANOVA at $\alpha = 0.05$.

2.3 Controlled Experiments for Bias Source

We conduct a sequence of controlled experiments to diagnose the reasons underlying bias: first we establish the existence of bias in the data and baseline model, then testing the hypothesis of label and representational bias, and finally demonstrating bias amplification when training on biased labels.

Experiment 0 - Establishing Anatomical Disparity and Benchmark Bias: First, to test for an underlying anatomical basis of bias, we performed a morphometric analysis of Gold-Standard labels (see Fig. 2). This reveals a *significant* disparity across age groups, where tumors in the *Young* cohort are 66% larger in volume and exhibit 70% greater variance than *Older* cohort. This provides evidence of an underlying anatomical *representational bias*. To further establish an initial baseline of real-world bias – a fairness audit on the complete cohort of automated silver-standard labels reveals a significant relation between age and segmentation quality (OLS Regression; Dice score: $R^2 = 0.0104$, $p = 0.0001$; HD95: $R^2 = 0.0093$, $p = 0.0009$; see Fig. 1). Formal fairness metrics support the finding (DPD: 0.0887; DIR: 0.699), indicating the *Young* group achieves a high performance at only $\approx 70\%$ the rate of the *Older* group. *This confirms that the Silver-Standard labels are indeed systematically biased, motivating our subsequent experiments.*

Experiment 1 - The “Biased Ruler” Effect: Having confirmed a bias in the Silver-Standard labels, we next investigate the effect of using biased labels for validation, here exempli-

fied by the Silver-Standard labels. This is particularly relevant as many segmentation datasets, both public benchmarks and those used for development of real-life (bio)medical imaging segmentation tools, are based on semi-automatic “ground truth” annotations that could be similarly biased. We use the Silver-Standard labels as an example of *observed* labels, and compare performance with respect to biased observed labels to the true performance quantified using the Gold-Standard labels, in this experiment considered *true* labels.

To this end, we train an M-BASELINE model under an age class-balanced protocol ($n = 349$ per group). The results (Table 2) show a statistically significant *observed* performance bias against the *Young* cohort (§ = 0.0559, $p = 0.0049$) when using the observed labels, which is 40% higher than the *true* bias (§ = 0.0399, $p = 0.0260$). This inflation of observed bias is also reflected in the formal fairness metrics (DPD: 0.0802 \rightarrow 0.1060; DIR: 0.8710 \rightarrow 0.8150). *This “Biased Ruler” effect quantitatively demonstrates how relying on flawed benchmarks for fairness auditing can misrepresent the model’s true performance disparities.*

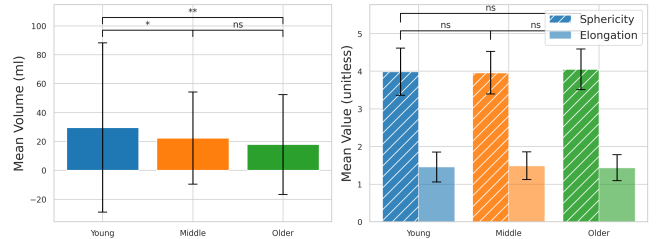


Fig. 2. (a) Tumor volume is, on average, larger and has higher variance in the *Young* group ($p < 0.01$ for Y-O). (b) In contrast, basic tumor shape metrics (sphericity and elongation) show no statistically significant difference.

Experiment 2 - Label Bias Sensitivity: We now investigate whether the true bias in M-BASELINE was caused by a superficial sensitivity to label quality. The M-SWAP-YOUNG and M-SWAP-OLDER replaced 100% of the *Tier 1* labels with their biased, Silver-Standard counterparts. This intervention had no meaningful effect on the model’s bias, see Table 2. The fairness gap remained stable, *refuting the hypothesis that the models’ fairness is fragile or primarily driven by a small subset of high-quality labels for any specific subgroup.*

Experiment 3 - Persistence of Representational Bias: Here, we test whether the true bias was caused by a quantitative imbalance in the distribution of *hard* cases. The M-DIFF-BAL model is trained on a dataset carefully balanced to provide each age group with an identical distribution of *easy* ($n = 143$) and *hard* ($n = 206$) cases. The results (Table 2) show that the difficulty-balancing intervention failed to eliminate the bias, leaving the fairness gap unchanged. This *refutes the hypothesis that a simple quantitative imbalance of difficult cases causes the bias.*

Experiment 4 - Training on Biased Labels Amplifies Bias: In this experiment, we train an age-group balanced model M-BIASED-INPUT on biased Silver-Standard labels. The results (in Table 2) confirm the hypothesis of **bias amplification**. The fairness gap widened by 66% relative to M-BASELINE (§ from 0.0399 \rightarrow 0.0661), and the bias became statistically severe ($p = 0.0006$). This amplification is even more evident when viewed through formal fairness metrics: the DIR dropped below the standard threshold to 0.7895.

3 Results and Conclusion

Our experimental study systematically diagnosed the mechanism behind age bias in breast cancer tumor segmentation. We first demonstrate (Experiment 0) that automated Silver-Standard segmentation labels are biased, and that the age-balanced model exhibits a statistically significant bias whose observed magnitude would be substantially inflated (40%) if we had only observed performance with respect to the automated Silver-Standard labels (Experiment 1). This effect has critical clinical implications: In real-world deployment, such a biased evaluation framework could mask true model performance, leading to undetected diagnostic failures and delayed treatment interventions. Moreover, if biased segmentation labels are used to validate model updates and guide clinical thresholds, this may systematically disadvantage younger patients by setting performance standards that appear adequate but actually mask age-related disparities. This underscores the urgent need for awareness of the effects of label bias in segmentation validation.

The following experiments (Experiments 2-3) refute the hypothesis that bias stems from superficial sensitivity to label quality, where replacing high-quality labels with biased counterparts had no meaningful effect on fairness gaps. The difficulty-balancing intervention then isolated representation

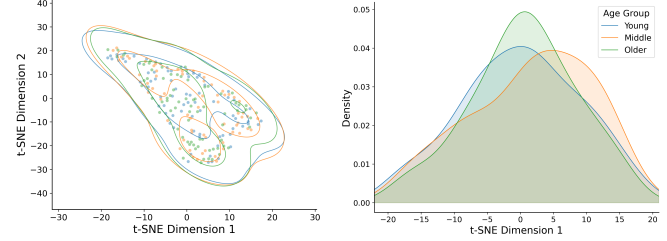


Fig. 3. Inspection of subgroup distribution shifts in the feature space projection. t-SNE (t-distributed stochastic neighbor embedding) [23] embeddings (for representative fold 0). *Left:* Scatter plot of the first two t-SNE dimensions. *Right:* Density distribution of the first t-SNE dimension. Both plots indicate a strong overlap, suggesting the model’s latent space does not strongly separate representations by age. Clustering metrics quantify this overlap across 5-folds (Mean \pm Std): For t-SNE, Silhouette = -0.0312 ± 0.0112 , Purity = 0.3943 ± 0.0193 , ARI (Adjusted Rand Index; $[-1, 1]$, higher is better agreement) = 0.0033 ± 0.0113 , NMI (Normalized Mutual Information; $[0, 1]$, higher is better agreement) = 0.0126 ± 0.0098 . Low ARI and NMI values confirm poor correspondence between the embedding structure and true age groups. t-SNE parameters: perplexity $\approx n/10$ (clipped to 5–50), LR 200, iters 1500, init = PCA.

as a definitive cause. The failure to completely eliminate bias point out that the problem is not the *quantity* of hard cases, but their qualitative nature. Our morphometric analysis provides supporting evidence: *Hard* cases for *Young* group are part of an anatomically distinct distribution. This aligns with clinical literature reporting that breast cancers in younger women are often larger and more aggressive [21, 22], and physiological factors mentioned in Section 1. Visual analysis of learned representations supports this conclusion (see Fig. 3).

More critically, training on biased Silver-Standard labels widened the fairness gap by 66% (Experiment 4). This creates dangerous clinical implications for modern AI development pipelines, where models are increasingly retrained on machine-generated labels to scale annotation efforts. For younger patients, this amplification may degrade segmentation quality and affect treatment planning.

In conclusion, we provide both a systematic diagnosis of age-related algorithmic bias and a demonstration of its serious clinical implications. The nature of this bias is learnable, amplified by label bias, and rooted in qualitative representational disparities, which demand fundamental changes to fairness practices in medical imaging. Future work must address qualitative representational interventions rather than rebalancing strategies. Additionally, rigorous auditing protocols using high-quality benchmarks must be established to detect and prevent bias propagation in automated data pipelines, ensuring that efforts to scale AI systems do not unintentionally scale their inequities.

4 Acknowledgment

This work was funded by the Novo Nordisk Foundation under project number 0087102.

5 References

- [1] Zikang Xu, Jun Li, Qingsong Yao, et al., “Addressing fairness issues in deep learning-based medical image analysis: a systematic review,” *npj Digital Medicine*, vol. 7, no. 1, pp. 286, 2024.
- [2] Emma A.M. Stanley, Raissa Souza, Matthias Wilms, et al., “Where, why, and how is bias learned in medical image analysis models? a study of bias encoding within convolutional networks using synthetic data,” *eBioMedicine*, vol. 111, pp. 105501, 2025.
- [3] Glenn Flores and Committee on Pediatric Research, “Racial and ethnic disparities in the health and health care of children,” *Pediatrics*, vol. 125, no. 4, pp. e979–e1020, 2010.
- [4] Amit Kumar Kundu, Florence X Doo, Vaishnavi Patil, et al., “Detecting and monitoring bias for subgroups in breast cancer detection ai,” *arXiv preprint arXiv:2502.10562*, 2025.
- [5] Keri Stephens, “How race and age impact ai mammo-gram results,” *AXIS Imaging News*, 2024.
- [6] Nora Eisemann, Stefan Bunk, Trasilas Mukama, et al., “Nationwide real-world implementation of ai for cancer detection in population-based mammography screening,” *Nature medicine*, vol. 31, no. 3, pp. 917–924, 2025.
- [7] Jordan Tschida, Mayanka Chandrashekar, Alina Peluso, et al., “Evaluating algorithmic bias on biomarker clas-sification of breast cancer pathology reports,” *JAMIA open*, vol. 8, no. 3, pp. ooaf033, 2025.
- [8] Jonas Gjesvik, Nataliia Moshina, Christoph I. Lee, et al., “Artificial intelligence algorithm for subclinical breast cancer detection,” *JAMA Network Open*, vol. 7, no. 10, pp. e2437402–e2437402, 10 2024.
- [9] Ok Hee Woo, Sung Eun Song, Su Jin Choe, et al., “In-vasive breast cancers missed by ai screening of mammo-grams,” *Radiology*, vol. 315, no. 3, pp. e242408, 2025.
- [10] Noel C. F. Codella, David Gutman, M. Emre Celebi, et al., “Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic),” in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 2018, pp. 168–172.
- [11] Timothy J. Hendrickson, Paul Reiners, Lucille A. Moore, et al., “Bibsnet: A deep learning baby image brain segmentation network for mri scans,” *bioRxiv*, 2023.
- [12] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, et al., “nnu-net: a self-configuring method for deep learning-based biomedical image segmentation,” *Nature meth-ods*, vol. 18, no. 2, pp. 203–211, 2021.
- [13] Deepa Krishnaswamy, Dennis Bontempi, David Clunie, et al., “Ai-derived annotations for the nlst and nslc-radiomics computed tomography imaging collections,” May 2023.
- [14] Lidia Garrucho, Claire-Anne Reidel, Kaisar Kushibar, et al., “Mama-mia: A large-scale multi-center breast cancer dce-mri benchmark dataset with expert segmen-tations,” *arXiv e-prints*, pp. arXiv–2406, 2024.
- [15] Rasmus OC Humlevik, Amalie A Svanøe, Turid Aas, et al., “Distinct clinicopathological features and treat-ment differences in breast cancer patients of young age,” *Scientific Reports*, vol. 15, no. 1, pp. 5655, 2025.
- [16] Simon Caton and Christian Haas, “Fairness in machine learning: A survey,” *ACM Comput. Surv.*, vol. 56, no. 7, Apr. 2024.
- [17] Alessandro Castelnovo, Riccardo Crupi, Greta Greco, et al., “A clarification of the nuances in the fairness met-rics landscape,” *Scientific reports*, vol. 12, no. 1, pp. 4209, March 2022.
- [18] Equal Employment Opportunity Commission et al., “Uniform guidelines on employee selection proce-dures,” *Fed Register*, vol. 1, pp. 216–243, 1990.
- [19] Khoa Tran and Simon S. Woo, “Fairness and robust-ness in machine unlearning,” in *Companion Proceed-ings of the ACM on Web Conference 2025*, New York, NY, USA, 2025, WWW ’25, p. 1336–1340, Association for Computing Machinery.
- [20] Clara Dismuke and Richard Lindrooth, “Ordinary least squares,” *Methods and designs for outcomes research*, vol. 93, no. 1, pp. 93–104, 2006.
- [21] Melinda A Maggard, Jessica B O’Connell, Karen E Lane, et al., “Do young breast cancer patients have worse outcomes?,” *Journal of Surgical Research*, vol. 113, no. 1, pp. 109–113, 2003.
- [22] Abdulkader M Albasri, “Clinicopathological character-istics of young versus older patients with breast cancer:

A retrospective comparative study from the madinah region of saudi arabia,” *Saudi Medical Journal*, vol. 42, no. 7, pp. 769, 2021.

- [23] Laurens van der Maaten and Geoffrey Hinton, “Visualizing data using t-sne,” *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.