

Personalized Federated Learning for Equitable Breast Cancer Detection in Underrepresented Pacific Islander Populations

Lydia J. Sollis^{1,2}, Patricia M. Young³, Arianna Bunnell^{1,2}, Brandon Quon¹, Brenda Y. Hernandez¹, Thomas K. Wolfgruber¹, and John A. Shepherd¹

¹ University of Hawai'i Cancer Center, Honolulu, HI, USA
lsollis@cc.hawaii.edu

² University of Hawai'i at Mānoa, Department of Information and Computer Sciences, Honolulu, HI, USA

³ Hawai'i Pacific University, Department of Computer Science and Engineering, Honolulu, HI, USA

Abstract. Deep learning models for mammographic breast cancer detection are developed and validated predominantly in cohorts from the continental US and Europe, while Native Hawaiian and Pacific Islander women, who experience later-stage diagnoses and poorer outcomes, remain underrepresented. Federated learning (FL) enables collaborative training without sharing patient data, but standard Federated Averaging produces a single global model that may underperform at individual sites differing in population, case mix, and image acquisition. We adapted GMIC, a multiple-instance classifier for mammographic breast cancer detection, for federated fine-tuning with NVIDIA FLARE and evaluated shared-global (FedAvg, FedProx) and personalized (FedBN, Ditto) strategies, plus a module-wise Ditto variant with per-module regularization. We constructed a three-site federation comprising a diverse Pacific Island cohort and cohorts based in the mainland US and Australia from the public RSNA Screening Mammography dataset. FedAvg, FedProx, and FedBN improved the largest site but left smaller sites at or below pre-trained performance; FedBN, despite personalizing batch-normalization statistics, was least equitable. Ditto improved the worst-performing site (test AUC 0.87 vs. 0.83 for the others and 0.84 for local-only training) while reducing cross-site AUC spread (0.019 vs. 0.030 for FedAvg) and matching the pooled performance of centralized training, without sharing data. Module-wise Ditto provided no measurable advantage over standard Ditto. To our knowledge, this is the first equity-motivated application of personalized FL to mammographic cancer detection in a Pacific Islander population, suggesting that personalized FL can improve robustness to site heterogeneity by allowing local adaptation while retaining the benefits of collaborative learning.

Keywords: Federated learning · Personalization · Breast cancer · Mammography · Health equity · Pacific Islander · GMIC

1 Introduction

Breast cancer is the most common malignancy in women worldwide [1], and screening mammography is the primary tool for early detection. Deep learning now matches radiologist-level performance in mammographic interpretation [2–4], yet models are overwhelmingly developed on majority-White, continental US, and European cohorts [5]. Such models can transfer poorly to sites whose patients, case mix, and imaging differ from the training distribution. This concern is especially acute for Native Hawaiian and Pacific Islander (NHPI) women, who face socioeconomic barriers and medical mistrust [6], resulting in later-stage diagnoses [7] and poorer survival [8].

Hawai‘i has no racial majority (37% Asian, 10% NHPI, 23% White) [9], and its population differs from mainland US and other western cohorts in imaging-relevant ways: higher breast density [10] and greater than two-fold variation in incidence and survival within subgroups of the aggregate “Asian” label [11, 12]. Federated learning (FL) [13] enables collaborative training without sharing patient data, but its standard algorithm (FedAvg) averages client parameters into one global model that, under such non-independent and identically distributed (non-IID) heterogeneity, may perform inadequately for some or all sites [14]. FedProx [14] adds a proximal penalty that keeps each client’s local update near the current global model, reducing client drift while still producing a single shared model.

Personalized FL instead lets each site maintain a model adapted to its local distribution while still benefiting from collaboratively learned parameters, replacing a single shared model with a family of related site-specific models. FedBN [15] keeps batch-normalization parameters local so each site adapts feature normalization to its own imaging distribution; Ditto [16] trains a full per-site model regularized toward the shared global FedAvg model, with a coefficient λ controlling the degree of personalization.

In this work we: (1) adapt GMIC [2], a high-resolution globally-aware classifier, for federated training with NVFlare [17] across a heterogeneous three-site federation; (2) evaluate shared-global (FedAvg, FedProx) and personalized (FedBN, Ditto) FL algorithms using a diverse Pacific Island site (the largest of our three cohorts) alongside US and Australian screening sites; (3) show that shared-global FL and FedBN’s batch-norm-only personalization improve the largest site but not the smaller sites, whereas Ditto’s per-site models improve the worst-performing site, equalize performance, and approach a centralized data-pooling ceiling without sharing data; (4) propose module-wise Ditto, an architecture-motivated shared-backbone/private-head variant, and report the negative result that per-module tuning performs on par with scalar Ditto; and (5) frame these findings as evidence for *when* federation helps equity: from a strong pretrained backbone, standard federation adds little over local fine-tuning and can disadvantage smaller sites, whereas per-site personalization recovers them, with the site most at risk set by data scale and cross-site heterogeneity.

2 Related Work

2.1 Deep Learning for Mammographic Breast Cancer Detection

Deep learning for mammography has progressed from patch-level classifiers to whole-image models at near-native resolution, reaching radiologist-level performance [4], and foundation-model approaches such as Mammo-CLIP attain expert-level accuracy with less labeled data [18]. Mammography AI has not been robustly validated in adequately powered, disaggregated Native Hawaiian and Pacific Islander cohorts, and commercial AI algorithms show race-associated performance differences [19], underscoring the risk of deploying models without population-specific evaluation.

We adopt GMIC [3] as our backbone: it processes mammograms at high resolution (2944×1920) using less GPU memory than ResNet-34 and yields weak lesion localization from image-level labels alone.

2.2 Federated Learning for Breast Cancer and Mammography

In mammography, FL has largely addressed breast-*density* classification [20] rather than cancer detection; the few cancer-detection efforts target class imbalance and non-IID distributions [21] or domain-harmonization preprocessing [22]. Critically, no prior mammography FL work has addressed *personalization for health equity* across racially heterogeneous populations: personalized methods (FedProx [14], FedBN [15], Ditto [16]) and fairness-aware FL [23] have been evaluated only in general settings, and none on breast-cancer screening or Pacific Islander populations. This paper addresses that gap.

3 Methods

3.1 GMIC Architecture

We adopted the GMIC framework [3, 2] for breast-level malignant classification. Given a grayscale image (2944×1920), GMIC proceeds in three stages: a reduced-width ResNet-22 processes the high-resolution image at low memory cost, producing saliency maps of suspicious findings; a full-width ResNet-18 extracts features from $K = 6$ ROI patches via gated attention; and a fusion module combines global and local representations.

3.2 Federated GMIC with NVIDIA FLARE

We implemented federated GMIC using NVFlare [17] via a custom executor that handled local training and server communication. Each client preprocessed identically: largest-connected-component breast cropping, orientation standardization (right-laterality views mirrored), and isotropic rescaling with zero-padding to the 2944×1920 input; ROIs were selected by GMIC’s saliency mechanism. Clients ran in Podman/Docker containers with GPU passthrough, enabling deployment across heterogeneous hardware (per site in Table 1).

3.3 Federated Learning Strategies

Centralized. We trained a single model on the pooled three-site data, a benchmark isolating decentralization from personalization.

Local-only. Each site trained GMIC on its own data, a baseline any federated method should beat.

FedAvg [13] averages the client weights (dataset-size-weighted) each round into a single shared global model.

FedProx [14] adds a proximal term $\frac{\mu}{2}\|w_k - w^t\|^2$ to each local objective, penalizing divergence from the current global w^t ; we fixed $\mu = 0.1$, treating the proximal strength as a prespecified algorithm parameter rather than an independently optimized hyperparameter.

FedBN [15] is FedAvg with batch-normalization parameters kept local.

Ditto [16] jointly maintains a global model w^* , trained by FedAvg, and a personalized model v_k for each client: $\min_{v_k} F_k(v_k) + \frac{\lambda}{2}\|v_k - w^*\|^2$. The regularization parameter λ controls personalization: larger values constrain the personalized model more strongly toward the global model, whereas smaller values permit greater local adaptation. Each site ultimately deploys its own v_k . We selected λ from $\{0.01, 0.05, 0.1, 0.5, 1, 2\}$ based on 12-round validation runs, choosing the value that maximized the worst-site validation AUC.

Module-wise Ditto. We extended Ditto by assigning a separate regularization strength to each major GMIC module, $F_k(v_k) + \sum_m \frac{\lambda_m}{2}\|v_k^m - w^{*m}\|^2$, where m denotes the whole-image global branch, high-resolution local backbone, or fusion/decision head. Based on the architecture, we defined an initial hypothesis-driven configuration of $\lambda_{\text{global}} = 0.01$, $\lambda_{\text{local}} = 0.5$, and $\lambda_{\text{fusion}} = 0$, reflecting the expectation that the fusion head would benefit from the greatest site-specific adaptation, whereas the feature-extraction branches would remain more closely coupled to the shared model. We then evaluated this hypothesis using planned one-factor-at-a-time validation sweeps in which the regularization strength of each module was varied while the others were held fixed (Supplementary Table S1b). Because none of the alternative settings produced a meaningful improvement in worst-site validation AUC, with all observed differences falling within the validation confidence interval, we retained the original architecture-motivated configuration.

3.4 Training Protocol

We fine-tuned GMIC from the released `sample_model_5` weights with a deep-supervision objective (binary cross-entropy on the global, local, and fusion outputs plus a β -weighted L_1 saliency penalty), a class-balanced sampler for class imbalance, and Adam with differential learning rates for the heads and backbone. The centralized and local-only models trained up to 40 epochs with batch size 32 and early stopping on breast-level validation AUC; the federated strategies instead ran 60 rounds of one local epoch per client, all from the same pretrained initialization. The training configuration was established through a hand-tuned,

Table 1: Per-site composition, acquisition, and local training hardware. Site A is the registry-linked Hawai’i cohort; Sites B and C are the US and Australian clients of the RSNA dataset, trimmed to a 3:1 control-to-case ratio to match Site A. [†]Scanner manufacturer not reported for Site C.

	Site A (HIPIMR)	Site B (RSNA-US)	Site C (RSNA-AU)
Exams (cancer)	2,188 (520)	996 (249)	936 (234)
Train/Val/Test	1,531/218/439	697/101/198	655/94/187
Mean age (y)	61.2	58.9	60.1
Density (BI-RADS)	available	available	n/a
Scanner(s)	Hologic ($\approx 96\%$)	7 units (Hologic)	3 units [†]
Hardware	V100 (DGX-1)	GCP VM, T4	Win10, RTX A4000

one-factor-at-a-time ablation on the pooled data; full hyperparameters and ablation results are provided in the Supplementary Material.

3.5 Datasets and Federated Sites

We constructed a three-site federation: a demographically distinct Hawai’i cohort (Site A) and two clinical sites containing subsets of the public RSNA Screening Mammography dataset [24], maintained as separate clients (Sites B and C) to capture realistic cross-site variation. Labels were breast-level per patient and laterality, assigned by matching each cancer’s laterality to the imaged breast; splits were patient-level (70% train/10% validation/20% test, stratified by label and age) with no patient shared across partitions. Table 1 summarizes per-site composition, acquisition, and training hardware.

Site A (HIPIMR). Site A was a registry-linked cohort: cancer cases were invasive and in situ breast cancers from the Hawai’i Tumor Registry with a mammogram within three months of diagnosis, matched 1:3 to controls on age, density, and scanner, yielding 2,188 examinations (520 cancer) after image-quality review. This cohort reflected Hawai’i’s diversity with no majority group: Native Hawaiian and Pacific Islander women were 16.2% (vs. 0.8% at the RSNA US site) and Asian women 58.0% (with Japanese, Filipino, Chinese, and Korean subgroups); 55.3% of graded breasts were dense (BI-RADS C/D). To match Site A’s class balance, the RSNA sites were trimmed to a 3:1 control-to-case ratio.

Site B (RSNA-US). The US site (Emory Healthcare, Atlanta) contributed 996 examinations (249 cancer, 167 invasive); composition, acquisition, and training hardware are summarized in Table 1.

Site C (RSNA-AU). The Australian site contributed 936 examinations (234 cancer, 174 invasive) under a population screening program (minimum age 40); details are in Table 1. The RSNA challenge reported lower algorithm sensitivity at the US site than here (52.0% vs. 68.1%) [25], and the US site (Site B) was likewise our lowest-performing.

Label heterogeneity. Site A labels were registry-confirmed invasive and in situ cancers from screening and diagnostic exams; RSNA labels were screen-detected, biopsy-confirmed cancers (including in situ disease). Both targeted image-visible malignancy but differed in label provenance and acquisition context, motivating personalized FL.

3.6 Evaluation

We reported breast-level AUC following the original GMIC protocol [3], averaging craniocaudal (CC) and mediolateral-oblique (MLO) view predictions per breast, with DeLong 95% confidence intervals [26]. The deployed round was selected by validation AUC (pooled for shared-global models, site-specific for personalized). Following Ditto [16], we measured equity as the sample standard deviation of per-site AUC across clients (the Spread column of Table 2); lower is more uniform. For the Site A subgroup analysis we report, at fixed clinical operating points, sensitivity at 90% specificity and specificity at 90% sensitivity.

4 Results

4.1 Detection Performance

Table 2 reports per-site and pooled AUC at one pooled-validation round; per-site best-validation rounds are in Supplementary Table S2.

A shared global model did not ensure equity. The centralized model set the benchmark at pooled test AUC 0.887 (95% CI 0.860–0.915), versus 0.857 pretrained (≈ 0.03 from fine-tuning). FedAvg, FedProx, and FedBN all achieved similar pooled AUCs of 0.858–0.861, but left Site B at 0.829–0.833, at or below its pretrained performance. For all three methods, Site B’s best round was round 0, corresponding to the pretrained checkpoint. FedProx tracked FedAvg (pooled 0.858 vs. 0.861). FedBN was least equitable: best at Site A (0.893) but worst among federated methods at Site C (0.822), with the largest spread (0.038). Local-only matched the benchmark at Site A (0.900) but not Site B (0.843).

Ditto achieved the highest worst-site AUC. With its regularization set to $\lambda = 0.05$ (the worst site, always Site B, favored weak regularization on validation and degraded as λ grew), Ditto was the only federated strategy that improved the straggler site, reaching a Site B AUC of 0.866 under pooled-validation selection (the own-best-validation view per site is in Supplementary Table S2). This exceeded local-only (0.843), centralized (0.845), and all shared-global methods, while matching Sites A and C. Ditto also achieved the second-lowest cross-site spread (0.019) and a pooled AUC of 0.889, matching the centralized benchmark without sharing data. Module-wise Ditto’s per-module regularization strengths were swept (Supplementary Table S1); at $\lambda_{\text{global}} = 0.01$, $\lambda_{\text{local}} = 0.5$, $\lambda_{\text{fusion}} = 0$ it performed similarly to scalar Ditto (spread 0.018). Given the modest case counts (Site B test $n = 396$; DeLong CI $\approx \pm 0.07$), the advantage of the Ditto variants was directionally consistent rather than individually significant.

Table 2: Per-site and pooled breast-level malignant AUC (95% DeLong CI) by strategy. The Pooled column is the AUC over the combined three-site test set; cross-site spread is the (sample, ddof = 1) standard deviation of the per-site AUCs. Pretrained is the released GMIC weights with no fine-tuning (the common starting point for all methods); Centralized is a single pooled-data model evaluated per origin site; Local-only is each site’s own model on its own test set (no single pooled model, so its Pooled/Spread cells are n/a).

Strategy	Site A (HIPIMR)	Site B (RSNA-US)	Site C (RSNA-AU)	Pooled	Spread (SD)
Pretrained (no fine-tuning)	0.870	0.842	0.854	0.857	0.014
Centralized (pooled data)	0.896	0.845	0.905	0.887	0.032
Local-only (own site)	0.900	0.843	0.888	n/a	n/a
FedAvg	0.887	0.829	0.842	0.861	0.030
FedProx	0.876	0.829	0.851	0.858	0.024
FedBN	0.893	0.833	0.822	0.859	0.038
Ditto	0.900	0.866	0.897	0.889	0.019
Module-wise Ditto	0.898	0.861	0.882	0.884	0.018

4.2 Subgroup Performance at Site A

Because Site A was the only cohort with per-examination race/ethnicity data, we evaluated group-specific performance there for the deployed Ditto model and pretrained baseline (Table 3). Under the pretrained model, NHPI women were the worst-served group (AUC 0.839), below both the Asian and overall Site-A performance. Personalization closed this gap: Ditto raised NHPI AUC to 0.898, reaching parity with the Asian subgroup and the overall cohort. The improvement is clearest at clinical operating points: at a fixed 90% specificity Ditto lifted NHPI sensitivity from 0.60 to 0.80, and at a fixed 90% sensitivity it doubled NHPI specificity from 0.39 to 0.79, the largest operating-point gain of any group. The equity benefit is therefore not merely a cross-site average but a gain for the underrepresented group itself. On Site A, every federated strategy improved NHPI performance over the pretrained baseline and reached comparable subgroup AUCs (Supplementary Table S3). These estimates rest on modest case counts (NHPI 20 positive breasts), so we report them as exploratory; because race is coded granularly for registry cancer cases but coarsely for matched controls, we pool Asian ethnicities and omit groups with too few positives.

5 Discussion and Conclusion

Personalized federated learning recovered the worst-performing site and closed the performance gap for the underrepresented subgroup, without sharing patient data. Ditto was the only strategy to improve worst-site performance while reducing cross-site disparity, reaching a pooled AUC comparable to centralized training, whereas shared-model approaches favored the larger site and left the

Table 3: Site A (HIPIMR) per-group performance at fixed clinical operating points: pretrained baseline vs. Ditto ($\lambda = 0.05$). Sn@Sp₉₀ is sensitivity at 90% specificity and Sp@Sn₉₀ is specificity at 90% sensitivity. n (positive breasts) in parentheses; small- n groups (†) are exploratory. PPV and false-positive rate are omitted because the case-enriched cohort ($\approx 12\%$ malignant vs. screening $\approx 0.5\%$) would not yield clinically representative values.

Group	Pretrained			Ditto		
	AUC	Sn@Sp ₉₀	Sp@Sn ₉₀	AUC	Sn@Sp ₉₀	Sp@Sn ₉₀
Overall (878, 104)	0.870	0.683	0.532	0.899	0.760	0.672
NHPI (142, 20) [†]	0.839	0.600	0.393	0.898	0.800	0.787
Asian (496, 66)	0.882	0.667	0.605	0.900	0.758	0.593
White (198, 11) [†]	0.825	0.727	0.492	0.870	0.818	0.690

smaller sites at or below their pretrained performance. Because module-wise Ditto performed similarly to scalar Ditto, we retained the simpler scalar variant.

Limitations. Each method was trained once, so we did not quantify optimization variance. Using one local epoch per round limited client drift and may have reduced differences among the regularization-based methods; these differences could grow with additional local epochs. All models were initialized from the same pretrained GMIC checkpoint because our data volume precluded training from random initialization; repeating each method across the five released GMIC ensemble members would provide distinct initializations for assessing variance and statistical significance. Per-site counts are modest (Site B test $n = 396$, DeLong CI $\approx \pm 0.07$), so Ditto’s advantage is directionally consistent but not statistically significant. Detailed per-examination race/ethnicity data were available only for Site A (RSNA provides site-level summaries only), which is also the best-served site. The equity question we most want to answer is per-group performance at the smaller, data-limited sites, where personalization has its largest effect; those sites lack the required demographic labels, so we leave this to future work. Matching the RSNA sites to Site A’s 3:1 class balance also narrows one axis of real-world cross-site heterogeneity. Finally, we evaluated a single architecture, so the relative performance of the methods and the null result for module-wise personalization may not generalize to other models.

To our knowledge, this is the first application of personalized FL to mammographic cancer detection with an explicit health-equity framing for Native Hawaiian and Pacific Islander populations. Larger cohorts with harmonized demographic annotation are needed to establish statistical significance and to test personalization across more architectures and clinical applications.

Acknowledgments. This work was supported by National Institutes of Health grants U54CA143728, R01CA263491 (HIPIMR), and P30CA071789.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. World Health Organization: Breast cancer. <https://www.who.int/news-room/fact-sheets/detail/breast-cancer> (2024), fact sheet, last accessed 2026/05/23
2. Shen, Y., Wu, N., Phang, J., Park, J., Liu, K., Tyagi, S., Heacock, L., Kim, S.G., Moy, L., Cho, K., Geras, K.J.: An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *Medical Image Analysis* **68**, 101908 (2021). <https://doi.org/10.1016/j.media.2020.101908>
3. Shen, Y., Wu, N., Phang, J., Park, J., Kim, G., Moy, L., Cho, K., Geras, K.J.: Globally-aware multiple instance classifier for breast cancer screening. In: *Machine Learning in Medical Imaging (MLMI)*. LNCS, vol. 11861, pp. 18–26. Springer (2019). https://doi.org/10.1007/978-3-030-32692-0_3
4. McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafiyan, H., Back, T., Chesus, M., Corrado, G.S., Darzi, A., et al.: International evaluation of an ai system for breast cancer screening. *Nature* **577**(7788), 89–94 (2020)
5. Miyawaki, I.A., Banerjee, I., Batalini, F., Jorge, C.A.C., Celi, L.A., Cobanaaj, M., Dee, E.C., Gichoya, J.W., Kaffey, Z., Lloyd, M.R., et al.: Global disparities in artificial intelligence-based mammogram interpretation for breast cancer: A scientometric analysis of representation, trends, and equity. *European Journal of Cancer* **220**, 115394 (2025)
6. Tapparra, K., Miller, R.C., Deville, C.: Navigating native hawaiian and pacific islander cancer disparities from a cultural and historical perspective. *JCO Oncology Practice* **17**(3), 130–134 (2021)
7. Yoshikawa, G.T., Miyazaki, K.S.Y., Acoba, J.D., Fujii, T.: Racial and survival disparities in inflammatory breast cancer (ibc) and non-ibc: a population-based study focused on native hawaiians and other pacific islanders. *Frontiers in Oncology* **14**, 1390080 (2024)
8. Nguyen, A., Duckworth, E., Pascua, D., Coles, B., Galiano, R.: Disparities in breast cancer treatment and reconstruction among native hawaiian and pacific islander women: systematic review and meta-analysis. *Journal of Surgical Oncology* **131**(6), 1002–1012 (2025)
9. U.S. Census Bureau: Hawaii state profile — race and ethnicity. american community survey 5-year estimates, 2019–2023. <https://data.census.gov/profile/Hawaii?g=040XX00US15> (2024), last accessed 2026/05/23
10. Kerlikowske, K., Bissell, M.C.S., Sprague, B.L., Tice, J.A., Tossas, K.Y., Bowles, E.J.A., Ho, T.Q.H., Keegan, T.H.M., Miglioretti, D.L.: Impact of bmi on prevalence of dense breasts by race and ethnicity. *Cancer Epidemiology, Biomarkers & Prevention* **32**(11), 1524–1530 (2023). <https://doi.org/10.1158/1055-9965.epi-23-0049>
11. Loo, L.W.M., Williams, M., Hernandez, B.Y.: The high and heterogeneous burden of breast cancer in Hawaii: A unique multiethnic U.S. population. *Cancer Epidemiology* **58**, 71–76 (2019). <https://doi.org/10.1016/j.canep.2018.11.006>
12. Giaquinto, A.N., Sung, H., Newman, L.A., Freedman, R.A., Smith, R.A., Star, J., Jemal, A., Siegel, R.L.: Breast cancer statistics 2024. *CA: A Cancer Journal for Clinicians* **74**(6), 477–495 (2024). <https://doi.org/10.3322/caac.21863>
13. McMahan, B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A.y.: Communication-efficient learning of deep networks from decentralized data. In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*. pp. 1273–1282 (2017)

14. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems* **2**, 429–450 (2020)
15. Li, X., Jiang, M., Zhang, X., Kamp, M., Dou, Q.: Fedbn: Federated learning on non-iid features via local batch normalization. *arXiv preprint arXiv:2102.07623* (2021)
16. Li, T., Hu, S., Beirami, A., Smith, V.: Ditto: Fair and robust federated learning through personalization. In: *International Conference on Machine Learning (ICML)*. pp. 6357–6368. PMLR (2021)
17. Roth, H.R., Cheng, Y., Wen, Y., Yang, I., Xu, Z., Hsieh, Y.T., Kersten, K., Harouni, A., Zhao, C., Lu, K., et al.: Nvidia flare: Federated learning from simulation to real-world. *arXiv preprint arXiv:2210.13291* (2022)
18. Ghosh, S., Poynton, C.B., Visweswaran, S., Batmanghelich, K.: Mammo-clip: A vision language foundation model to enhance data efficiency and robustness in mammography. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. pp. 632–642. Springer (2024)
19. Nguyen, D.L., Ren, Y., Jones, T.M., Thomas, S.M., Lo, J.Y., Grimm, L.J.: Patient characteristics impact performance of ai algorithm in interpreting negative screening digital breast tomosynthesis studies. *Radiology* **311**(2), e232286 (2024)
20. Roth, H.R., Chang, K., Singh, P., Neumark, N., Li, W., Gupta, V., Gupta, S., Qu, L., Ihsani, A., Bizzo, B.C., et al.: Federated learning for breast density classification: A real-world implementation. In: *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning (DART/DCL), MICCAI Workshops*. pp. 181–191. LNCS, Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-60548-3_18
21. Jiménez-Sánchez, A., Tardy, M., González Ballester, M.A., Mateus, D., Piella, G.: Memory-aware curriculum federated learning for breast cancer classification. *Computer Methods and Programs in Biomedicine* **229**, 107318 (2023). <https://doi.org/10.1016/j.cmpb.2022.107318>
22. Tzortzis, I.N., Gutierrez-Torre, A., Sykiotis, S., Agulló, F., Bakalos, N., Doulamis, A., Doulamis, N., Berral, J.L.: Towards generalizable federated learning in medical imaging: A real-world case study on mammography data. *Computational and Structural Biotechnology Journal* **28**, 106–117 (2025)
23. Xing, H., Sun, R., Ren, J., Wei, J., Feng, C.M., Ding, X., Guo, Z., Wang, Y., Hu, Y., Wei, W., Ban, X., Xie, C., Tan, Y., Liu, X., Cui, S., Duan, X., Li, Z.: Achieving flexible fairness metrics in federated medical imaging. *Nature Communications* **16**(1) (2025). <https://doi.org/10.1038/s41467-025-58549-0>
24. Trivedi, H.M., Vazirabad, M., Kitamura, F.C., Chen, Y., Frazer, H.: Open-source dataset for the rsna screening mammography cancer detection challenge. *Radiology: Artificial Intelligence* p. e250375 (2026)
25. Chen, Y., Partridge, G.J.W., Vazirabad, M., Ball, R.L., Trivedi, H.M., Kitamura, F.C., Frazer, H.M.L., Retson, T.A., Yao, L., Darker, I.T.: Performance of algorithms submitted in the 2023 rsna screening mammography breast cancer detection ai challenge. *Radiology* **316**(2), e241447 (2025)
26. DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L.: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**(3), 837–845 (1988)