

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 · John A. Shepherd¹

¹ University of Hawai'i Cancer Center, Honolulu, HI, USA · ² University of Hawai'i at Mānoa, Dept. of Information & Computer Sciences · ³ Hawai'i Pacific University, Dept. of Computer Science & Engineering
Isollis@cc.hawaii.edu

Strasbourg, France
October 1, 2026

INTRODUCTION

Why Personalized Federated Learning?

Deep learning now matches radiologist-level mammography performance [4], but is developed almost entirely on continental-US and European cohorts [5]. Native Hawaiian and Pacific Islander (NHPI) women face later-stage diagnoses and poorer survival [7,8], shaped in part by cultural and structural barriers to care [6], yet remain absent from model training and validation.

Hawai'i has no racial majority, and differs from mainland cohorts in breast density and imaging-relevant population mix [9–11]. Our own study cohort reflects that diversity directly:

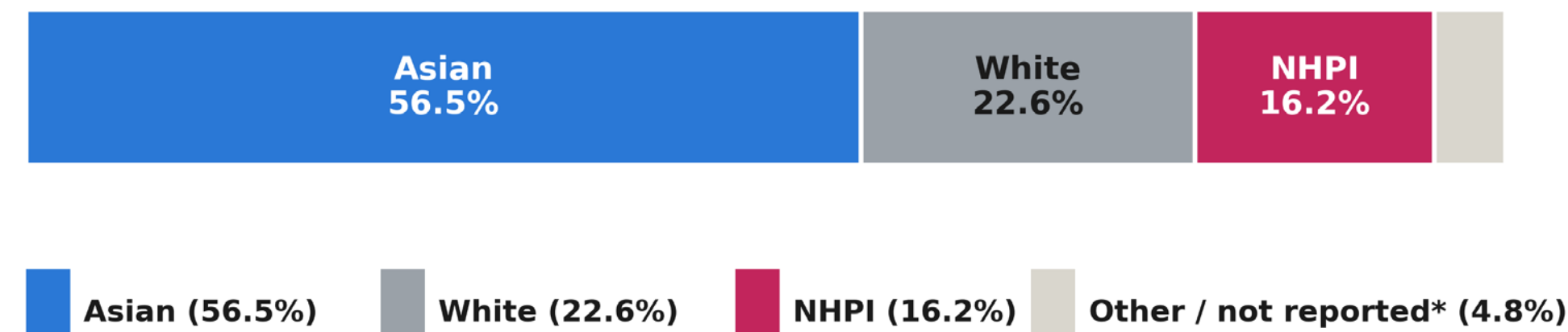


Fig. 1. *Includes small-n groups pooled for privacy; Site A (HIPIMR) test-set composition by self-reported race/ethnicity, n=878 breasts (this study).

Federated learning (FL) enables multi-site training without sharing patient data, but standard FedAvg [13] produces one global model that can underperform at sites that differ in population, case mix, and acquisition. Personalized FL instead lets each site keep a model adapted to its own distribution while still benefiting from shared learning.

How Ditto Personalizes

Ditto [16] jointly trains a shared global model w^* (via FedAvg [13]) and a personalized model v_k per site:

$$\text{minimize over } v_k : F_k(v_k) + (\lambda/2) \cdot \|v_k - w^*\|^2$$

Larger λ pulls a site's model back toward the shared global model; smaller λ allows more local adaptation. A single λ , swept over {0.01, 0.05, 0.1, 0.5, 1, 2} on validation data and chosen to maximize worst-site AUC, is applied at every site; each site deploys its personalized v_k , not the global model.

Related Work: The Gap

Mammography FL has primarily targeted breast density classification [20], class imbalance, and domain-harmonization preprocessing [21,22], not cancer detection through an equity lens. Ditto [16], a personalized FL aggregation method, has been shown to improve test set performance by ~6% while reducing variance by ~10% on benchmark datasets, but has not been used in breast cancer screening or with a Pacific Islander population. This is the first equity-motivated application of personalized FL to breast cancer detection.

Study Objectives

- Adapt GMIC [2,3], a high-resolution, multiple-instance learning (MIL) mammography classifier, for federated fine-tuning with NVIDIA FLARE [17].
- Evaluate shared-global (FedAvg [13], FedProx [14]) vs. personalized (FedBN [15], Ditto [16]) strategies across a heterogeneous 3-site federation.
- Test whether personalization recovers the performance of the worst site and closes the gap for underrepresented subgroups.
- Propose module-wise Ditto (shared backbone / private head) as an architecture-motivated variant.
- Ask when federation actually helps equity, versus when it can quietly disadvantage smaller sites.

METHODS

Three-Site Federation

Table 1. Per-site composition, acquisition, and local training hardware.

	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
NHPI representation	16.2%	0.8%	n/a
Dense breasts (BI-RADS C/D)	55.3%	available	n/a
Scanner(s)	Hologic (~96%)	7 units (Hologic)	3 units
Local hardware	V100 (DGX-1)	GCP VM, T4	RTX A4000

Site A: institutional registry (HIPIMR, University of Hawai'i Cancer Center). Sites B/C: public RSNA screening mammography dataset [24].

Dataset Composition Across Sites

- Site A (HIPIMR): Controls matched 3:1 to registry-confirmed invasive + in-situ cancers on age, density, and scanner; Hawaii population
- Sites B/C (public RSNA dataset [24]): screen-detected, biopsy-confirmed cancers, trimmed to the same 3:1 control-to-case ratio; mainland US (site B) and Australian (site C) population

Federated Setup

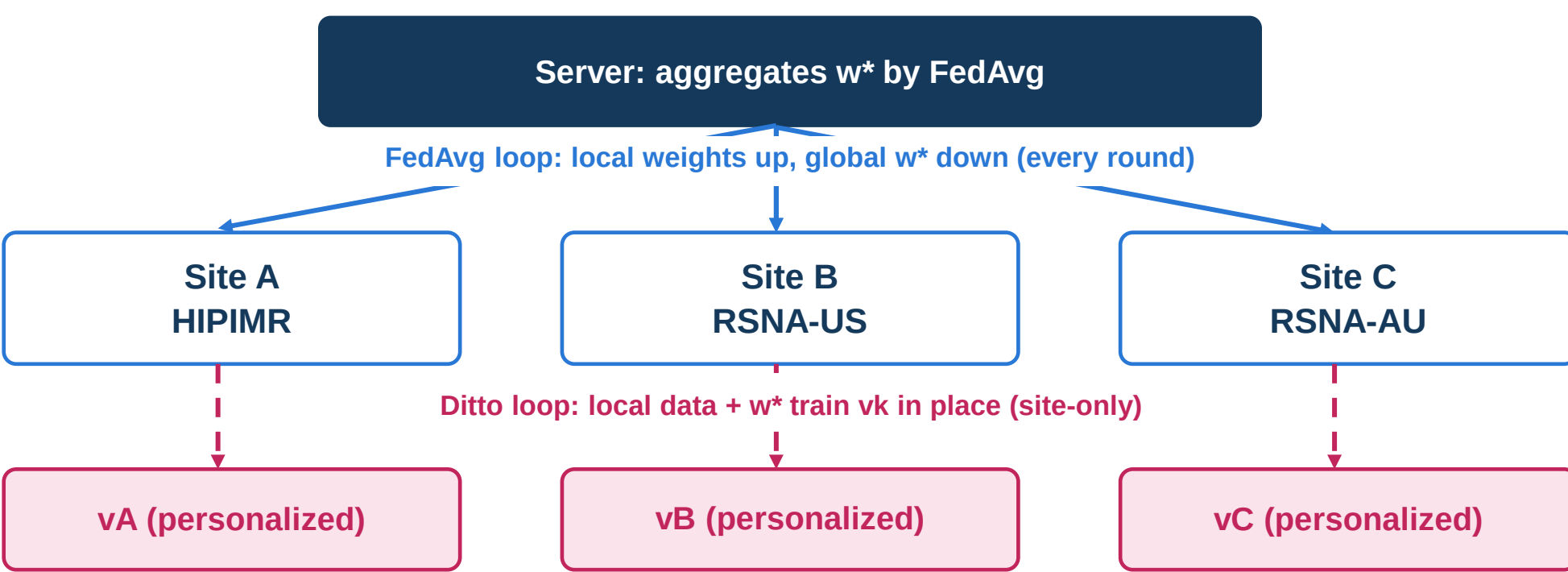


Fig. 2. Ditto's two training loops. Blue: the shared FedAvg loop through the server, each round. Orange: a loop that runs only at each site, fine-tuning a personalized v_k toward w^* on that site's own data; v_k is never sent to the server.

Training Strategies Compared

Table 2. Training strategies evaluated, from the pretrained starting point to full personalization.

Training Strategy	Definition
Centralized Benchmark	Single model on pooled data from all 3 sites
Local-only	Each site trains GMIC independently with just their local data
FedAvg [13]	Dataset-size-weighted average creates shared global model
FedProx [14]	FedAvg + proximal penalty ($\mu=0.1$) limiting client drift
FedBN [15]	FedAvg with batch normalization weights kept local per site
Ditto [16]	Global model + regularized per-site personalized model v_k
Module-wise Ditto	Ditto with a separate λ per GMIC module (global / local / fusion)

GMIC Architecture & Training

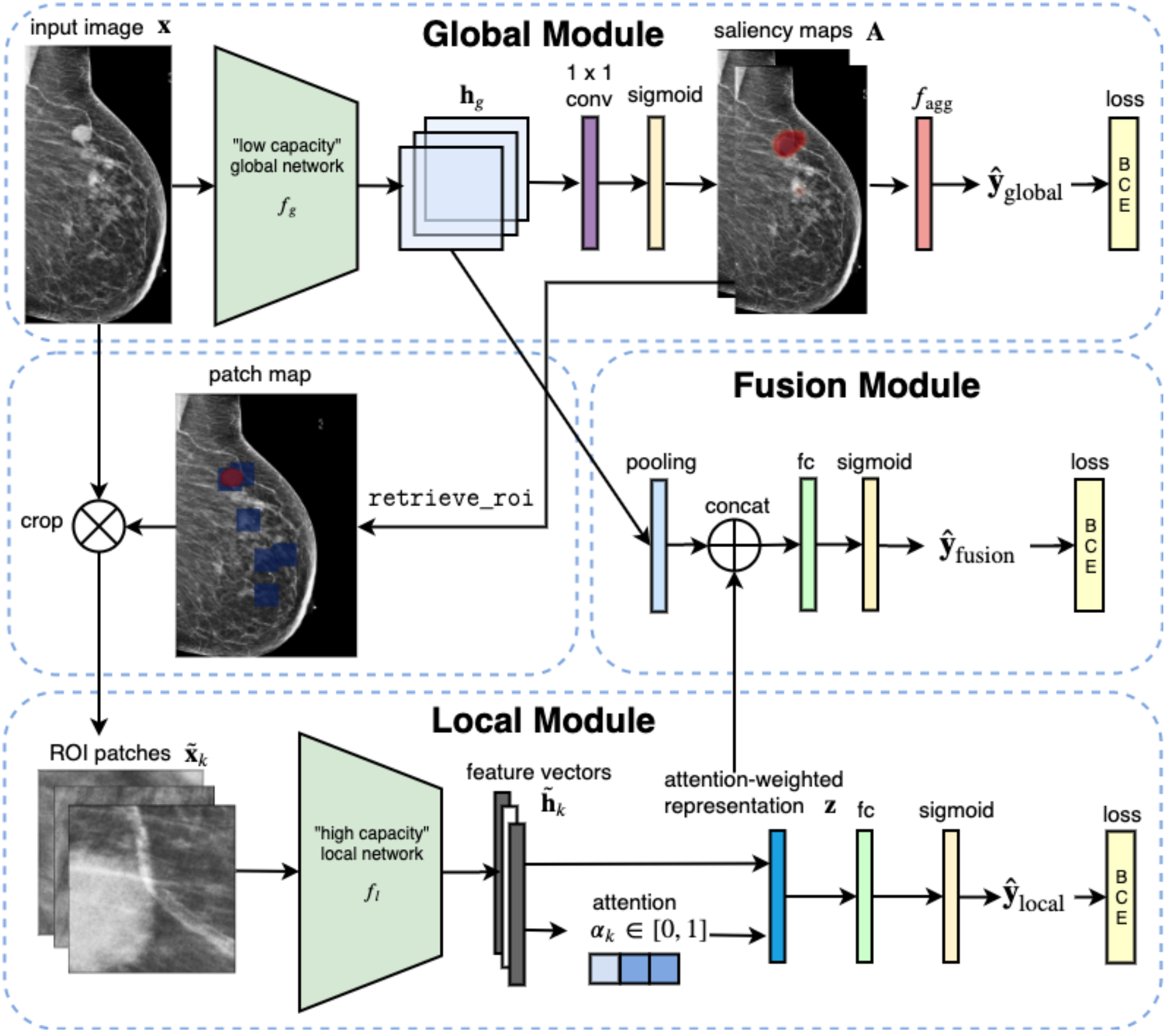


Fig. 3. GMIC's three-module architecture at 2944×1920 resolution: a global module scans the whole image for saliency, a local module examines the K=6 highest-attention ROI patches at full capacity, and a fusion module combines both. Adapted from [2,3].

- Class imbalance (7.2:1 view-level pos:neg) handled via class-balanced sampler, positive-class weight 1.0.
- Adam: lr 3×10^{-5} (heads) / 3×10^{-7} (backbone), weight decay 1×10^{-6} , 2-epoch initial backbone freeze.
- Local models: up to 40 epochs, batch size 32, early stopping on validation AUC (patience 6).
- Federated: 60 rounds $\times$ 1 local epoch/round from the same pretrained init; FedProx $\mu=0.1$.

Evaluation & Equity Metric

- Breast-level AUC, averaging CC + MLO view predictions per breast, with DeLong 95% confidence intervals [26].
- Deployed round chosen by validation AUC: pooled for shared-global models, site-specific for personalized ones.
- Equity = cross-site spread: sample SD of per-site AUC (lower = more uniform).
- Site A subgroup analysis at fixed operating points: sensitivity at 90% specificity, specificity at 90% sensitivity.

RESULTS

Personalization Improves the Worst Site

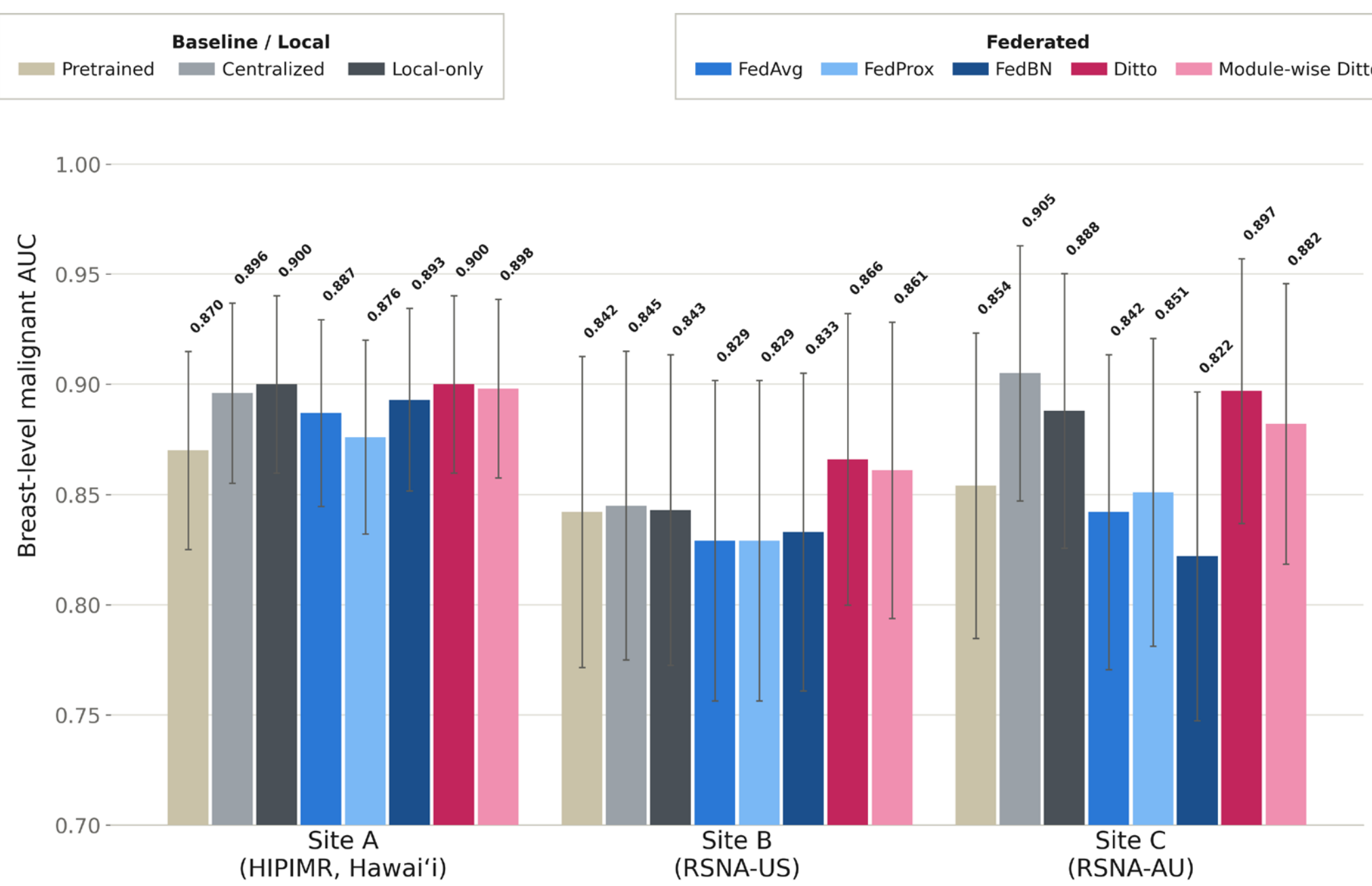


Fig. 4. Per-site breast-level AUC by strategy, with 95% CIs; Ditto and module-wise Ditto are the only federated strategies that lift Site B, the straggler, above pretrained.

Ditto and module-wise Ditto are the only federated methods that simultaneously lift Site B above the pretrained baseline, reduce cross-site spread, and match the centralized pooled ceiling's AUC (0.889 Ditto vs. 0.887 centralized).

Ditto λ Selection (Validation Sweep)

Table 3. Ditto validation-AUC sweep across regularization strength λ (12-round validation runs). All differences fall within the validation CI (≤ 0.007 AUC); $\lambda=0.05$ (highlighted) maximizes worst-site AUC and is deployed throughout.

λ	Worst-site AUC	Mean AUC
0.01	0.8556	0.8948
0.05	0.8562	0.8945
0.1	0.8560	0.8944
0.5	0.8549	0.8923
1	0.8540	0.8909
2	0.8521	0.8901

Site A Subgroup Performance

Table 4. Site A (HIPIMR) per-group AUC and Sensitivity/Specificity at fixed operating points, pretrained vs. deployed Ditto ($\lambda=0.05$); n = (group size, positive breasts). NHPI (highlighted) was the worst-served group pretrained; †small-n groups are exploratory.

Group	Pretrained AUC	Sn@Sp90	Sp@Sn90	Ditto AUC	Sn@Sp90	Sp@Sn90
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

At fixed 90% specificity, Ditto raised NHPI sensitivity 0.60 to 0.80; at fixed 90% sensitivity it raised NHPI specificity 0.39 to 0.79, the largest operating-point gain of any group. Every federated strategy raised NHPI performance by a similar margin (Supp. Table S3), so this gain reflects fine-tuning generally rather than an effect unique to Ditto's personalization; Ditto is shown because it is the deployed model with the best overall site-and-subgroup equity.

DISCUSSION & CONCLUSION

Key Takeaways

- A shared global model does not ensure equity: FedAvg, FedProx, and FedBN all left Site B at or below its pretrained performance.
- Ditto was the only strategy to simultaneously improve the worst site, reduce cross-site spread, and match centralized pooled performance, all without sharing patient data.
- Module-wise (per-module λ), Ditto showed no advantage vs scalar Ditto.
- Personalization improves equity and overall performance where standard methods fail.

Limitations & Future Work

- Single training run per method; per-site counts modest (Site B test n=396): Ditto's advantage is directionally consistent but not individually significant.
- Detailed race/ethnicity data available only at Site A, the largest site.
- All methods initialized from one pretrained checkpoint; single architecture evaluated.
- Matching the RSNA sites to Site A's 3:1 class balance also narrows one axis of real-world cross-site heterogeneity.
- Next: Federate with larger cancer centers, harmonized demographic labels at all sites, multiple initializations, unique λ per site, additional architectures, larger datasets.

Acknowledgments
Supported by NIH grants U54CA143728, R01CA263491 (HIPIMR), and P30CA071789. The authors declare no competing interests.

Contact
Lydia J. Sollis · Isollis@cc.hawaii.edu · Shepherd Research Lab, University of Hawai'i Cancer Center

References
[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); [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: Medical Learning in Medical Imaging (MLMI), LNCS, vol. 11861, pp. 18–26. Springer (2019); [4] McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, 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., Cobanaj, 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 (2024); [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); [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); [12] Gilaquinto, 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); [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-iiid 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); [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); [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); [22] Tzortzis, I.N., Gutierrez-Torre, A., Sykiotis, S., Aguiló, F., Bakalos, N., Doulami, A., Doulami, 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); [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); [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)