FedVendorBC: Accurate Multi-vendor Federated Learning for Privacy-preserving and Generalizable Breast Cancer Diagnosis using Ultrasound Images

Faisal Ahmed[1,2⋆], David Sánchez[1], Zouhair Haddi[2], and Josep Domingo-Ferrer[1]

[1] Universitat Rovira i Virgili, Dept. of Computer Engineering and Mathematics, ComSCIAM-Center for Computational Science and Applied Mathematics, CYBERCAT-Center for Cybersecurity Research of Catalonia, Tarragona, Catalonia.
[2] NVISION Systems and Technologies SL, Barcelona, Catalonia.
faisal.ahmed@nvision.es

Abstract. Manual ultrasound-based breast cancer diagnosis is labor-intensive, time-consuming, and prone to errors, a burden that AI-assisted diagnostic systems can help alleviate. However, developing accurate and generalizable diagnostic models requires large-scale multi-vendor data collected across multiple hospitals, which are difficult to obtain due to privacy concerns, data ownership restrictions, and regulatory constraints. Federated learning (FL) enables privacy-preserving and regulation-compliant collaborative training. Nevertheless, conventional FL often underperforms in multi-vendor breast ultrasound settings due to substantial data heterogeneity across vendors. To tackle this issue, we propose FedVendorBC, an FL framework tailored for accurate, privacy-preserving training of breast cancer diagnostic models across multi-vendor ultrasound imaging settings. Experiments show that FedVendorBC achieves relative improvements in diagnostic performance of 7.85% in recall, 5.55% in F1-score, 2.51% in auROC, and 12.72% in sensitivity at a 5% false-positive rate over standard baselines.

Keywords: Breast cancer diagnosis · Data privacy · Data heterogeneity · Federated learning · Multi-vendor generalization

1 Introduction

Breast cancer is a leading cause of cancer-related mortality among women, where early diagnosis significantly improves outcomes. Ultrasound (US) imaging is widely used for detection due to its cost-effectiveness and real-time capability, but manual interpretation is time-consuming and expertise-dependent. Accurate AI-assisted diagnosis from US images is thus essential to reduce this burden. However, developing accurate and generalizable deep learning models requires

⋆ Corresponding author.

large-scale, multi-institutional datasets covering diverse populations and clinical settings. Such data collection is challenging due to privacy constraints, data ownership, and regulatory restrictions.

Federated learning (FL) [1] has emerged as a promising paradigm for collaborative training of AI models across multiple institutions that does not require them to share their (sensitive) data. However, FL on multi-vendor datasets remains challenging because US images originating from different vendors are inherently heterogeneous. This heterogeneity arises from variations in image-acquisition hardware and reconstruction pipelines and is further amplified by differences in imaging protocols and diverse demographics of patients. These compound distribution shifts can significantly degrade the convergence and generalization of the model if not explicitly addressed. To tackle this issue, we propose FedVendorBC, a federated learning framework for training AI models for breast cancer diagnosis using multi-vendor US images. Empirical results conducted in a realistic multi-vendor clinical setting show that FedVendorBC offers superior performance than three baseline FL methods.

2 FedVendorBC

FedVendorBC incorporates two innovative strategies along with the vanilla FL workflow: i) it employs a gradient correction mechanism to align all parameters of local models in the same direction before aggregating them at the server, in order to overcome the gradient conflict that arises from training each local model on images of a different vendor; and ii) it regulates local training on the vendor-specific data at the client side using an early-stopping strategy guided by clinically meaningful evaluation criteria.

Local training regularization at the client level. In clinical breast ultrasound practice, benign lesions substantially outnumber malignant cases, resulting in a class imbalance. This can bias learning toward the majority (benign) class and may degrade diagnostic reliability, leading to excessive false-positive predictions that result in unnecessary biopsies, increased healthcare costs, and additional patient burden. To ensure training is performed under clinically realistic operating conditions and to promote clinically meaningful performance while mitigating overfitting, FedVendorBC regulates local training by using an early-stopping strategy based on sensitivity measured at a fixed 5% false-positive rate (FPR) on the local validation set. This low-FPR operating point aligns with practical clinical requirements, where the false-positive rate must remain limited while maintaining a high sensitivity to cancer detection.

More specifically, after each local training epoch, FedVendorBC evaluates the current model on the client's vendor-specific validation set and terminates the training if the sensitivity at 5% FPR does not improve for a predefined patience period. The model achieving the highest validation sensitivity at this operating point is retained as the final local model and is sent to the server.

Gradient correction at the server. Due to non-IID data distributions across clients –arising from US images acquired with devices from different vendors– the locally learned model parameters can diverge and conflict with one another, degrading the quality of the global model. To tackle this problem, FedVendorBC employs a gradient correction mechanism inspired by the *projecting conflicting gradients (PCGrad)* strategy [8]. FedVendorBC treats each vendor's local model as a separate task and projects a vendor's gradient onto the normal plane of the gradient of other vendors with conflicting gradients before local parameter aggregation. Specifically, in round t, FedVendorBC computes the local gradient g_t^k for each client k by determining the difference between the locally updated model parameters θ_t^k and the previous global model parameters θ_{t-1}. Then it flattens these gradient vectors into a vector form to enable pairwise gradient correction and projection. The gradient of client k is formally defined as $g_t^k = \mathrm{vec}\big(\theta_t^k - \theta_{t-1}\big)$, where $\mathrm{vec}(\cdot)$ denotes the vectorization (flattening) operator. Subsequently, for each client k, FedVendorBC randomly selects another client $j \neq k$ from the remaining $K-1$ clients and computes the inner product $\langle g_t^k, g_t^j \rangle$ between their respective gradient vectors. If the inner product is negative, that is, $\langle g_t^k, g_t^j \rangle < 0$, the two gradients are interpreted as being in conflict, indicating that they point in opposite directions. In this scenario, g_t^k is projected onto the normal plane of g_t^j to eliminate the conflicting component. The gradient is corrected as follows:

$$\bar{g}_t^k = \begin{cases} g_t^k - \dfrac{\langle g_t^k, g_t^j \rangle}{\|g_t^j\|_2^2}\, g_t^j, & \text{if } \langle g_t^k, g_t^j \rangle < 0, \\ g_t^k, & \text{otherwise.} \end{cases} \tag{1}$$

The gradient $\bar{g}_t^k$ is then reshaped back to match the parameter θ_t^k tensor structure of the target model and used to obtain the corrected local model parameter $\bar{\theta}_t^k$, as $\bar{\theta}_t^k = \theta_{t-1} + \mathrm{unvec}(\bar{g}_t^k)$, where $\mathrm{unvec}(\cdot)$ denotes the inverse vectorization operator that reshapes a vector back into the original parameter tensor structure. The server finally aggregates these corrected parameters to compute the global model update for the next training round $t+1$ using the Federated Averaging (FedAvg) [5] algorithm.

3 Experiments

Datasets and Model. We curated five datasets from public sources: *BUS-BRA* [4], *BUSI* [2], *QAMEBI* [3], *BUS-UCLM* [6] and *UDIAT* [7]. The BUS-BRA dataset contains images acquired using ultrasound systems from *GE Healthcare* and *Toshiba Medical Systems* (now Canon Medical Systems). Images in the BUSI dataset were also acquired using ultrasound systems from *GE Healthcare*. BUS-UCLM and UDIAT both include images acquired using ultrasound machines from *Siemens Healthineers*. The QAMEBI dataset comprises images captured using ultrasound system from *Supersonic Imagine*. Based on the manufacturer specifications, we organized the data into four vendor-specific datasets. We preprocessed all images using a standard transformation pipeline used in

[1]. We retrained ResNet-18 as the prediction model, modifying its input layer to accept two channels: a grayscale US image and its corresponding mask. We also replaced its final layer with a single output neuron for binary classification between benign and malignant.

Results and Discussion. We conducted leave-one-vendor-out validation to evaluate the effectiveness of FedVendorBC and compared it against three standard FL baselines (FedAvg, FedProx, and SCAFFOLD) using the following evaluation metrics: accuracy, precision, recall, sensitivity at 5% false-positive rate (S-5-F), F1-score, and area under the receiver operating characteristic curve (auROC). The results for four unseen external vendors across all metrics, as well as the average results across the vendors, are presented in Table 1.

Table 1. Performance of FedVendorBC and baselines on external vendor datasets. Averages are computed over per-vendor point estimates; best result per metric in bold.

Vendor	Method	Accuracy	Precision	Sensitivity	Specificity	S-5-F	F1-score	auROC
GE	FedVendorBC	**87.16**	**76.53**	**85.31**	**88.00**	**72.63**	**0.807**	**0.929**
	FedAvg	84.77	72.18	83.85	85.19	62.62	0.776	0.915
	FedProx	84.73	73.62	80.11	86.84	60.61	0.767	0.909
	SCAFFOLD	84.47	70.62	86.65	83.48	61.95	0.778	0.912
Siemens	FedVendorBC	**78.59**	64.96	**86.36**	74.29	**37.50**	**0.742**	**0.852**
	FedAvg	73.33	58.94	82.39	68.34	31.25	0.687	0.817
	FedProx	77.17	66.15	73.30	79.31	37.50	0.695	0.832
	SCAFFOLD	78.59	**69.23**	71.59	**82.45**	33.52	0.704	0.816
Supersonic	FedVendorBC	**81.90**	80.92	**86.18**	77.06	**51.22**	**0.835**	**0.872**
	FedAvg	78.88	83.64	74.80	**83.49**	42.28	0.790	0.866
	FedProx	79.74	**83.93**	76.42	**83.49**	**51.22**	0.800	0.868
	SCAFFOLD	79.74	83.33	77.24	82.57	41.46	0.802	0.839
Toshiba	FedVendorBC	**89.21**	90.77	86.76	91.55	**75.00**	**0.887**	**0.944**
	FedAvg	84.89	**94.34**	73.53	**95.77**	73.53	0.826	0.903
	FedProx	82.01	77.22	**89.71**	74.65	57.35	0.830	0.899
	SCAFFOLD	82.01	82.09	80.88	83.10	33.82	0.815	0.836
Avg±SD	FedVendorBC	**84.22 ± 4.85**	**78.30 ± 10.70**	**86.15 ± 0.61**	82.72 ± 8.35	**59.09 ± 17.93**	**0.818 ± 0.060**	**0.899 ± 0.044**
	FedAvg	80.47 ± 5.52	77.28 ± 15.21	78.64 ± 5.23	**83.20 ± 11.30**	52.42 ± 19.15	0.770 ± 0.059	0.875 ± 0.044
	FedProx	80.91 ± 3.22	75.23 ± 7.43	79.88 ± 7.10	81.07 ± 5.22	51.67 ± 10.10	0.773 ± 0.058	0.877 ± 0.034
	SCAFFOLD	81.20 ± 2.67	76.32 ± 7.43	79.09 ± 6.25	82.90 ± 0.47	42.69 ± 13.47	0.775 ± 0.049	0.851 ± 0.042

FedVendorBC shows a substantial generalization improvement over FedAvg across all evaluation metrics except specificity, without incurring additional costs in terms of communication efficiency and bandwidth usage. Performance improvements are more substantial for sensitivity, F1-score, auROC and sensitivity at a fixed 5% false-positive rate, which are clinically significant metrics for breast cancer diagnosis. Compared to standard baselines, FedVendorBC achieves relative performance gains of 7.85% in recall, 5.55% in F1-score, 2.51% in auROC, and 12.72% in sensitivity at a 5% false-positive rate. These gains indicate improved cancer-detection performance under clinically realistic low-false-positive operating conditions. A vendor-specific analysis shows that GE HealthCare and Siemens Healthineers, whose datasets exhibit a more realistic ratio of malignant to benign samples, demonstrate consistently superior diagnostic performance across all evaluation metrics compared to FedAvg. The data distributions for these two vendors closely reflect real-world clinical practice, where benign lesions substantially outnumber malignant cases, further highlighting the practical rele-

vance of FedVendorBC. In contrast, for SuperSonic Imagine and Toshiba Medical Systems, FedVendorBC shows a relatively lower specificity, as also reflected in the overall average specificity score. This reduced specificity can be attributed to the atypical class distributions in these datasets, where the balance between malignant and benign samples deviates from realistic clinical prevalence.

4 Conclusions

FedVendorBC addresses the data heterogeneity that arises in collaborative trained AI models using multi-vendor US imaging systems for breast cancer diagnosis. By explicitly accounting for inter-vendor heterogeneity, FedVendorBC enables robust and stable learning across hospitals operating scanners from different manufacturers. Future work will extend FedVendorBC to additional imaging modalities and cancer types, and include an ablation study to identify which component drives the performance gains.

Acknowledgments. This work is funded by the European Commission (GA 101073222), the Government of Catalonia (ICREA Acadèmia Prizes to J. Domingo-Ferrer and D. Sánchez), and MCIN/AEI grant PID2024-157271NB-I00 "CLEARING-IT".

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Ahmed, F., Sánchez, D., Domingo-Ferrer, J., Haddi, Z.: Fedbc: Privacy-preserving breast cancer diagnosis from ultrasound images using federated learning. In: Deep Breast Workshop on AI and Imaging for Diagnostic and Treatment Challenges in Breast Care. pp. 31–40. Springer (2025)
2. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in brief **28**, 104863 (2020)
3. Ardakani, A.A., Mohammadi, A., Mirza-Aghazadeh-Attari, M., Acharya, U.R.: An open-access breast lesion ultrasound image database: Applicable in artificial intelligence studies. Computers in Biology and Medicine **152**, 106438 (2023)
4. Gómez-Flores, W., Gregorio-Calas, M.J., Coelho de Albuquerque Pereira, W.: Busbra: a breast ultrasound dataset for assessing computer-aided diagnosis systems. Medical Physics **51**(4), 3110–3123 (2024)
5. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282. PMLR (2017)
6. Vallez, N., Bueno, G., Deniz, O., et al.: Bus-uclm: Breast ultrasound lesion segmentation dataset. Scientific Data **12**, 242 (2025)
7. Yap, M.H., Pons, G., Marti, J., Ganau, S., Sentis, M., Zwiggelaar, R., Davison, A.K., Marti, R.: Automated breast ultrasound lesions detection using convolutional neural networks. IEEE journal of biomedical and health informatics **22**(4), 1218–1226 (2017)
8. Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., Finn, C.: Gradient surgery for multi-task learning. Advances in neural information processing systems **33**, 5824–5836 (2020)