FedBC: Privacy-preserving breast cancer diagnosis from ultrasound images using federated learning

Faisal Ahmed[1,2][0000−0002−7601−3962], David Sánchez[1][0000−0001−7275−7887], Josep Domingo-Ferrer[1][0000−0001−7213−4962], and Zouhair Haddi[2][0000−0001−9835−4795]

[1] Universitat Rovira i Virgili, Dept. of Computer Engineering and Mathematics, CYBERCAT-Center for Cybersecurity Research of Catalonia, Tarragona, Catalonia.
{david.sanchez,josep.domingo}@urv.cat
[2] NVISION Systems and Technologies SL, Barcelona, Catalonia.
{faisal.ahmed,zouhair.haddi}@nvision.es

Abstract. Accurate diagnosis of breast cancer in dense breasts requires expert radiologists to examine multiple ultrasound images per patient. This diagnosis procedure is tedious, time-consuming and prone to misdiagnosis due to human fatigue. AI-aided diagnosis systems can help alleviate this burden. However, vast amounts of data from multiple hospitals, diverse patient demographics, imaging scanners, and protocols are required to develop accurate, robust, and generalizable AI models. Obtaining such a mixture of data is quite challenging due to privacy concerns, data ownership issues, and regulatory constraints. To address this problem, we introduce a privacy-preserving and regulatory-compliant method for breast cancer diagnosis from ultrasound images using federated learning. This allows hospitals to keep their data on their premises while collaboratively training the AI model, only exchanging the model parameters trained on their private data. Experimental evaluation within a real federated data setting shows that our method achieved higher diagnostic accuracy than standardized centralized training while ensuring compliance with data privacy regulations.

Keywords: Breast cancer diagnosis · Federated learning · Data privacy

1 Introduction

Breast cancer causes the highest mortality among women worldwide among all cancer-related deaths. Early diagnosis of breast cancer significantly improves patient outcomes and minimizes the risk of mortality [21]. Ultrasound (US) imaging is a widely used modality for the detection and diagnosis of breast cancer, particularly in resource-constrained settings, due to its cost-effectiveness, non-invasiveness, and ability to provide real-time imaging [20]. However, manual breast cancer diagnosis from US images is time-consuming and requires expert radiologists. In addition, with the increasing number of cases of breast cancer, the

volume of US images that radiologists must interpret also increases, which can lead to diagnostic delays and errors. Therefore, a precise AI-assisted diagnosis of breast cancer from US images is essential to support medical decision-making.

Recent advances in deep learning have shown significant success in automating medical image analysis for diagnosis [19,17,23] and prediction [12,3,6]. However, deep learning requires vast amounts of data from multiple hospitals, imaging scanners, and diverse patient demographics to develop accurate, robust, and generalizable models. Such a wealth of data is often difficult to obtain due to privacy concerns, data ownership issues, and regulatory restrictions such as the Health Insurance Portability and Accountability Act (HIPAA) [1] and the EU General Data Protection Regulation (GDPR) [2]. Moreover, processing large volumes of data in a single institution requires a high-performance computing infrastructure, which is costly and often unavailable in resource-constrained settings, such as those in developing countries.

Federated learning (FL) [16] has emerged as a promising paradigm for training collaborative models between multiple institutions without the need to share data directly. By enabling decentralized learning, FL intrinsically preserves data privacy, distributes computational load across multiple institutions or edge devices, and leverages multi-institutional datasets to enhance model generalization. In our setting, FL can help overcome the limitations associated with developing deep learning models for the diagnosis of breast cancer using US images.

To our knowledge, the feasibility of applying FL for breast cancer diagnosis using US images has not yet been explored. Our study aims to address this research gap through the following key contributions:

1. We propose a privacy-preserving and regulatory-compliant framework that leverages FL to diagnose breast cancer from US images. Our framework enhances model generalizability by integrating multi-institutional datasets encompassing diverse patient demographics, imaging protocols, and US scanners from different vendors.
2. We curate multiple data sets that feature variations in imaging protocols, patient demographics, number of images, and US systems, reflecting real-world data heterogeneity between different hospitals.
3. We report empirical results that compare our FL setting with conventional centralized learning (CL).

The remainder of the paper is organized as follows. Section 2 discusses related work on AI-based breast cancer diagnosis. Section 3 describes our method for the training of FL-based breast cancer models. Section 4 details the datasets and experimental setup, while Section 5 reports and discusses the results obtained. Conclusions and several lines of future research are described in Section 6.

2 Related work

Several solutions have been proposed to automate the diagnosis of breast cancer from US images. Most of them operate in a centralized setting, where a single computing server aggregates all images from different sources and performs

model training and validation. Most of these solutions follow a common workflow comprising three steps: *image preprocessing*, *lesion segmentation*, and *lesion classification*.

Preprocessing techniques include: (1) *data augmentation* to mitigate class imbalance and improve model generalization; (2) *simple linear iterative clustering* [11] to address complex artifacts and noise; (3) *anisotropic diffusion* [8], *interference-based despeckling* [8], and *bilateral filtering* [24] to reduce noise and speckle noise while preserving edges and fine details; (4) *morphological opening and closing operations* [24] on the original image using the binary mask to refine tumor boundaries; (5) *sigmoid filtering* [8] and *fuzzy enhancement* [24] to improve image quality and contrast, making lesions more distinguishable from surrounding tissues. For lesion segmentation, related works use U-Net [15], fully convolutional networks [14] or UNet variants [22,11]. For classification, [11] and [15] adopted a deep transfer learning approach, with the former fine-tuning a pre-trained VGG19 model and the latter fine-tuning a pre-trained ResNet50 model. On the other hand, three studies [8,14,22] train a wavelet neural network, a custom convolutional neural network (CNN), and a CNN with Inception blocks and an attention mechanism, respectively, from scratch. Few studies have also explored traditional machine learning techniques. [8] utilized *wavelet neural networks*, along with manually extracted morphological and texture features. Stacking ensembles of multi-layer perceptrons and support vector machines, combined with transfer learning features of segmented lesions obtained from pre-trained VGG16, VGG19, and InceptionV3 models, were used in [8].

The works discussed above have several limitations: (1) They require aggregating patients' US images in a central computing platform from multiple data sources (*i.e.*, hospitals), potentially compromising patient privacy and violating regulatory compliance requirements such as HIPAA, and GDPR; (2) The models are trained and validated on single or limited source data, and their performance across diverse or out-of-domain datasets and varying imaging conditions has not been extensively evaluated; (3) They rely on multiple models for segmentation and classification along with extensive preprocessing, which is computationally expensive and potentially limits the pipeline's applicability in resource-constrained environments and edge devices.

3 Methods

Figure 1 depicts the privacy-preserving training setup for our FL-based breast cancer diagnosis model. Each hospital retains its private dataset of breast US images, corresponding masks, and labels. They perform model updates locally and share the updated model parameters with a server, which coordinates the model's training across all hospitals over several communication rounds.

In the communication round t, the server distributes the global model parameters θ_t to all participating hospitals. Each hospital initializes its local model with θ_t as a starting point and updates it using its private data, producing local models with parameters $\theta_t^1, \theta_t^2, \ldots, \theta_t^K$. These updated parameters are sent back

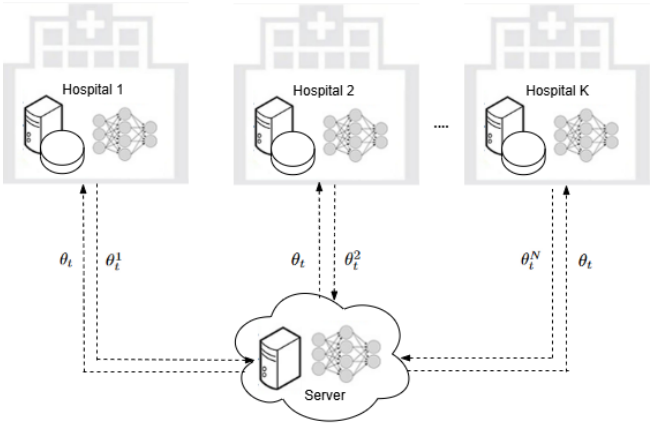


Fig. 1. FL framework for privacy-preserving breast cancer diagnosis using US images

to the server, which aggregates them to refine the global model. This process repeats over multiple rounds until convergence. Only model updates are exchanged between the server and hospitals, thereby ensuring that no raw data are shared.

Model architecture for breast cancer diagnosis. We utilized the ResNet-18 [10] architecture as a breast cancer prediction model. The input layer of the network was modified to handle a two-channel input $\mathbf{x} \in \mathbb{R}^{H \times W \times 2}$ –consisting of a grayscale breast US image and its corresponding mask, both with height H and width W. In a similar manner, the final layer of the network was restructured to perform binary classification by replacing it with a single output neuron. This neuron employs a sigmoid activation function, which maps the network's output to a probability value between 0 and 1, enabling the model to distinguish between two classes –*benign* and *malignant*– based on the predicted probability.

Model parameter update at hospital premises. The local training objective of hospital k at training round t is to update the global model parameter θ_t, which maps input data X to target outputs Y in the hospital's private dataset while minimizing a given loss function $\mathcal{L}(\theta_t)$. Formally, for the private dataset $D_k = \{(x_i, y_i)\}_{i=1}^{n^k}$ of hospital k, the optimization objective is

$$\theta_t^k = \arg\min_{\theta_t} \mathbb{E}_{(x,y)\sim D_k} \mathcal{L}(\theta_t(x), y), \tag{1}$$

where $\theta_t(x)$ is the predicted output of the model, y is the ground truth label, $\mathcal{L}$ is the loss function and $\mathbb{E}$ denotes the expectation over the distribution of the data of hospital k.

Since breast cancer diagnosis is a binary classification problem and the model architecture employs a sigmoid activation function in the classification layer, we use binary cross-entropy as the loss function $\mathcal{L}$. For the dataset of hospital k

with n_k samples, where $\mathbf{x}_i$ is the i-th input and $y_i \in \{0, 1\}$ is its corresponding label, the cross-entropy loss for a single sample, given model parameter θ_t, is:

$$\mathcal{L}_{\text{CE}}(\theta_t, \mathbf{x}_i, y_i) = -\left[y_i \log p_i + (1 - y_i) \log(1 - p_i)\right], \tag{2}$$

where $p_i = \sigma(\theta_t^T \mathbf{x}_i)$, and $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function.

To prevent overfitting, the loss function incorporates an L2 regularization term that penalizes large parameter values. The final loss function over n_k samples of hospital k combines the classification loss and the regularization term, defined as

$$\mathcal{L}(\theta_t, \mathbf{x}_i, y_i) = \underbrace{\frac{1}{n_k} \sum_{i=1}^{n_k} \mathcal{L}_{\text{CE}}(\theta_t, \mathbf{x}_i, y_i)}_{\text{classification loss}} + \underbrace{\frac{\lambda}{2} \|\theta_t\|_2^2}_{\text{regularization term}}, \tag{3}$$

where λ controls the regularization strength.

The loss $\mathcal{L}$ is minimized, and the model parameter is updated using stochastic gradient descent as follows:

$$\theta_t^k = \theta_t - \eta \nabla \mathcal{L}(\theta_t; D_k), \tag{4}$$

where $\mathcal{L}(\theta_t; D_k)$ is the regularized binary cross-entropy loss function in Eq. (3), $\nabla \mathcal{L}(\theta_t; D_k)$ is the gradient computed on the local dataset D_k, and η is the local learning rate.

Aggregation of parameters at the server. After performing multiple local updates within their premises during training round t, hospitals transmit their updated model parameters, θ_t^k, to the server. The server then aggregates these parameters to compute the global model update for the federated training round $t + 1$ using the Federated Averaging (FedAvg) [16] algorithm as follows:

$$\theta_{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} \theta_t^k, \tag{5}$$

where n_k represents the number of data samples held by hospital k, and n denotes the total number of samples across all participating hospitals.

4 Experimental setup

Datasets. We curated three datasets from public sources: *BUS-BRA* [9], *BUSI* [5], and *QAMEBI* [7]. These datasets were collected from hospitals in various regions of the world, using different US systems at different times and under varying image acquisition protocols. We further partitioned the datasets into federated hospitals or domains based on ultrasound system specifications, resulting in six distinct groups, referred to as hospitals A through F in Table 1.

All the images were preprocessed using a standard transformation pipeline. Each image was first resized to a fixed resolution of 128×128 pixels to ensure uniform input dimensions. The intensity values of the resized images were then normalized to the range $[-1, 1]$, using a mean of 0.5 and a standard deviation of 0.5.

Table 1. Features of the federated datasets

Hospital	Source	US Machine	Frequency (MHz)	# Patients	# Images
A	BUS-BRA [9]	Toshiba Aplio 300	12–14	72	902
B	BUS-BRA [9]	GE Logiq 5	10–12	490	1064
C	BUS-BRA [9]	GE Logiq 7	10–14	488	809
D	BUS-BRA [9]	GE U-Systems	12–14	14	25
E	BUSI [5]	LOGIQ E & LOGIQ E9 Agile	1–5	-	697
F	QAMEBI [7]	AirPlorer Ultimate	5–18	-	232

Implementation details and evaluation metrics. Model training was conducted at each hospital site using a mini-batch gradient descent optimizer with a batch size of 64 image samples. The initial learning rate was set at 0.001 and a StepLR scheduler was used to adjust the learning rate during training. Each hospital performed local training for 10 epochs per communication round, and FL was executed over 20 communication rounds.

The hardware setup consisted of a remote server equipped with an Intel Xeon processor, 64GB of RAM, and an NVIDIA A10 GPU with 24GB of VRAM. The software environment included the Ubuntu 22.04.3 LTS operating system, CUDA 12.2, and NVIDIA driver version 570.124.06. Models were implemented and trained using PyTorch 1.9.0, while Python 3.10.12 was used to develop the FL pipelines. We employed five classification metrics for performance evaluation: *accuracy*, *precision*, *recall*, *F1-score*, and the *area under the ROC curve* (auROC).

5 Results and discussion

Firstly, we investigate how FL enables each hospital to develop a more generalized model that is specifically tailored to its local data distribution while benefiting from the collective knowledge of the data of the other participating hospitals. To demonstrate this benefit, we compare the performance of individually and independently trained models on each hospital's data with that of a global model jointly trained by all hospitals using FL and evaluated on each hospital's test data. More specifically, we split each hospital's dataset into training and testing sets with an 80:20 ratio. An individual model is trained solely on each hospital's training set and evaluated on its corresponding test set. On the other hand, FL is performed collaboratively by using the training sets from

all hospitals. The resulting global model is then individually evaluated in each hospital's test set.

Table 2. FL vs individual training performance across hospitals. Best figures in bold.

Hospital	Training type	Accuracy↑	Precision↑	Recall↑	F1-Score↑	auROC↑
A	Federated	**89.29**	**0.85**	**0.92**	**0.88**	**0.91**
	Individual	82.72	0.73	0.64	0.68	0.86
B	Federated	**89.51**	**0.86**	**0.77**	**0.81**	**0.93**
	Individual	82.72	0.73	0.64	0.68	0.86
C	Federated	**81.22**	0.70	**0.85**	**0.76**	0.88
	Individual	79.56	**0.82**	0.55	0.66	**0.90**
D	Federated	**100.00**	**1.00**	**1.00**	**1.00**	**1.00**
	Individual	60.00	0.50	0.50	0.50	0.67
E	Federated	**97.69**	**1.00**	**0.93**	**0.96**	0.98
	Individual	96.95	1.00	0.88	0.94	**1.00**

The results of this evaluation are presented in Table 2, and show that FL consistently outperforms individual training on all metrics in all hospitals. This clearly indicates that federated training improves generalization and robustness to hospital-specific data distributions. This result could be expected, as each hospital trains a model solely on its local data in the individual training setting, which is limited in both size and diversity and may lead to underfitting, overfitting, or poor generalization. In contrast, FL allows hospitals to learn from each other's data while preserving their privacy. This enables training on a larger, more diverse dataset than any single hospital could access, improving parameter estimation and reducing prediction variance. In this sense, model aggregation acts like an ensemble, further boosting model generalization. Together, these factors enhance the model's performance on hospital-specific data distributions.

Secondly, we compare the leave-one-domain-out validation results of FL with those of CL. For this experiment, data from five out of six domains (A to E) are used to simulate five federated hospitals, while the remaining domain (F) serves as an unseen test set to evaluate the global model produced by FL. To enable a fair comparison with CL, data from the same five domains are aggregated into a single combined dataset to train a standalone model. The held-out domain is then used exclusively for evaluation, mirroring the setup used in the FL scenario. This process is repeated six times, with each domain taking a turn as the unseen test set. The remaining five domains are used to train the same model in both FL and CL settings. After all six iterations, performance metrics (*e.g.*, accuracy, precision, recall, F1-score and auROC) are averaged to provide a robust estimate of the generalization performance of each approach. Table 3 reports the results on six unseen sites in terms of five metrics.

The FL approach achieves comparable CL performance on average across all evaluation metrics while offering significant additional benefits: it enhances data privacy, eliminates the demanding need to aggregate data in a central repository, and reduces the demand for highly centralized computing resources by distributing the computational load for model training across participating hospitals. In general, these results demonstrate that FL is a viable solution for developing deep learning models for breast cancer diagnosis in US images.

Table 3. Performance metrics on held-out of domains. Best average figures in bold.

Unseen domain	Setting	Accuracy↑	Precision↑	Recall↑	F1-Score↑	auROC↑
A	FL	82.73	0.81	0.85	0.83	0.91
	CL	80.58	0.83	0.80	0.80	0.92
B	FL	87.76	0.87	0.70	0.77	0.93
	CL	86.53	0.86	0.80	0.82	0.92
C	FL	80.82	0.67	0.78	0.72	0.88
	CL	76.27	0.76	0.79	0.75	0.88
D	FL	100.00	1.00	1.00	1.00	1.00
	CL	100.00	1.00	1.00	1.00	1.00
E	FL	89.95	0.83	0.87	0.85	0.96
	CL	89.34	0.88	0.87	0.87	0.94
F	FL	78.88	0.78	0.83	0.81	0.85
	CL	78.02	0.78	0.77	0.77	0.84
Average	FL	**86.69**	0.83	**0.84**	**0.83**	**0.92**
	CL	85.12	**0.85**	**0.84**	**0.83**	**0.92**

6 Conclusions

We proposed a privacy-preserving and regulatory-compliant framework that leverages FL for breast cancer diagnosis from US images guided by lesion masks. Empirical results and comparative analyses against traditional CL and individual training approaches demonstrate the framework's ability to support multi-institutional collaboration in training a highly generalized model for breast cancer diagnosis. FL also overcomes the reliance on a single high-power computing platform, making it well suited for resource-constrained settings.

In future research, we will extend this framework to address data heterogeneity [13] in different hospitals caused by variations in image acquisition protocols, US systems, patient demographics, and number of images. We also plan to incorporate the MemberShield [4] method to further enhance the robustness of the framework against membership inference attacks [18].

Acknowledgments. Partial support for this work has been received from Marie Skodowska-Curie Doctoral Networks Actions (HORIZON-MSCA-2021-DN-01-01; 101073222), the Government of Catalonia (ICREA Acadèmia Prizes to J. Domingo-Ferrer and to D. Sánchez, and grant 2021SGR-00115), MCIN/AEI/ 10.13039/501100011033 and "ERDF A way of making Europe" under grant PID2021-123637NB-I00 "CURLING", and the EU's NextGenerationEU/PRTR via INCIBE (project "HERMES" and INCIBE-URV cybersecurity chair).

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

References

1. Health insurance portability and accountability act of 1996 (1996)
2. Regulation (EU) 2016/679 of the European parliament and of the council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (2016)
3. Afshar, P., Mohammadi, A., Plataniotis, K.N., Oikonomou, A., Benali, H.: From handcrafted to deep-learning-based cancer radiomics: challenges and opportunities. IEEE Signal Processing Magazine **36**(4), 132–160 (2019)
4. Ahmed, F., Sánchez, D., Haddi, Z., Domingo-Ferrer, J.: Membershield: A framework for federated learning with membership privacy. Neural Networks **181**, 106768 (2025)
5. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in brief **28**, 104863 (2020)
6. Allugunti, V.R.: Breast cancer detection based on thermographic images using machine learning and deep learning algorithms. International Journal of Engineering in Computer Science **4**(1), 49–56 (2022)
7. 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)
8. Bourouis, S., Band, S.S., Mosavi, A., Agrawal, S., Hamdi, M.: Meta-heuristic algorithm-tuned neural network for breast cancer diagnosis using ultrasound images. Frontiers in Oncology **12**, 834028 (2022)
9. 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)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
11. Inan, M.S.K., Alam, F.I., Hasan, R.: Deep integrated pipeline of segmentation guided classification of breast cancer from ultrasound images. Biomedical Signal Processing and Control **75**, 103553 (2022)
12. Khan, S., Islam, N., Jan, Z., Din, I.U., Rodrigues, J.J.C.: A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. Pattern Recognition Letters **125**, 1–6 (2019)
13. Li, Q., Diao, Y., Chen, Q., He, B.: Federated learning on non-iid data silos: An experimental study. In: 2022 IEEE 38th international conference on data engineering (ICDE). pp. 965–978. IEEE (2022)

14. Long, B., Guan, Y., Holden, M.: A two-stage neural network model for breast ultrasound image classification. In: 2023 IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE). pp. 129–133. IEEE (2023)
15. Luo, Y., Huang, Q., Li, X.: Segmentation information with attention integration for classification of breast tumor in ultrasound image. Pattern Recognition **124**, 108427 (2022)
16. 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)
17. Sethi, R., Mehrotra, M., Sethi, D.: Deep learning based diagnosis recommendation for covid-19 using chest x-rays images. In: 2020 Second international conference on inventive research in computing applications (ICIRCA). pp. 1–4. IEEE (2020)
18. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: 2017 IEEE symposium on security and privacy (SP). pp. 3–18. IEEE (2017)
19. Sivaranjini, S., Sujatha, C.: Deep learning based diagnosis of parkinson's disease using convolutional neural network. Multimedia tools and applications **79**(21), 15467–15479 (2020)
20. Thigpen, D., Kappler, A., Brem, R.: The role of ultrasound in screening dense breasts—a review of the literature and practical solutions for implementation. Diagnostics **8**(1), 20 (2018)
21. Wheeler, S.B., Rocque, G., Basch, E.: Benefits of breast cancer screening and treatment on mortality. Jama **331**(3), 199–200 (2024)
22. Yi, S., Chen, Z., Yi, L., She, F.: Cas: Breast cancer diagnosis framework based on lesion region recognition in ultrasound images. Journal of King Saud University-Computer and Information Sciences **35**(8), 101707 (2023)
23. Zhang, T., Tan, T., Samperna, R., Li, Z., Gao, Y., Wang, X., Han, L., Yu, Q., Beets-Tan, R.G., Mann, R.M.: Radiomics and artificial intelligence in breast imaging: a survey. Artificial Intelligence Review **56**(Suppl 1), 857–892 (2023)
24. Zhuang, Z., Yang, Z., Raj, A.N.J., Wei, C., Jin, P., Zhuang, S.: Breast ultrasound tumor image classification using image decomposition and fusion based on adaptive multi-model spatial feature fusion. Computer methods and programs in biomedicine **208**, 106221 (2021)