ODELIA

Open Consortium for Decentralized Medical Artificial Intelligence

HORIZON-HLTH-2021-CARE-05-02

Deliverable D2.1

REPORT COMPARING THE PERFORMANCE OF LOCALLY TRAINED MODELS AGAINST SWARM-TRAINED MODELS IN BREAST CANCER SCREENING

Lead beneficiary	University of Cambridge (CAM)
Author(s)	CAM: Fiona Gilbert, Nick Payne, Lorena Escudero UKA: Gustav Müller-Franzes, Debora Jutz TUD: Oliver Saldanha
Reviewer(s)	RSH: Julia Camps, Maria Mercedes Gozalbo Mestres, Beatriz Benito Robles
Dissemination level	PU
Type	R
Delivery date	19 December 2024

ODELIA is funded by the European Union's Horizon Europe Framework under Grant Agreement 101057091

Funded by the European Union

TABLE OF CONTENTS

DISCLAIMER

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them

EXECUTIVE SUMMARY

Work Package 2.2 (WP2.2) of the ODELIA project aimed to develop AI models for breast cancer detection using MRI, demonstrating the potential of swarm learning (SL) to improve model performance across diverse clinical environments. The study revealed significant variability in performance of locally trained models due to site-specific factors, but SL training improved results for most sites, highlighting its potential to address cross-site variability. Despite challenges, such as dataset size limitations and the need for further exploration of model architectures, the collaborative approach of SL shows promise for enhancing breast cancer diagnosis in multi-institutional settings.

INTRODUCTION

Work Package 2.2 (WP2.2) of the ODELIA project focuses on demonstrating the clinical applicability of swarm learning (SL) for breast cancer detection using MRI. The primary objective of WP2.2 is to develop, train, and evaluate artificial intelligence (AI) models capable of accurately identifying breast cancer in MRI data based on case-level annotations. These models aim to enhance breast cancer screening and diagnosis while ensuring consistent performance across diverse clinical environments.

This report summarizes the outcomes of WP2.2, where each partner institution trained its own local model for breast cancer detection in MRI data. These models were developed using locally available datasets, reflecting the real-world variability in MRI acquisition protocols and patient demographics. Additionally, the locally trained models were exchanged in a standardized format, enabling each model to be tested on the local test sets from all other institutions.

Rather than developing independent model architectures, a standardized model framework established at UKA was adopted to ensure compatibility and facilitate the aggregation of results across institutions. This approach was chosen, as different model architectures would have hindered effective comparison of findings without considering the difference in model architectures.

The datasets used for training by each partner were also integrated into ODELIA's swarm learning framework. The SL training has been simulated at UKA to verify the accuracy and reliability of the developed code.

METHODS

Study Design

This study employed a multi-center design to collect breast MRI data from eight distinct research institutions. Each site adhered to institutional ethical guidelines, and the study protocol was approved by the ethics committee of each participating institution.

Image Acquisition

Breast MRI data were collected from eight collaborating research institutions across Europe (**Table 1**), comprising a total of 1,013 patients. Imaging parameters, such as resolution and spacing, varied across institutions (**Table 2**). The median in-plane resolution was 512 × 448 pixels, with an average of 112 slices per scan.

Image Preprocessing

Preprocessing was standardized across sites. Subtraction images were generated by computing the difference between the first post-contrast image and the pre-contrast image. All images were reoriented to a canonical orientation and resampled to a uniform voxel spacing of 0.7 × 0.7 × 3.0 mm. Subsequently, images were separated into left and right breasts and cropped or padded to a resolution of 256 × 256 × 32 pixels. Signal intensities were clipped to the 99th percentile, then z-normalized by subtracting the mean and dividing by the standard deviation.

Image Labelling

Radiologists from each institution labelled the imaging data according to the following categories:

- No Lesion: No lesions with contrast enhancement
- Benign Lesion: Lesion with contrast enhancement but confirmed to be benign either by biopsy or follow-up.
- Malignant Lesion (DCIS): Ductal Carcinoma in Situ
- Malignant Lesion (Invasive): Invasive Carcinoma
- Malignant Lesion (unknown): Malignant lesion of unknown specific type

For cases where both benign and malignant lesions were present, the breast was labelled as malignant. For this preliminary study, we simplified the classification to differentiate between cancerous and non-cancerous cases due to the limited number of samples.

Model Training

Each site divided the unilateral breasts into a training-validation and a test set with an 80% to 20% split, stratified by label distribution and grouped by patients. The training-validation set was further partitioned into 80% for training and 20% for validation. This results into a final training split of 64%, validation split of 16% and test set of 20% of the collected data.

To ensure consistency, a ResNet18(1) model with three-dimensional convolutional kernels was selected as the baseline model, implemented using the MONAI library(2). Training was conducted using a batch size of 2 with input resolution set to 224 × 224 × 32 voxels, utilizing 16-bit precision to accelerate training. Weighted sampling addressed class imbalances, while data augmentation techniques, including random flipping, Gaussian noise addition, and random rotations, mitigated overfitting. Data augmentation was implemented using the TorchIO library(3).

Early stopping was employed, terminating training if validation AUC values did not improve for 50 consecutive epochs. The model checkpoint with the highest validation AUC was selected for testing. Optimization was performed using the AdamW optimizer(4), with a learning rate of 1e-4 and a weight decay of 1e-2. Cross-entropy loss was used as the loss function.

Swarm Learning

The swarm learning framework is a pivotal component of our collaborative effort to enhance breast cancer detection through AI. Partners from TUD together with UKA have established a dedicated GitHub repository that houses the SL code, which has been set up and is currently operational at each participating institution. The first objective of our study was to evaluate whether weakly supervised DL workflows can effectively detect breast cancer in MRI data using only one per-volume label for each patient. To this end, we carried out weakly supervised prediction workflows on 3D-CNN models (**Figure 4A**). Given that data sharing typically presents a major hurdle in training radiology image analysis pipelines, we hypothesized that SL could alleviate this problem by keeping the dataset distributed throughout different partners. Therefore, we first simulated an SL setup (**Figure 4B**) with three nodes set up in one laboratory, each controlling 40%, 30%, and 10% of the data, respectively. Second, we conducted real-world SL experiments, trained on multicentric data, and externally validated them on two test cohorts (**Figure 4C**).

Due to varying levels of local expertise, not all institutions are yet fully familiar with how to effectively run the SL processes in all centres. So we conducted the results on only selected centers for the real world swarm learning study.

To address this challenge and facilitate a smoother transition towards full implementation, we initiated a simulation of the SL framework at UKA. In this setup, we consolidated datasets from all participating institutions—excluding those from VHIO and USZ—and created distinct SL instances on our machine for each institution. Each instance was trained using its respective local dataset, allowing us to test the efficacy of our SL approach while providing valuable insights into model performance across different data sources.

This preliminary phase not only enabled us to validate the functionality of our SL framework but also served as an opportunity to provide hands-on training and support for partners who had hitherto been less experienced with running SL protocols. By fostering collaboration and knowledge sharing during this initial stage, we aim to empower all partner institutions to engage confidently in swarm learning as we move forward.

Statistical Analysis and Hardware

All statistical analyses were performed using Python v3.10. The primary endpoint for classification performance was the area under the receiver operating characteristic curve (AUROC), calculated using the scikit-learn library(5). Differences in AUROC values between models were assessed for statistical significance using DeLong's test(6), employing a fast implementation(7). Confidence intervals for the ROC curves were estimated through the bootstrap method with 1,000 iterations. To compare models, the median patient score from five repetitions of each model was used in DeLong's test, with a significance threshold of $p<0.05$ indicating better performance. AUROCs are reported as mean ± standard deviation. Additional evaluation metrics, including F1 score, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV), were applied to the best-performing model. Sensitivity measures the ability to identify true positives, specificity assesses the ability to identify true negatives, PPV estimates the likelihood of a positive test result being accurate, and NPV evaluates the accuracy of negative test results. At three centers, we deployed SL on different hardware configurations for our computational tasks. Duke hosted in Dresden utilized an operating system version of Ubuntu 22.04.3, coupled with 128 GB RAM, and powered by an NVIDIA Quadro RTX 6000 GPU. Meanwhile, USZ operated on Ubuntu 22.04.4, with a more robust setup comprising 256 GB RAM and two NVIDIA GeForce RTX 4090 GPUs. CAM, on the other hand, employed Ubuntu 20.04.6, featuring 62 GB RAM and an NVIDIA RTX 6000 GPU. Furthermore, each system was connected to at least 10 MBit/sec Internet connection, ensuring consistent and reliable network connectivity throughout the study.

RESULTS

Patient Characteristics

Of the 1,013 patients included, 615 (61%) were cancer-free, while 398 (39%) had cancer. All sites contributed more cancer-free patients than patients with cancer, reflecting the reality of breast cancer screening, where a majority of patients are healthy. An exception to this trend was observed at RSH (**Figure 1**). After separating the examinations into the left and right breast side, the 2,026 unilateral examinations consisted of 1,621 (80%) breasts without cancer and 405 (20%) with cancer.

Local Training

The ResNet model was trained on each site's local training dataset, and its performance was evaluated in two scenarios.

First, the performance on local test splits was evaluated, where the model was tested on the test data from the same site. AUC values ranged widely, from .37 ± .10 (MHA) to .98 ± .01 (CAM), with a mean AUC of .58 ± .14, indicating substantial variability in performance depending on the site (**Figure 2**).

Second, performance across different sites was evaluated by testing models on the test sets of other sites after they were trained on one specific site. For instance, the model trained on CAM's training set was evaluated on UKA's test set. AUC values in this cross-site evaluation ranged from .36 ± .10 to .96 ± .02, demonstrating that no site-specific model generalized effectively across all other sites, underscoring the challenges posed by cross-site variability (**Figure 3**).

Swarm Training

Using the swarm learning framework, the ResNet101 model was trained across multiple sites.

To validate our findings in a real-world scenario, we set up an SL training network spanning three institutions in three countries: Universitaetsspital Zurich (USZ) in Switzerland, Cambridge University Hospitals (CAM) in the United Kingdom, and the Duke dataset, residing in Dresden, Germany. We trained a 3D-ResNet101 model architecture across the three sites and validated it in two separate sites, Mitera Hospital Athens (MHA) in Greece and University Hospital Aachen (UKA) in Germany. Local models trained on either site were also tested on the UKA and MHA datasets (**Table 3-4**).

The model achieved an AUC of .77 ± .03 on the aggregated test datasets.

Real-world training and validation in an international SL network

We found that on the first external test cohort, UKA, the local models trained on Duke, USZ and CAM achieved AUROCs of 0.743 [±0.025], 0.538 [±0.033], and 0.703 [±0.025] respectively. In comparison, the models trained in the SL setup outperformed all the locally trained models, with an AUROC of 0.807 [±0.024]. The swarm model was significantly better than the local models (p= 0.035, 0.001, 0.001, respectively **(Figure 5A**). The swarm models validated on the UKA cohort achieved an F1 score of 0.624 [±0.029], surpassing the local models at Duke, USZ, and CAM, which attained scores of 0.507 [±0.086], 0.452 [±0.027], and 0.495 [±0.073], respectively.

To investigate the generalizability further, we externally validated all models on a second test dataset from MHA. Here, the local models trained on Duke, USZ, and CAM achieved AUROCs of 0.729 [±0.024], 0.520 [±0.040], and 0.673 [±0.036], respectively. Comparatively, models trained in the SL setup outperformed all the locally trained models, with an AUROC of 0.821 [±0.013]. (**Figure 5B**). The swarm model validated on the MHA cohort had an F1 score of 0.596 [±0.036], which is better than the local models. Additional matrices for both local and swarm models, such as sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV), are also documented in **Table 3-4.**

DISCUSSION

The findings of this study highlight the complexities and challenges associated with training deep learning models for cancer detection across diverse clinical settings. Despite advancements in deep learning, our results underscore the inherent limitations of locally optimized models and the potential of swarm learning to address some of these challenges.

The locally trained ResNet models demonstrated significant variability in performance, as evidenced by the wide range of AUC values (37 ± 10 to 98 ± 1) observed across sites. This variability reflects differences in imaging characteristics, patient demographics, and data quality among sites. Notably, the CAM site achieved the highest local AUC, most likely due to the fact that patients with cancer and those without cancer were drawn from distinct internal cohorts, which may have simplified the model's task. Overall this indicates that site-specific factors, such as imaging protocols or population homogeneity, may play a critical role in influencing model performance.

The inability of locally trained models to perform well on test sets from other sites underscores the challenge of cross-site variability. This variability may stem from differences in imaging equipment, or patient characteristics, which create distinct feature distributions across sites. Interestingly, certain sites demonstrated better compatibility with one another, suggesting that similarities in imaging protocols or patient populations might facilitate cross-site generalization. This finding emphasizes the need for strategies that account for inter-site differences to enable robust and generalized model performance.

Swarm learning demonstrated promise in addressing some of these challenges by aggregating knowledge across multiple sites. SL training enhanced the performance for most individual sites, suggesting that the collaborative learning approach helps mitigate the limitations of site-specific data variability. However, the lack of improvement for CAM and RUMC indicates that SL may not always be effective in handling data distributions that are out-of-distribution. Computer-based image analysis of radiology examinations, particularly MRI data, is challenging due to the need for time-consuming and expensive manual annotations of ground truth labels. Additionally, data sharing between medical institutions often faces obstacles due to patient privacy, data ownership, and legal requirements. In this study, we aimed to address these two hurdles by developing DL classification models for radiology. By utilizing weakly supervised DL—which relies on readily available patient labels instead of detailed manual annotations—we sought to reduce the dependency on strong labels. Furthermore, we incorporated SL to enable collaborative DL model training, as a means to eliminate the necessity for data exchange between collaborating institutions while still benefiting from training on each dataset. Applying this combined strategy to breast MRI datasets, we demonstrated the practical feasibility of integrating weakly supervised learning with SL for cancer detection.

Our findings highlight the practical application of combining weakly supervised learning with SL. Weakly supervised DL allows for efficient processing of radiology images using patient labels that can be semi-automatically generated, reducing the need for extensive manual annotations. While we observed that larger models with more parameters showed better performance even with limited data—a trend consistent with observations in other non-medical domains. Concurrently, SL facilitated inter-institutional collaboration without direct data exchange, addressing data privacy and ownership concerns. Although our approach holds promise for large-scale studies spanning multiple hospitals, further research with larger and more diverse datasets is necessary to fully realize this potential.

Gender Dimension

The gender dimension is crucial in the development and evaluation of neural networks for detecting breast cancer in MRI images, as datasets predominantly consist of female patients who are primarily affected by the disease. This reality aligns with the condition's prevalence, making it difficult to explicitly apply a broader gender framework in this context.

Models trained on local data may inherently reflect gender-specific biases or demographic characteristics unique to the local population, potentially limiting their generalizability to broader contexts. In contrast, the swarm learning framework enables collaborative training across diverse datasets from multiple institutions, incorporating a wider variety of gender-related biological and imaging variations. This diversity enhances the model’s robustness and applicability across populations, ensuring that the trained neural networks are inclusive and equitable in performance. More concrete, even without explicit representation, the large sample size likely includes some individuals across a broader gender spectrum, as per statistical probabilities.

Future work should explore gender dimensions more deeply to enhance inclusivity and ensure the continued robustness of diagnostic models across diverse populations. Moreover, including a higher gender variety could effectively address diagnostic challenges, such as variations in breast density or hormonal influences, ultimately promoting gender-sensitive advancements in breast cancer detection.

Limitations

One limitation of our exemplary model training is the relatively small size of the datasets available at each site, which may constrain the ability of locally trained models to learn robust and generalizable patterns. The limited data also likely contributes to the variability in model performance observed across sites. To address this limitation, participating sites continue to actively expand their datasets by adding new cases, which will increase the diversity and volume of data available for training and evaluation.

This study evaluated only the ResNet architecture, leaving open the possibility that other model architectures might perform better in this context. Exploring alternative architectures, such as Vision Transformers or ensemble approaches, could provide insights into whether specific models are better suited to handle cross-site variability. This limitation will be addressed by upcoming ODELIA challenges, facilitating a comprehensive evaluation.

Another limitation is the computational expense associated with higher-dimensional models like 3D-ResNet, potentially posing challenges in resource-limited environments. For future research, we aim to address these limitations by substantially increasing the patient count and including a broader range of centers worldwide. We also plan to investigate strategies to mitigate the impact of noisy labels and improve model performance. While our study provides a proof of concept demonstrating the feasibility of integrating SL with weakly supervised learning, further work is necessary to enhance predictive performance and validate the approach in larger, more diverse populations before it can be considered for developing clinical-grade DL systems in radiology image analysis for MRI-based cancer screening.

CONCLUSION

In conclusion, while locally optimized models can provide high performance on their respective datasets, their poor generalizability underscores the necessity of collaborative approaches such as swarm learning. By aggregating knowledge across multiple sites, SL holds promise for improving breast cancer diagnosis.

TABLES & FIGURES

Table 1: Image acquisition. Age is given in years for mean ± standard deviation.

Site	Country	Patients	Age [years]	Manufacturer	Field Strength [T]	Fat suppression	2D/3D
CAM	England	302	57±7	General Electric	1.5, 3.0	Yes	3D
MHA	Greece	100	48±11	Siemens	3.0	Yes	3D
RSH	Spain	100	NA	Philips	1.5	No	NA
RUMC	Netherlands	100	NA	Siemens	1.5, 3.0	No	3D
UKA	Germany	100	55±9	Philips	1.5	No	2D
UMCU	Netherlands	161	NA	Philips	1.5, 3.0	Yes	3D
USZ	Switzerland	100	NA	NA	NA	NA	NA
VHIO	Spain	50	NA	NA	NA	NA	NA

Table 2: Image resolution and spacing. All values are given as median [minimum, maximum].

Site	Resolution XY	Resolution Z	Spacing XY	Spacing Z
CAM	512 [512, 512]	112 [68, 116]	0.68 [0.68, 0.74]	2.00 [1.40, 2.00]
MHA	424 [360, 456] 256 [256, 256]	104 [104, 122]	0.86 [0.79, 0.98]	2.00 [2.00, 2.00]
RSH	432 [400, 448]	130 [110, 130]	0.87 [0.72, 0.87]	1.50 [1.50, 1.50]
RUMC	512 [384, 512] 120 [120, 448]	256 [144, 256]	0.69 [0.66, 0.94] 1.30 [0.80, 1.30]	0.69 [0.66, 1.10]
UKA	512 [512, 560]	27 [25, 31]	0.64 [0.55, 0.78]	3.00 [3.00, 3.50]
UMCU	384 [352, 672]	200 [106, 222]	0.89 [0.51, 0.97]	0.90 [0.90, 1.60]
USZ	448 [384, 490] 448 [384, 448]	104 [104, 144]	0.76 [0.62, 0.94]	1.50 [1.20, 1.95]
VHIO	480 [416, 512]	104 [60, 204]	0.80 [0.62, 0.87]	1.60 [1.00, 2.20]

Table 3: Prediction performance of different centers used for benchmarking breast cancer tumor prediction on the external validation UKA cohort.

External Validation on UKA dataset by real-world training with 3D-ResNet101				
	Duke	**USZ**	**CAM**	**SWARM**
AUROC	0.743 [±0.025]	0.538 [±0.033]	0.703 [±0.025]	0.807 [±0.024]
F1 Score	0.507 [±0.086]	0.452 [±0.027]	0.495 [±0.073]	0.624 [±0.029]
Sensitivity	0.635 [±0.144]	0.430 [±0.06]	0.589 [±0.078]	0.767 [±0.021]
Specificity	0.815 [±0.074]	0.824 [±0.077]	0.826 [±0.096]	0.813 [±0.028]
PPV	0.404 [±0.099]	0.308 [±0.021]	0.398 [±0.076]	0.559 [±0.023]
NPV	0.840 [±0.068]	0.821 [±0.055]	0.859 [±0.084]	0.831 [±0.017]

Table 4: Prediction performance of different centers used for benchmarking breast cancer tumor prediction on the external validation MHA cohort.

External Validation on the MHA dataset by real-world training with 3D-ResNet101				
	Duke	**USZ**	**CAM**	**SWARM**
AUROC	0.729 [±0.024]	0.520 [±0.040]	0.673 [±0.036]	0.821 [±0.013]
F1 Score	0.517 [±0.089]	0.537 [±0.075]	0.446 [±0.073]	0.596 [±0.036]
Sensitivity	0.645 [±0.132]	0.403 [±0.118]	0.584 [±0.06]	0.744 [±0.02]
Specificity	0.774 [±0.089]	0.796 [±0.108]	0.774 [±0.014]	0.804 [±0.012]
PPV	0.453 [±0.049]	0.395 [±0.025]	0.382 [±0.104]	0.553 [±0.033]
NPV	0.806 [±0.074]	0.840 [±0.05]	0.873 [±0.037]	0.839 [±0.037]

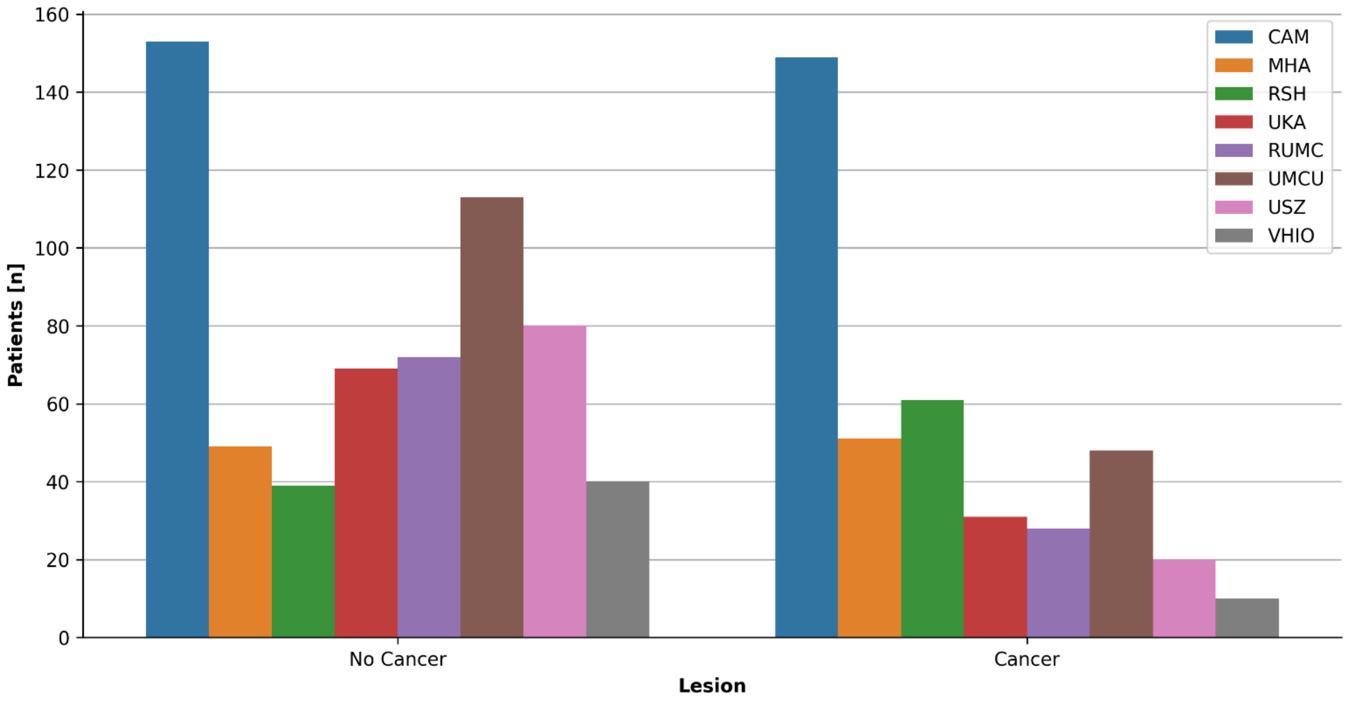

Figure 1: Distribution of patients with and without cancer across all sites.

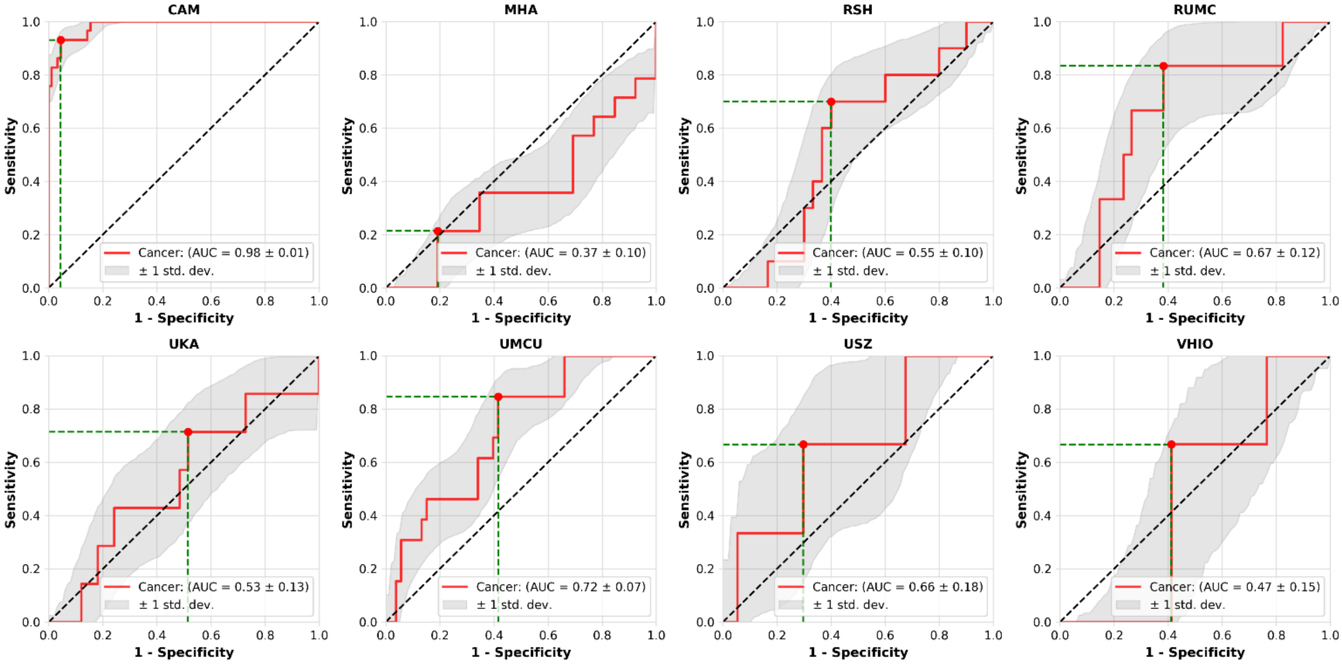

Figure 2: Classification performance when the model was trained and tested locally.

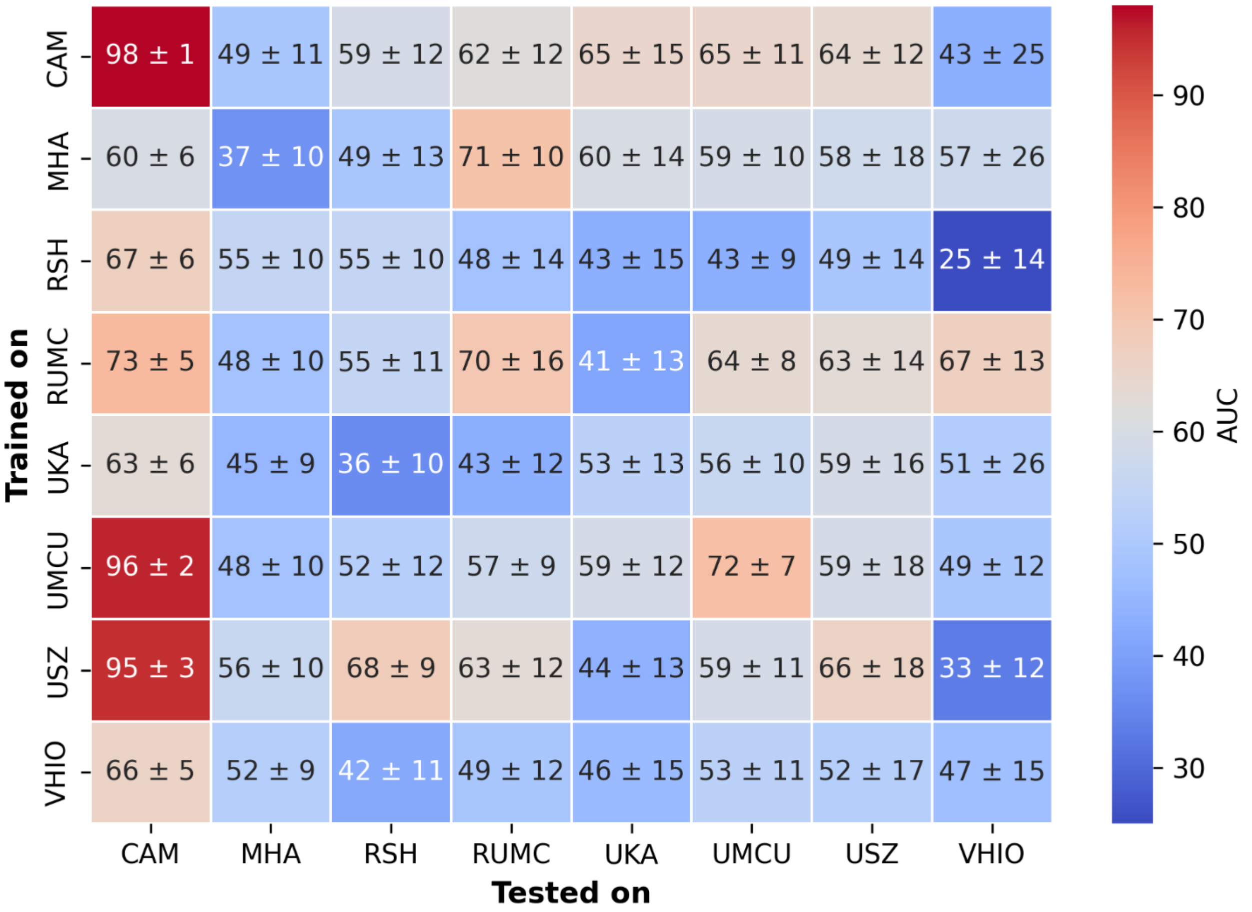

Figure 3: Classification performance when the model was trained locally on and tested across sites. Performance was measured by the Area Under the Receiver Operating Curve (AUC).

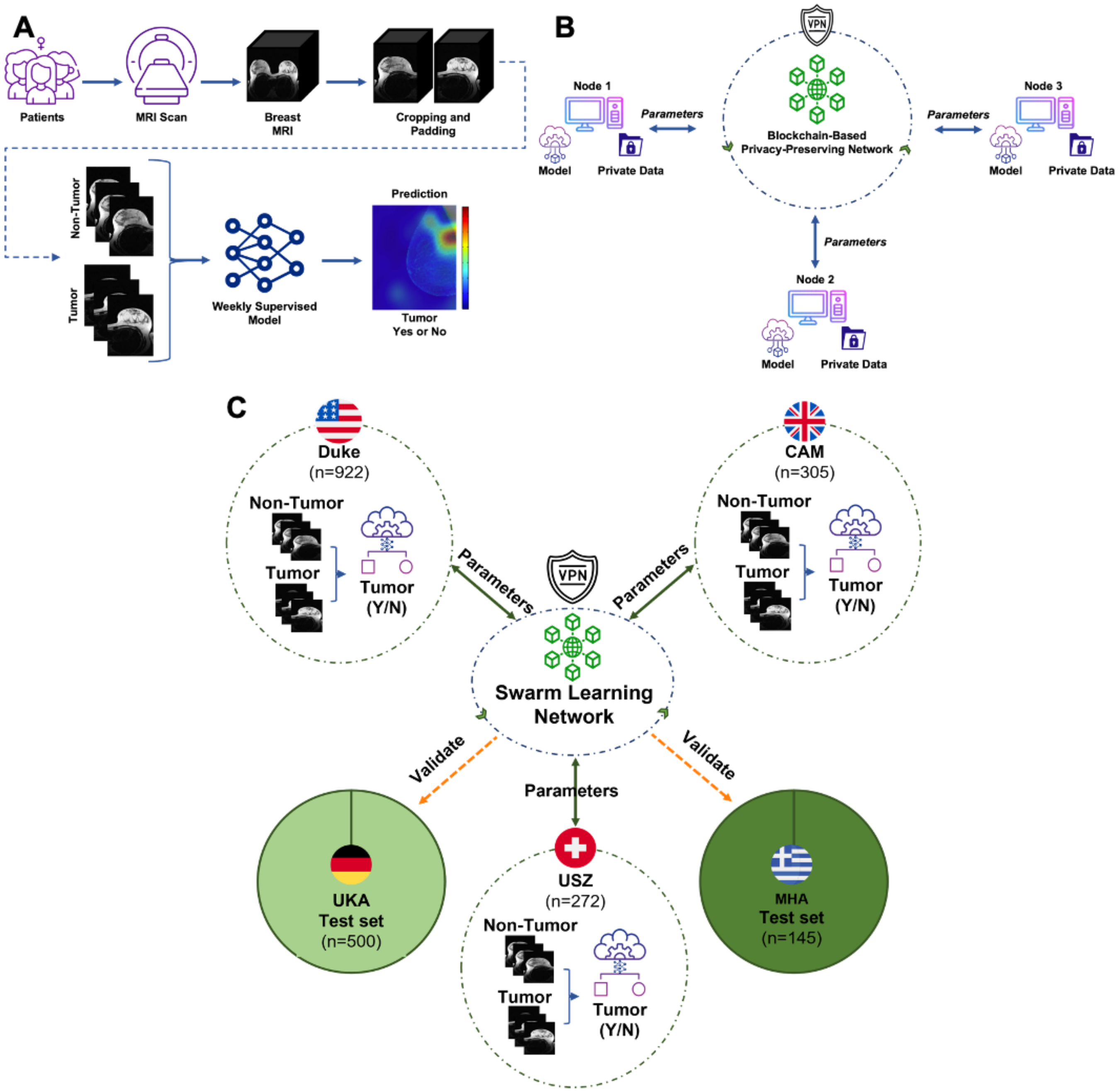

Figure 4: Schematic of the Weakly Supervised Learning (WSL) and Swarm Learning (SL) workflow. **(A)** Schematic representation of the Deep Learning-based WSL workflow for breast cancer tumor detection on Magnetic Resonance Imaging (MRI) data, **(B)** Overview of the SL setup for a 3-node network **(C)** Combined representation of real-world SL-based WSL for Breast Cancer Tumor Detection and Data Split Ratio.

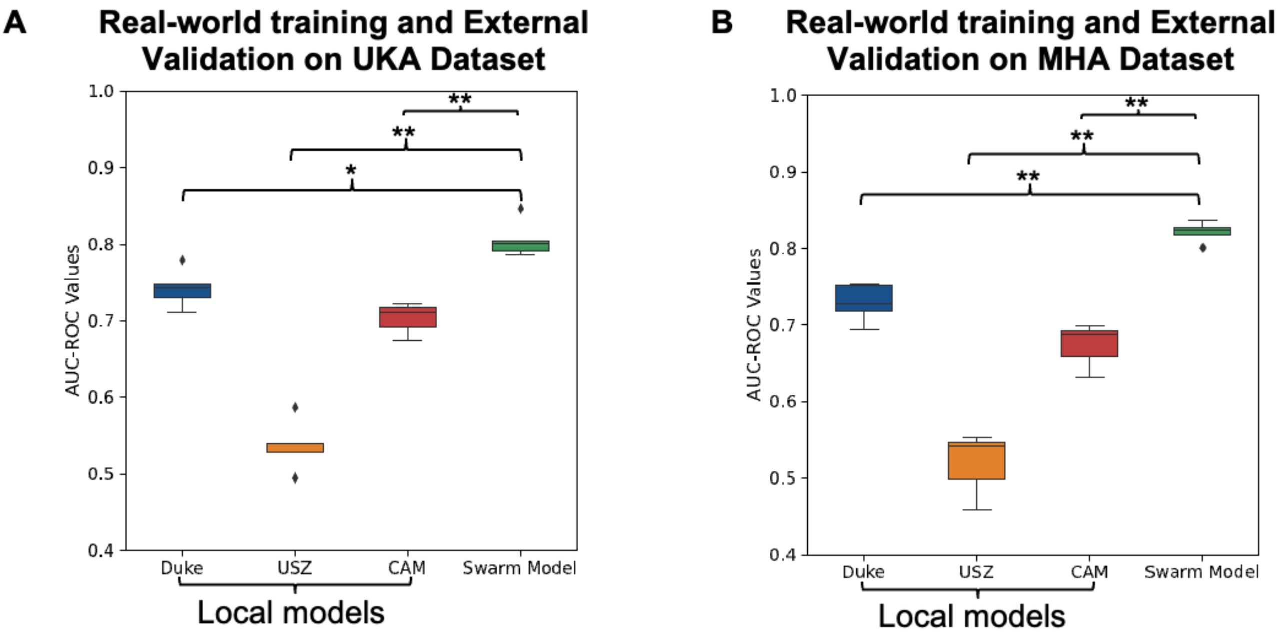

Figure 5: Benchmarking models on internal and external validation. **(A)** Classification performance for prediction of the tumor using 3D-Resnet101 model trained using real-world swarm learning across three cohorts: Duke, USZ, and CAM. Its classification performance was evaluated on an external validation cohort, UKA, for tumor prediction. Local model performance was assessed using AUROC and DeLong's test to compare it with swarm models. The significance level was set at p < 0.05 (*P < 0.05, **P < 0.001), and median patient scores from five repetitions determined superior performance. **(B)** Classification performance for the prediction of tumors using the 3D-Resnet101 model was trained using real-world swarm learning across three cohorts: Duke, USZ, and CAM. Its classification performance was evaluated on an external validation cohort, MHA, for tumor prediction. Local model performance was assessed using AUROC and DeLong's test to compare it with swarm models. The significance level was set at p < 0.05 (*P < 0.05, **P < 0.001), and median patient scores from five repetitions determined superior performance. The training cohort from Duke is consistently represented by the dark blue color throughout the figure.

REFERENCES

1. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE; 2016. p. 770–778. doi: 10.1109/CVPR.2016.90.

2. Cardoso MJ, Li W, Brown R, et al. MONAI: An open-source framework for deep learning in healthcare. arXiv; 2022; doi: 10.48550/ARXIV.2211.02701.

3. Pérez-García F, Sparks R, Ourselin S. TorchIO: A Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. arXiv; 2020; doi: 10.48550/ARXIV.2003.04696.

4. Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. 2017; doi: 10.48550/arXiv.1711.05101.

5. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. 2011;12:2825–2830.

6. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics. 1988;44(3):837. doi: 10.2307/2531595.

7. Sun X, Xu W. Fast Implementation of DeLong's Algorithm for Comparing the Areas Under Correlated Receiver Operating Characteristic Curves. IEEE Signal Process Lett. 2014;21(11):1389–1393. doi: 10.1109/LSP.2014.2337313.