# Beyond the Sandbox: Real-World Federated Learning for MRI Prostate Cancer Detection

Ashkan Moradi[1]
ashkan.moradi@ntnu.no
Bendik Skarre Abrahamsen[1]
bendik.s.abrahamsen@ntnu.no
Jeroen Geerdink[2]
j.geerdink@zgt.nl
Derya Yakar[3]
d.yakar@umcg.nl
Henkjan Huisman[4]
henkjan.huisman@radboudumc.nl
Tone Frost Bathen[1,5]
tone.f.bathen@ntnu.no
Mattijs Elschot[1,6]
mattijs.elschot@ntnu.no

[1] Department of Circulation and Medical Imaging, Norwegian University of Science and Technology, Trondheim, Norway
[2] Department of Information and Organization, Hospital Group Twente, Almelo ,The Netherlands
[3] Medical Imaging Center, Department of Radiology, University Medical Center Groningen, The Netherlands
[4] Diagnostic Image Analysis Group, Department of Medical Imaging, Radboud University Medical Center, Nijmegen, The Netherlands
[5] Department of Radiology and Nuclear Medicine, St. Olavs Hospital, Trondheim University Hospital, Norway
[6] Central Staff, St. Olavs Hospital, Trondheim University Hospital, Norway

**Abstract**

The success of computer-assisted methods based on deep learning for the early detection of prostate cancer (PCa) has attracted significant attention in recent years. Successful training of such models, however, requires access to large amounts of patient data. This data is typically not available at a single institution, and sharing patient-specific information raises privacy concerns. Federated learning (FL) thus emerges as a viable solution that enables decentralized model training without the need to share patient data. In this study, we design and implement a real-world FL solution for the detection of clinically significant prostate cancer in biparametric MRI. We benchmark the performance by conducting centralized model training and simulated FL experiments using the Flower FL and NVIDIA FLARE frameworks, and compare these results with those from a real-world FL solution implemented using the Rhino Federated Computing platform. The results showed that FL-based models outperformed local models and achieved performance comparable to the centralized model. Furthermore, the real-world FL model closely replicated the performance of simulated FL models and demonstrated that the choice of FL implementation framework had minimal impact on performance across external test

sets. Additionally, we provided a time performance analysis for simulated and real-world FL scenarios, highlighting the longer execution time and challenges in implementing FL in practice. Code available here https://github.com/AshknMrd/moradi2025realworldFL/.

# 1 Introduction

The high global mortality rate associated with prostate cancer (PCa) underscores the importance of detecting clinically significant prostate cancer (csPCa) [16, 23]. Recent guidelines from the European Association of Urology recommend magnetic resonance imaging (MRI) as the initial diagnostic test for PCa prior to biopsy [5]. Consequently, MRI analysis has become a critical component of the PCa diagnostic pathway [25]. Simultaneously, advances in medical technology and imaging have led to an exponential increase in available data, driving the adoption of deep learning-based image analysis models in clinical practice. However, training these models to achieve high accuracy requires access to large, diverse datasets, which are often unavailable at individual institutions. As a result, collaboration through data sharing is necessary but raises significant privacy concerns. Federated learning (FL) has emerged as a promising solution, enabling distributed model training without sharing private data [10]. In this study, we propose an FL solution for csPCa detection using biparametric MRI (bpMRI) data. We evaluate the performance of FL across various frameworks in both simulated and real-world environments. Their performance is compared to single-institution models, from now referred to as local models, and centralized models, with a detailed analysis of associated challenges and impacts.

In healthcare applications, where protecting patient data is essential, FL holds great promise. By enabling collaborative model training across multiple clients, FL can enhance diagnostic accuracy and clinical outcomes [4]. Moreover, by eliminating the need for data centralization, FL offers practical solutions for digital health, with successful applications in various oncological contexts, including breast, lung, and prostate cancer [1]. FL frameworks have been applied in healthcare for brain tumor segmentation, demonstrating the generalizability of FL models and addressing challenges related to data sharing among institutions in both simulated and real-world settings [13, 21, 22]. In a simulated setting, a federated approach for training a lung nodule detection model on horizontally distributed data is investigated in [9], while [26] introduces a customized FL framework for identifying csPCa and classifying skin lesions. Similarly, [8] proposes an FL-based method for prostate cancer diagnosis and Gleason grading using pathological images. Recent studies have focused on the technical developments and potential of FL in both simulated and real-world healthcare settings [15, 24]; however, evaluating its effectiveness, model performance, and generalizability in specific real-world oncological contexts remains a challenge.

In the prostate domain using MRI data, a simulated FL framework was implemented for automated classification of csPCa, addressing cross-client variation by mapping raw images from individual clients onto a shared image space prior to federated training [27]. A real-world implementation of prostate gland segmentation on T2-weighted MRI data was presented in [19], while [14] proposed a versatile FL framework for real-world cross-site training and evaluation of customized deep learning-based MRI PCa detection. This approach collaboratively trains a 3D UNet-based model with a region-of-interest classification head on diverse annotated prostate MRI data for lesion classification and detection. Kades et al. [7] introduced a framework for clinical implementation of FL based on a standardized digital infrastructure across multiple university hospitals. This framework enables the

adaptation of nnU-Net [6] for FL settings using the Kaapana framework [20], and evaluates it for MRI prostate segmentation. Moradi *et al*. [11, 12] benchmarked the impact of simulated FL environments and configuration optimization in MRI csPCa detection. However, the real-world implementation and its challenges have not been addressed. In this work, we investigate such real-world FL implementations and compare them with two simulated FL frameworks based on Flower FL [2] and NVIDIA FLARE (NVFlare) [17]. Addressing a gap in the literature, we evaluate the impact of real-world FL on both patient- and lesion-level accuracy of csPCa detection, with the main contributions summarized as follows:

- Design and implementation of a deep learning-based solution, a 3D U-Net model architecture, for detecting csPCa on bpMRI data in a federated setting.
- Comprehensive evaluation of the model performance at both the patient and lesion levels on independent test sets, with comparisons among locally trained models, a centralized approach, and simulated and real-world FL models.
- Investigation of real-world FL effectiveness compared to simulated implementations across different frameworks, highlighting key results in performance and time, and addressing implementation challenges.

# 2 Materials and Method

This study proposes a deep learning-based solution for detecting csPCa that trains a U-Net model in an FL setting on bpMRI data. The bpMRI dataset comprised T2-weighted images, high b-value diffusion-weighted images, and apparent diffusion coefficient maps. We conducted simulated FL experiments using the Flower FL [2] and NVFlare [17] frameworks, along with a real-world FL model training on the Rhino Federated Computing Platform (FCP). Simulated FL means that the clients are in the same cloud infrastructure, even if they are on different machines not placed in different geographical locations or institutions. In contrast, real-world FL involves clients in different geographical locations within different cloud infrastructures. An overview of the differences between the simulated and real-world implementations of the FL experiments is provided in Figure 1.

## 2.1 Methodology

The csPCa is defined as ISUP grade $\geq 2$ cancer lesions.[1] To perform lesion detection, the server assigns clients a global model designed using nnU-Net [6], which configures a 3D U-Net-based pipeline tailored to the input data geometry and available computational resources. Each client uses T2-weighted, high b-value, and apparent diffusion coefficient sequences as input and trains its local model to generate a voxel-level probability map indicating the likelihood of each voxel being cancerous or non-cancerous. Following the Prostate Imaging: Cancer Artificial Intelligence (PI-CAI) Grand Challenge guidelines [18], these probability maps are then post-processed to extract lesions and predict the patient-level likelihood of harboring csPCa.

We designed three training setups: local and centralized training, simulated FL, and real-world FL. In local training, each client trains its model independently on its own data for E local epochs. Centralized training aggregates data from all clients at a single site,

[1] International Society of Urological Pathology (ISUP)

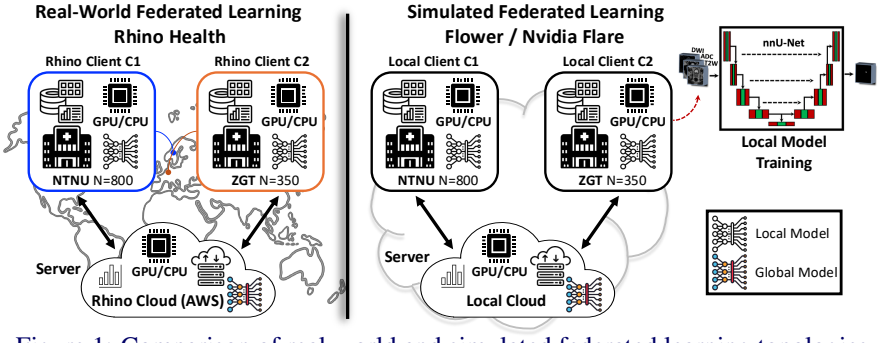


Figure 1: Comparison of real-world and simulated federated learning topologies.

where a nnUNet [6] model is trained for E epochs. For simulated FL, we used two clients hosted on separate machines, each equipped with an NVIDIA A40-48GB GPU, within the local cloud infrastructure at the Norwegian University of Science and Technology (NTNU). Training begins with the central server, located within the same cloud infrastructure as the clients, distributing an initial model to all participating clients. This model is identical to the one used for local training. Each client then trains the model on its own data for E local epochs and transmits the updated model parameters back to the server. Using the FedAvg algorithm [10], the server aggregates local models into a global model and redistributes it to the clients to initiate another round of training. This process was iterated for R federated rounds to progressively refine the global models. The entire FL simulation was conducted within a secure on-premises local cloud, using Flower [2] and NVFlare [17], two widely recognized FL frameworks. Additionally, we conducted a real-world implementation of FL model training using the Rhino FCP. In this setup, client C1 was hosted by NTNU in Norway on an NVIDIA A40-48GB GPU, and client C2 was hosted by Ziekenhuis Groep Twente in the Netherlands on an NVIDIA RTX A5000-24GB GPU. The Rhino client configurations were installed on each site, and both clients were connected to the Rhino Server, hosted on Amazon Web Services (AWS). The training scripts for the Rhino clients were implemented based on NVFlare [17], as required by Rhino FCP, and the experiments were conducted via the Rhino FCP interface. After training, the model parameters were downloaded locally and evaluated on various test sets.

For real-world FL implementation, established platforms such as Rhino FCP are preferable for production systems that handle private data, as they provide off-the-shelf software and infrastructure to facilitate deployment, along with validated implementations of privacy-preserving mechanisms. In contrast, relying on ad hoc or purely open-source setups increases the risk of implementation errors and insufficient safeguards, which can compromise data confidentiality.

## 2.2 Experiments Design and Evaluation

The local and centralized experiments were conducted with E = 500 local epochs, where each epoch is defined as a fixed number of training steps over mini-batches. This number of epochs was chosen to ensure client convergence. In FL experiments, each client performed E local epochs, shared model updates with the server, and repeated this process for R federated rounds. To ensure a fair comparison between the local and FL models, we kept $E \times R$ constant and equal to the number of iterations used in the local experiments. Accordingly, the FL

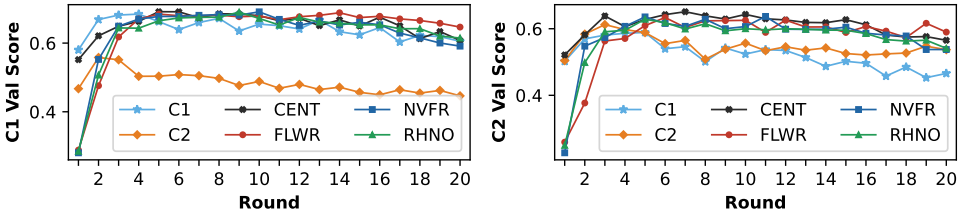


Figure 2: Comparison of average validation PI-CAI scores across five folds for local models (C1, C2), centralized training (CENT), simulated FL with Flower (FLWR) and NVFlare (NVFR), and real-world FL with Rhino FCP (RHNO).

setup used E = 25 local epochs and R = 20 federated rounds. To allow direct comparisons between the local and FL models after each round, we saved the local and centralized models at intervals of 25 epochs. All scenarios were evaluated using five-fold cross-validation, where each fold used a different subset of the data for validation. The ensemble of the resulting five models was then evaluated on various test sets.

To evaluate the trained models, we followed the PI-CAI Grand Challenge guidelines [18]. The area under the receiver operating characteristic curve (AUROC) was used to assess patient-level diagnostic performance, while average precision (AP) was employed to evaluate lesion-level detection performance. AP summarizes the precision–recall curve as the weighted mean of precision values at each decision threshold. Overall model performance for csPCa detection was quantified using the PI-CAI score, calculated as (AUROC + AP)/2, which combines both patient- and lesion-level metrics. The process of computing the patient-level metric from voxel-level predictions and extracting meaningful lesions from raw outputs are adopted from [3].

**Lesion Extraction:** Lesion extraction was performed in accordance with the PI-CAI Grand Challenge guidelines [18]. The voxel-level predicted probability map output from the model contains voxel values representing the likelihood of csPCa. These probability maps undergo post-processing to extract lesion candidates, where dynamic thresholding is employed to select the candidates. In this method, the threshold is adjusted based on the distribution of probability values within each probability map. After thresholding, connected component analysis is performed to group neighboring voxels that exceed the threshold into distinct 3D clusters. To further refine lesion extraction, these potential lesions undergo size filtering, and the top k lesions are selected based on confidence scores. Further details regarding the process are provided in the supplemental materials, Section **1.1**.

## 3 Experimental Results

We utilized the public training and development dataset from the PI-CAI Grand Challenge, which comprises 1500 annotated bpMRI scans from patients at three different centers. FL experiments were conducted with two clients: client C1 used data from Radboud University Medical Center, consisting of 800 male patients (mean age: 64.6±7.1 years; range: 35–92), and client C2 used data from Ziekenhuis Groep Twente, comprising 350 male patients (mean age: 66.6±7.4 years; range: 43–89). Data from the third center, University Medical Center Groningen (UMCG), containing 350 male patients (mean age: 66.9±6.8 years; range:

Table 1: Dataset distribution for csPCa detection across different clients.

| Scenario | | Dataset Detail | Vendor |
|---|---|---|---|
| Local C1 | 800 | (640/160 - train/val) | Siemens |
| Local C2 | 350 | (280/70 - train/val) | Siemens |
| Centralized | 1150 | (920/230 - train/val) | Siemens |
| UMCG | 350 | (test) | Siemens & Philips |
| In-House | 200 | (test) | Siemens |

45–83), was reserved as an external test set for model evaluation. Additionally, we used an independent in-house dataset of 200 male patients (mean age: 64.4±6.9 years; range: 44–76) to further evaluate the trained models. This dataset included patients that were part of the hidden test set in the PI-CAI challenge. For all patients in the independent in-house dataset, the presence of csPCa lesions was confirmed through histopathology, and the absence of csPCa in patients without biopsy or with negative biopsy results was verified via a 3-year follow-up. Each client used the split of 80/20 for training and validation. In the centralized experiment, data from both clients were combined at a single site, with the training and validation sets defined as the union of the respective sets from both clients. The data distribution details is summarized in Table 1.

The experiments were performed using Python 3.10, Flower 1.17, nnU-Net 2.5, CUDA 12.4, running on Ubuntu 22.04.5 LTS. For both the simulated and real-world FL implementations using NVFlare, version 2.5 was used. For all experiments, the data was pre-processed by resampling all data to the same resolution (0.5mm × 0.5mm × 3.0mm/voxel) and cropping to 20 slices, each of size 256 × 256 voxels. We employed the default nnU-Net configuration as the Stochastic Gradient Descent optimizer with an initial learning rate of 0.01, momentum of 0.99, and FedAvg uses weighted averaging based on each client's local dataset size. The batch size was 3, and each epoch used 250 training steps over mini-batches. Additionally, the MR images were normalized independently using instance-wise z-score normalization. Following the PI-CAI Grand Challenge guidelines [13] and baseline models, the model training experiments employed cross-entropy loss and evaluated model performance using the PI-CAI score, which includes both patient- and lesion-level evaluation.

## 3.1 Results and Discussion

Figure 2 shows the average validation PI-CAI scores across five folds for both clients. The mean of these two curves, as shown in Figure 3, is used as the metric to determine the best-performing model. As illustrated in Figure 3, the optimal performance occurs early in the total number of iterations: C1 at E=100, C2 at E=75, centralized training at E=150, simulated Flower FL at R=6, simulated NVFlare FL at R=5, and real-world Rhino FL at R=5. The optimal models are then selected and evaluated on the in-house and UMCG test sets. Specifically, we report the performance of the ensemble model, created by averaging the prediction maps of five models trained on separate folds at the optimal round/epoch.

To examine the patient- and lesion-level performance of the ensembled best models, Tables 2 and 3 present the AUROC and AP metrics across various scenarios on the in-house and UMCG test sets, respectively. These results show that FL-based models outperform locally trained models and closely match the performance of centralized model, without requiring data sharing between institutions. This consistent improvement in performance

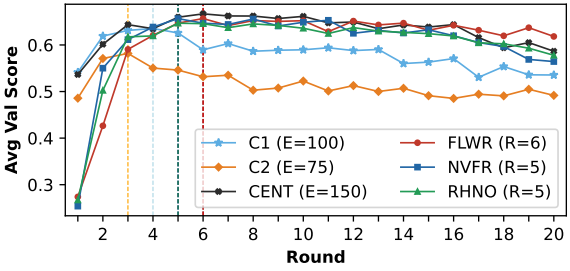


Figure 3: Average validation PI-CAI scores across clients and folds per round for local (C1, C2) and centralized (CENT) models, simulated FL with Flower (FLWR) and NVFlare (NVFR), and real-world FL (RHNO). Dashed lines indicate the best performing round for each model. Each round corresponds to 25 local epochs; thus, a peak at R=4 equals E=100 for local and centralized models.

Table 2: Performance metrics and 95% confidence interval for the ensemble of the 5-fold cross-validated best models, evaluated on the in-house test set.

| Model | AUROC | AP | PI-CAI Score |
|---|---|---|---|
| Local C1 | 0.89 [0.85, 0.94] | 0.48 [0.38, 0.60] | 0.69 [0.62, 0.76] |
| Local C2 | **0.92** [0.87, 0.95] | 0.36 [0.25, 0.48] | 0.64 [0.58, 0.71] |
| Centralized | 0.91 [0.86, 0.95] | 0.53 [0.41, 0.65] | 0.72 [0.65, 0.79] |
| Flower FL | 0.90 [0.85, 0.94] | **0.54** [0.43, 0.66] | 0.72 [0.65, 0.79] |
| NVFlare FL | **0.92** [0.88, 0.96] | **0.54** [0.42, 0.66] | **0.73** [0.66, 0.80] |
| Rhino FL | 0.90 [0.85, 0.94] | 0.53 [0.41, 0.66] | 0.72 [0.65, 0.79] |

compared to the local model across various test sets demonstrates the generalizability of the FL models when evaluated on unseen data. These tables also demonstrate that real-world FL implementations can achieve performance comparable to simulated FL models, and that the choice of FL framework has minimal impact on performance.

To further assess the generalizability of the trained models, Figure 4 presents the ensembled PI-CAI scores of different models evaluated on the in-house and UMCG test sets. The FL models consistently demonstrate strong generalizability, outperforming the local models on both unseen test sets in nearly every round. Although the same approach was used to select the best model in all experiments based on validation performance, the selected checkpoint for the centralized model at E=150 does not correspond to its peak performance on the external test sets. This may explain why, in Table 2, the centralized model performs slightly worse than the FL models. Comparing the results in Tables 2 and 3, it is evident that performance is lower on the UMCG test set, which may be due to differences in scanners and image acquisition protocols. In addition, since both the simulated FL with NVFlare and the real-world FL use NVFlare as the implementation tool, we expected to observe more identical performances; however, such marginal performance differences may be due to the use of different seed numbers in the implementations. The performance of the real-world FL implementation (Rhino FL) is comparable to that of the simulated FL approaches (NVFlare FL and Flower FL) and generally surpasses the non-federated local models (Local C1 and Local C2) across all metrics.

We also analyzed the time required to complete each round in the successful FL exper-

Table 3: Performance metrics and 95% confidence interval for the ensemble of the 5-fold cross-validated best models, evaluated on the UMCG test set.

| Model | AUROC | AP | PI-CAI Score |
|---|---|---|---|
| Local C1 | 0.71 [0.65, 0.76] | 0.31 [0.23, 0.41] | 0.51 [0.44, 0.58] |
| Local C2 | 0.71 [0.65, 0.77] | 0.29 [0.20, 0.39] | 0.50 [0.43, 0.57] |
| Centralized | **0.75** [0.69, 0.80] | **0.36** [0.28, 0.47] | **0.56** [0.49, 0.63] |
| Flower FL | 0.72 [0.66, 0.77] | 0.34 [0.24, 0.43] | 0.53 [0.46, 0.59] |
| NVFlare FL | 0.73 [0.67, 0.79] | 0.35 [0.27, 0.46] | 0.54 [0.47, 0.62] |
| Rhino FL | 0.72 [0.66, 0.77] | 0.33 [0.25, 0.44] | 0.52 [0.46, 0.60] |

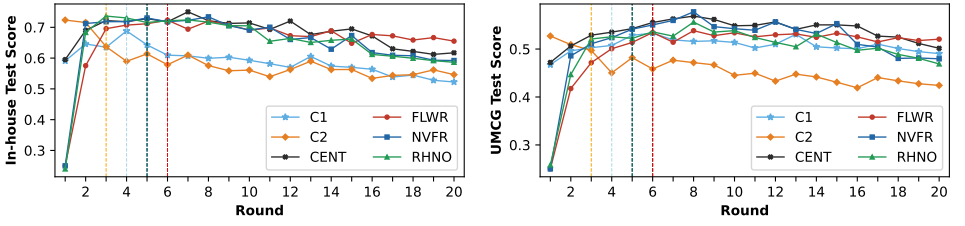


Figure 4: Comparison of the ensembled PI-CAI scores for trained models evaluated on various test sets. Dashed lines indicate the round at which the average validation PI-CAI score across clients and folds peaked for each curve.

iments. Figure 5 (left) represents the average time per round and its variation for simulated FL using Flower and NVFlare, and real-world FL using Rhino FCP. As expected, the real-world FL setup required more time to complete each round due to the geographical separation between the server and clients, and because data transfers had to pass through different client-side firewalls. Meanwhile, the simulated FL experiments exhibited shorter and more consistent round completion times, due to stable connections within the on-premises cloud system where all components were co-located. To examine the time analysis and investigate the cause of the longer rounds and their variability in the real-world FL, Figure 5 (right) shows the average time taken to upload model updates from each client to the server for the NVFlare-based FL models. The delay appears to be caused by an unstable connection between the server and client C2. However, we observed similar issues with client C1 as well, which might be due to server-side congestion or NVFlare 2.5 connectivity issues that appear in the real-world implementation requiring external connectivity. Such problems might not exist in other versions. We note that the subsequent release of NVFlare 2.6 brought significant communication enhancements, including native tensor transfer and model streaming improvements, which address known sources of network strain and may resolve such issues.

Figure 6 demonstrates the qualitative performance of different models in detecting csPCa on patients from the in-house test set. Each row represents one test case, while each column shows the image and prediction for different models as an overlay. In the first case, all local, centralized, and federated models successfully predict the csPCa lesion, comparable to the ground truth (GT) lesion. However, in the second case, clients trained separately using only their own local data fail to detect the lesion, while employing the FL technique generalizes the trained model and enables successful detection of the csPCa lesion.

A primary limitation of this study is the exclusive use of publicly available data for model

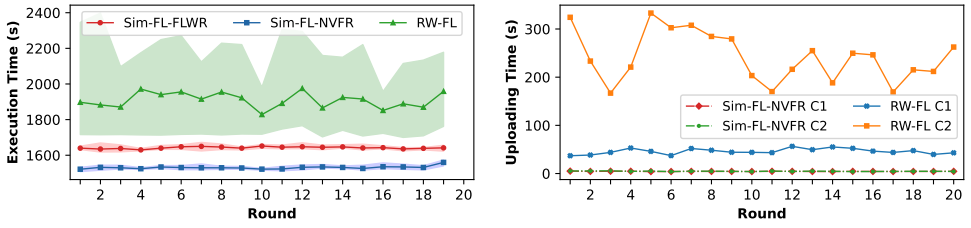


Figure 5: Time analysis per training round for simulated FL based on Flower (Sim-FL-FLWR), simulated FL based on NVFlare (Sim-FL-NVFR), and real-world FL based on Rhino (RW-FL).

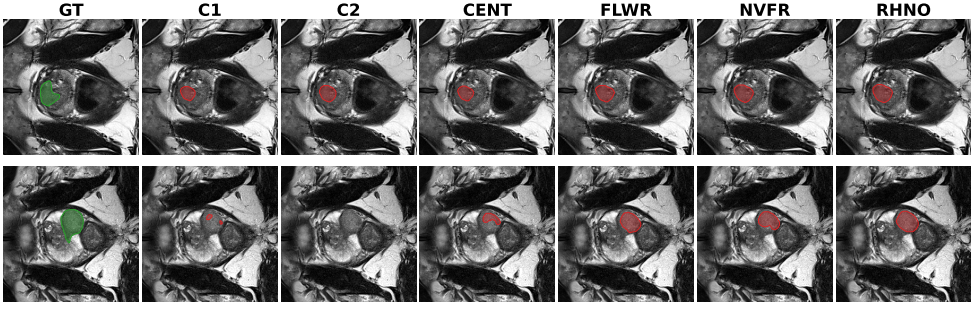


Figure 6: Qualitative performance comparison of predictions from different models on test patients from the in-house test set. Each row shows the lesions predicted by different models for a single test case.

training, in accordance with agreements among participating partners during the proof-of-concept phase. Incorporating private data and involving more clients in future work may enhance the model's generalizability and clinical applicability. The limited number of clients in this experiment reflects the constraints of implementing the FL scenario in a real-world setting. Adding more clients would require involving additional partners, establishing new agreements, and performing extra hardware installations and client configurations, efforts that demand more time and financial resources and are planned for future work. Furthermore, due to the use of only public data, more advanced privacy-preserving techniques, such as differential privacy, were not explored. Another major challenge, not captured in the time analysis, was the instability of the connection between the server and clients during the real-world FL experiment. Although the FL process was successfully initiated, the connection with one client was lost after several rounds, resulting in a timeout and experiment failure. Ensuring more stable connections, from the hardware and software perspective, and allocating experiment-specific server resources may help mitigate such issues.

# 4 Conclusion

This work proposed a real-world implementation of FL for detecting csPCa using bpMRI data. We evaluated the effectiveness, performance, and implementation challenges of real-world FL in comparison with local, centralized, and simulated FL models. The FL-based

models outperformed local models and achieved performance comparable to that of the centralized model, without requiring data sharing. Furthermore, the real-world FL model closely replicated the performance of simulated FL models and demonstrated that the choice of implementation framework had minor impact on performance across external test sets. Additionally, time analysis showed longer execution times in real-world FL scenario due to FL contributors being deployed in separate cloud environments.

## Acknowledgment

The authors gratefully acknowledge Rhino Federated Computing for providing access to their platform and for their valuable support throughout the implementation of this work. This work was supported by the Norwegian Cancer Society and Prostatakreftforeningen under the Project FLIP.AI – Federated Learning to Improve Prostate cancer imaging with Artificial Intelligence with Project code: 215951.

## References

[1] Anshu Ankolekar, Sebastian Boie, Maryam Abdollahyan, Emanuela Gadaleta, Seyed Alireza Hasheminasab, Guang Yang, Charles Beauville, Nikolaos Dikaios, George Anthony Kastis, Michael Bussmann, et al. Advancing breast, lung and prostate cancer research with federated learning. a systematic review. *npj Digital Medicine*, 8 (1):1–12, 2025.

[2] Daniel J Beutel, Taner Topal, Akhil Mathur, Xinchi Qiu, Javier Fernandez-Marques, Yan Gao, Lorenzo Sani, Kwing Hei Li, Titouan Parcollet, Pedro Porto Buarque de Gusmão, et al. Flower: A friendly federated learning research framework. *arXiv preprint arXiv:2007.14390*, 2020.

[3] Joeran S Bosma, Anindo Saha, Matin Hosseinzadeh, Ivan Slootweg, Maarten de Rooij, and Henkjan Huisman. Semisupervised learning with report-guided pseudo labels for deep learning–based prostate cancer detection using biparametric mri. *Radiology: Artificial Intelligence*, 5(5):e230031, 2023.

[4] Alexander Chowdhury, Hasan Kassem, Nicolas Padoy, Renato Umeton, and Alexandros Karargyris. A review of medical federated learning: Applications in oncology and cancer research. In *International MICCAI Brainlesion Workshop*, pages 3–24. Springer, 2021.

[5] European Association of Urology (EAU). EAU guidelines on prostate cancer. 2019.

[6] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

[7] Klaus Kades, Jonas Scherer, Maximilian Zenk, Marius Kempf, and Klaus Maier-Hein. Towards real-world federated learning in medical image analysis using kaapana. In *MICCAI: DeCaF*, pages 130–140. Springer, 2022.

[8] Fei Kong, Xiyue Wang, Jinxi Xiang, Sen Yang, Xinran Wang, Meng Yue, Jun Zhang, Junhan Zhao, Xiao Han, Yuhan Dong, et al. Federated attention consistent learning models for prostate cancer diagnosis and gleason grading. *Comput. Struct. Biotechnol. J.*, 23:1439, 2024.

[9] Lixin Liu, Kefeng Fan, and Mengzhen Yang. Federated learning: a deep learning model based on resnet18 dual path for lung nodule detection. *Multimedia Tools and Applications*, 82(11):17437–17450, 2023.

[10] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proc. 20th Int. Conf. Artif. Intell. Stat.*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282, 2017.

[11] Ashkan Moradi, Fadila Zerka, Joeran Sander Bosma, Derya Yakar, Jeroen Geerdink, Henkjan Huisman, Tone Frost Bathen, and Mattijs Elschot. Federated learning for prostate cancer detection in biparametric MRI: optimization of rounds, epochs, and aggregation strategy. In *SPIE Medical Imaging*, volume 12927, pages 412–421, 2024.

[12] Ashkan Moradi, Fadila Zerka, Joeran Sander Bosma, Mohammed RS Sunoqrot, Bendik S Abrahamsen, Derya Yakar, Jeroen Geerdink, Henkjan Huisman, Tone Frost Bathen, and Mattijs Elschot. Optimizing federated learning configurations for MRI prostate segmentation and cancer detection: A simulation study. *Radiology: Artificial Intelligence*, 7(5):e240485, 2025. doi: 10.1148/ryai.240485.

[13] Sarthak Pati, Ujjwal Baid, Brandon Edwards, Micah Sheller, Shih-Han Wang, G Anthony Reina, Patrick Foley, Alexey Gruzdev, Deepthi Karkada, Christos Davatzikos, et al. Federated learning enables big data for rare cancer boundary detection. *Nature communications*, 13(1):7346, 2022.

[14] Abhejit Rajagopal, Ekaterina Redekop, Anil Kemisetti, Rushikesh Kulkarni, Steven Raman, Karthik Sarma, Kirti Magudia, Corey W Arnold, and Peder EZ Larson. Federated learning with research prototypes: application to multi-center MRI-based detection of prostate cancer with diverse histopathology. *Academic Radiology*, 30(4): 644–657, 2023.

[15] Ashish Rauniyar, Desta Haileselassie Hagos, Debesh Jha, Jan Erik Håkegård, Ulas Bagci, Danda B Rawat, and Vladimir Vlassov. Federated learning for medical applications: A taxonomy, current trends, challenges, and future research directions. *IEEE Internet Things J.*, 11(5):7374–7398, 2023.

[16] Prashanth Rawla. Epidemiology of prostate cancer. *World J. Oncol.*, 10(2):63, 2019.

[17] Holger R Roth, Yan Cheng, Yuhong Wen, Isaac Yang, Ziyue Xu, Yuan-Ting Hsieh, Kristopher Kersten, Ahmed Harouni, Can Zhao, Kevin Lu, et al. Nvidia flare: Federated learning from simulation to real-world. *arXiv preprint arXiv:2210.13291*, 2022. https://github.com/NVIDIA/NVFlare.

[18] Anindo Saha, Joeran S Bosma, Jasper J Twilt, Bram van Ginneken, Anders Bjartell, Anwar R Padhani, David Bonekamp, Geert Villeirs, Georg Salomon, Gianluca Giannarini, et al. Artificial intelligence and radiologists in prostate cancer detection on MRI

(PI-CAI): an international, paired, non-inferiority, confirmatory study. *The Lancet Oncology*, 2024.

[19] Karthik V Sarma, Stephanie Harmon, Thomas Sanford, Holger R Roth, Ziyue Xu, Jesse Tetreault, Daguang Xu, Mona G Flores, Alex G Raman, Rushikesh Kulkarni, et al. Federated learning improves site performance in multicenter deep learning without data sharing. *J. Am. Med. Inform. Assoc.*, 28(6):1259–1264, 2021.

[20] Jonas Scherer, Marco Nolden, Jens Kleesiek, Jasmin Metzger, Klaus Kades, Verena Schneider, Michael Bach, Oliver Sedlaczek, Andreas M Bucher, Thomas J Vogl, et al. Joint imaging platform for federated clinical data analytics. *JCO clinical cancer informatics*, 4:1027–1038, 2020.

[21] Micah J Sheller, G Anthony Reina, Brandon Edwards, Jason Martin, and Spyridon Bakas. Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th Int. Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018*, pages 92–104. Springer, 2018.

[22] Micah J Sheller, Brandon Edwards, G Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R Colen, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Springer Scientific Reports*, 10(1):1–12, 2020.

[23] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer J. Clin.*, 71(3):209–249, 2021.

[24] Zhen Ling Teo, Liyuan Jin, Nan Liu, Siqi Li, Di Miao, Xiaoman Zhang, Wei Yan Ng, Ting Fang Tan, Deborah Meixuan Lee, Kai Jie Chua, et al. Federated machine learning in healthcare: A systematic review on clinical applications and technical architecture. *Cell Reports Medicine*, 5(2), 2024.

[25] J. C. Weinreb, J. O. Barentsz, P. L. Choyke, F. Cornud, M. A. Haider, K. J. Macura, D. Margolis, M. D. Schnall, F. Shtern, C. M. Tempany, H. C. Thoeny, and S. Verma. PI-RADS prostate imaging – reporting and data system: 2015, version 2. *European Urology*, 69(1):16–40, 2016.

[26] Jeffry Wicaksana, Zengqiang Yan, Xin Yang, Yang Liu, Lixin Fan, and Kwang-Ting Cheng. Customized federated learning for multi-source decentralized medical image classification. *IEEE J. Biomed. Health Inform.*, 26(11):5596–5607, 2022.

[27] Zengqiang Yan, Jeffry Wicaksana, Zhiwei Wang, Xin Yang, and Kwang-Ting Cheng. Variation-aware federated learning with multi-source decentralized medical image data. *IEEE J. Biomed. Health Inform.*, 25(7):2615–2628, 2020.