frontiers in Digital Health

TYPE Original Research
PUBLISHED 13 March 2026
DOI 10.3389/fdgth.2026.1691088

OPEN ACCESS

EDITED BY
Paolo Crippa,
Marche Polytechnic University, Italy

REVIEWED BY
Michel Audette,
Old Dominion University, United States
Colin Vanden Bulcke,
Université Catholique de Louvain, Belgium

*CORRESPONDENCE
Stijn Denissen
✉ stijn.denissen@vub.be

RECEIVED 22 August 2025
REVISED 29 January 2026
ACCEPTED 05 February 2026
PUBLISHED 13 March 2026

CITATION
Denissen S, Laton J, Grothe M, Vaneckova M, Uher T, Kudrna M, Horáková D, Baijot J, Penner I-K, Kirsch M, Motýl J, De Vos M, Chén OY, Van Schependom J, Sima DM and Nagels G (2026) Real-world federated learning for brain imaging scientists. Front. Digit. Health 8:1691088. doi: 10.3389/fdgth.2026.1691088

Real-world federated learning for brain imaging scientists

Stijn Denissen[1,2,3]*, Jorne Laton[1], Matthias Grothe[4], Manuela Vaneckova[2], Tomáš Uher[4], Matěj Kudrna[2], Dana Horáková[5], Johan Baijot[1], Iris-Katharina Penner[6], Michael Kirsch[7], Jiří Motýl[5], Maarten De Vos[8,9], Oliver Y. Chén[10,11], Jeroen Van Schependom[1,12], Diana Maria Sima[1,3] and Guy Nagels[1,13]

[1]AIMS Lab, Center for Neurosciences, UZ Brussel, Vrije Universiteit Brussel, Brussels, Belgium, [2]Department of Radiology, First Faculty of Medicine, Charles University, General University Hospital, Prague, Czechia, [3]icometrix, Leuven, Belgium, [4]Department of Neurology, University Medicine Greifswald, Greifswald, Germany, [5]Department of Neurology and Center of Clinical Neuroscience, First Faculty of Medicine, Charles University, General University Hospital, Prague, Czechia, [6]Department of Neurology, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland, [7]Institute for Diagnostic Radiology and Neuroradiology, University Medicine of Greifswald, Greifswald, Germany, [8]Department of Electrical Engineering (ESAT), KU Leuven, Leuven, Belgium, [9]Development and Regeneration, KU Leuven, Leuven, Belgium, [10]Département Médecine de Laboratoire et Pathologie (DMLP), Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland, [11]Faculté de Biologie et de Médecine (FBM), Université de Lausanne, Lausanne, Switzerland, [12]Department of Electronics and Informatics (ETRO), Vrije Universiteit Brussel, Brussels, Belgium, [13]St Edmund Hall, University of Oxford, Oxford, United Kingdom

Background: Federated learning (FL) has the potential to boost deep learning in neuroimaging but is rarely deployed in real-world scenarios, where its true potential lies. We propose FLightcase, a new FL toolbox tailored for brain research, and evaluate it on a real-world FL network to predict the cognitive status in patients with multiple sclerosis (MS) from brain magnetic resonance imaging (MRI).

Methods: We first trained a DenseNet neural network to predict age from T1-weighted brain MRI on three open-source datasets: IXI (586 images), SALD (491 images), and CamCAN (653 images). These were distributed across the three centres in our FL network: Brussels (BE), Greifswald (DE), and Prague (CZ). We benchmarked this federated model with a centralised version. The best-performing brain age model was then fine-tuned to predict performance on the symbol digit modalities test (SDMT) of patients with MS (Brussels: 96 images, Greifswald: 756 images, Prague: 2,424 images). Shallow transfer learning (TL) was compared with deep transfer learning, in which weights were updated either in the last layer or across the entire network, respectively.

Results: Federated training outperformed centralised training, predicting age with a mean absolute error (MAE) of 6.08 versus 7.02. Federated training yielded Pearson correlations (all $p < .001$) between true and predicted age of0.88 (IXI, Brussels), 0.91 (SALD, Greifswald), and 0.93 (CamCAN, Prague). Fine-tuning of the centralised model to SDMT was most successful with a deep TL paradigm (MAE = 9.19) compared to shallow TL (MAE = 11.05). Across Brussels, Greifswald, and Prague, deep TL predicted SDMT with MAEs of 10.71, 9.67, and 8.98, respectively, and yielded Pearson correlations between true and predicted SDMT of.25 ($p = 0.282$), 0.40 ($p < 0.001$), and 0.50 ($p < 0.001$).

Conclusion: Real-world federated learning using FLightcase is feasible for neuroimaging research in MS, enabling access to large MS imaging databases without sharing data. The federated SDMT-decoding model is promising and could be improved in the future by adopting FL algorithms that address the non-IID data issue and consider other imaging modalities. We hope our detailed real-world experiments and open-source distribution of FLightcase will prompt researchers to move beyond simulated FL environments.

KEYWORDS

BIDS, brain, brain age, cognition, deep learning, federated learning, multiple sclerosis

Introduction

Deep learning is gaining traction as a tool to study the brain's function and structure (1, 2). Since the comprehensive Nature review by Lecun et al. (3), interest among brain researchers has surged; more than 10,000 papers have been published on the topic in the past decade, compared just over 600 by the end of 2014.

Deep learning has led to major breakthroughs in brain analysis. Brain structures can now be accurately quantified using segmentation models (4, 5), aiding doctors in diagnosis and treatment evaluation. Deep neural networks also reduce the error in predicting age from brain magnetic resonance imaging (MRI) to the order of 2 years (6). This yields accurate biological aging clocks that capture deviations from healthy aging patterns and serve as efficient communication tools, allowing brain damage to be expressed in terms of "how much older the brain looks." Deep learning models can indeed capture the brain's complexity by creating high-dimensional, data-driven representations beyond human understanding. However, there is a catch. To create reliable representations, deep learning models require big training datasets, typically in the order of tens of thousands of images (6). As many dedicated researchers have invested significant effort in collecting datasets, we need to reconsider how they can be reused and combined to unlock the full potential of deep learning in brain image analysis.

The conventional way of training deep learning models involves centralising datasets. This is feasible for certain domains like brain age, as it relies on healthy control datasets that are publicly shared. Indeed, age-labelled T1-weighted images are widely available via initiatives such as the UK Biobank (7) and many repositories in OpenNeuro (8). In most other domains, however, particularly those working with sensitive patient data, data sharing is difficult. Barriers to data sharing are numerous, spanning in technical, motivational, economic, political, legal, and ethical terms (9). The latter three domains are especially relevant for centralising neuroimaging data. However, the first condition is trust: "In the absence of trust, providers could anticipate potential misinterpretation, misuse or intentional abuse of the data" (9). Second, the global data protection regulation (GDPR) enforces strict guidelines on data sharing. This can complicate the procedure, lower incentives, or block sharing entirely. This is especially true for the type of data that brain scientists work with, as faces can be reconstructed from MR images. Lastly, data sharing is relatively static. Data from routine clinical practice is continuously generated, and sharing those periodically is time-consuming and inefficient.

This conventional, centralised view on machine learning was challenged by McMahan et al. (10). They introduced the concept of federated learning (FL), in which models are trained at local institutions. Instead of sharing data, models are shared. Moreover, the computational load and data storage are spread across multiple centres. In the following years, brain researchers started experimenting with FL in simulated settings, primarily using open-source data. New algorithms were proposed (11), performance with respect to centralised training was explored (12), and federated learning toolboxes were designed (13). While FL in principle solves all previously mentioned issues, it generates new challenges that hinder brain scientists from deploy it in a real-world setting. These include financial constraints [e.g., graphical processing units (GPUs)], hardware and software differences, connectivity issues, and data heterogeneity. To date, only a few groups have succeeded, mainly in brain tumour segmentation (14, 15), where access to bigger datasets boosted model performance (15). While other real-world examples exist, it is often unclear from published methods whether they involved a simulation or real-world FL on geographically distributed data.

In this work, we aimed to pioneer real-world FL in our modelling domain, decoding cognitive performance from structural brain MRI in patients with multiple sclerosis (MS), using data from three international centres in Brussels (BE), Greifswald (DE), and Prague (CZ). When initiating the practical setup in early 2023, the most pressing challenge was the lack of software capable of orchestrating real-world federated learning in our neuroimaging context and of handling clinical datasets that differed widely in format. Therefore, we designed a simple, glass-box FL framework that allows easy real-world deployment in an international context: "FLightcase." FLightcase is designed to work with the Brain Imaging Data Structure [BIDS (16)], which has become the standard data organisational format for brain imaging data over the past decade (17). BIDS is enforced by data sharing platforms such as OpenNeuro (8), and we sought to continue this trend. By doing so, we aim to encourage brain AI researchers to adopt decentralised model training and leverage larger, multi-institutional datasets to boost generalisability.

The contributions of this paper are as follows:

- We introduce FLightcase, a simple, open-source, BIDS-compliant FL toolbox for neuroimaging.
- We demonstrate the real-world readiness of FLightcase on a real-world FL network with geographically distributed data from Brussels (BE), Greifswald (DE), and Prague

(CZ). The modelling goal was to predict cognitive function from brain MRI in MS using transfer learning (TL) from a pre-trained brain age model. We believe this is possible as cognitive deterioration naturally occurs with aging (18) and as we previously found a link between brain age and cognition in MS (19). If successful, the model could be employed as a cognitive screening tool on routine brain imaging, which is recommended at least yearly in people with MS (20). In the real-world experiments, we addressed two research questions:
  - Does federated training match centralised training in terms of model performance? This question was addressed in brain age modelling, as it relies on open-source data that can be centralised.
  - Does deep transfer learning (updating all network weights) outperform shallow transfer learning (updating only the weights of the fully connected layer) for predicting cognitive impairment in MS?
- We discuss challenges and solutions in setting up a real-world federated learning network to encourage the method in the field.

Methods

FLightcase in brief

FLightcase is a federated learning toolbox that was specifically designed to work with the brain imaging data structure (BIDS) (16). We aim to facilitate model training for brain researchers and stimulate researchers to use this data structure. Its basic communication relies on sending two files sequentially. The first file contains the machine learning information to be transmitted and the second is a text file marking transmission completion. Files are sent between computers via secure copy protocol (SCP), which requires all participating computers to be UNIX-based. In this paper, the terms "computer" and "node" are used interchangeably. The SCP command requires two localisers:

- The IP address of the receiving computer is required. An example of a secure network where computers are assigned an IP address, and are therefore reachable, is a virtual private network (VPN). This was used in our real-world example (cfr. infra).
- The receiver location for the file within the computer is also necessary. To facilitate this, we use the concept of an "FL workspace." The FL workspace is a folder dedicated to the federated learning experiment.

These and other settings are stored in a JavaScript Object Notation (JSON) file per computer. They are different for server and clients, for which templates can be found in Supplementary Table S1 and S2 respectively. For example, the client settings define the location of the BIDS dataset, while the server settings define the expected clients. To orchestrate the FL process, the server moreover stores an "FL plan" JSON file containing training preferences (e.g., number of FL rounds). Lastly, the server defines the model architecture that will be trained in a separate Python file.

For a detailed overview of the FLightcase software, refer to the "FLightcase unpacked" section of the Supplementary Material. FLightcase is publicly available in our AIMS-VUB GitHub repository (https://github.com/AIMS-VUB/FLightcase, branch "FL_POC") and is deployed on the Python Package Index (PyPI, https://pypi.org) for easy installation. The federated experiments in this study used FLightcase v0.1.17. FLightcase can be tested using the simulation provided in the Supplementary Material, section "How to test FLightcase."

FLightcase v0.1.17 depends on the following Python modules: torch v2.5.1 (21), pandas v2.2.3 (22), monai v1.4.0 (23), scikit-learn v1.6.0 (24), tqdm v4.67.1 (25), nibabel v5.3.2 (26), paramiko v3.5.0 (27), scp v0.15.0 (28), matplotlib v3.10.0 (29), scipy v1.14.0 (30), numpy v1.26.4 (31), click v8.1.8 (32), and twine v6.0.1 (33).

The federated learning network

To prove the real-world effectiveness of FLightcase, we tested it in an FL network across our AIMS labs in Brussels (BE), the Greifswald University Hospital (DE), and the General University Hospital Prague (CZ). The network (Figure 1) consisted of four computers, of which one is the server that coordinates the project and the other three serve as clients on which models are trained using the local data. The two Brussels computers were located in the same office and connected to the network of the Department of Electronics and Informatics (ETRO) at VUB. The computers in Greifswald and Prague were connected to this network via a VPN. Models were shared via secure copy protocol (SCP) with secure shell (SSH).

All client computers were equipped with graphical processing units (GPUs): Brussels: NVIDIA GeForce RTX 4090 (24GB), Greifswald: NVIDIA GeForce RTX 3090 (24GB); and Prague: NVIDIA GeForce RTX 4090 (24GB). The operating system of each computer (server and clients) was Debian GNU/Linux 12 ("bookworm"), and Python version 3.11.2 was used consistently.

Real-world FLightcase testing

Testing FLightcase on the real-world FL network involved two steps. Step 1 involved training a DenseNet convolutional network (34) to predict age from T1-weighted brain MRI, similar to the work of James Wood and colleagues (35). As step 1 was modelled on open-source data and thus allowed centralising of data, federated training was benchmarked against centralised training (explained at the end of the Methods section). Step 2 involved transfer learning of the best brain age model from step 1 [lowest overall test mean absolute error (MAE)] to predict cognitive impairment in people with MS.

The FL plans containing the hyperparameters for federated training are provided in Supplementary Table 4. We used a batch size of 10. Federated averaging (FedAvg) was used consistently, given its popularity in the field (36, 37) and in line with our emphasis on simplicity to enable real-world FL. At the end of each federated learning round, the FL server selected a sample of two out of three client models and performed a weighted average across both. The weight of each client model

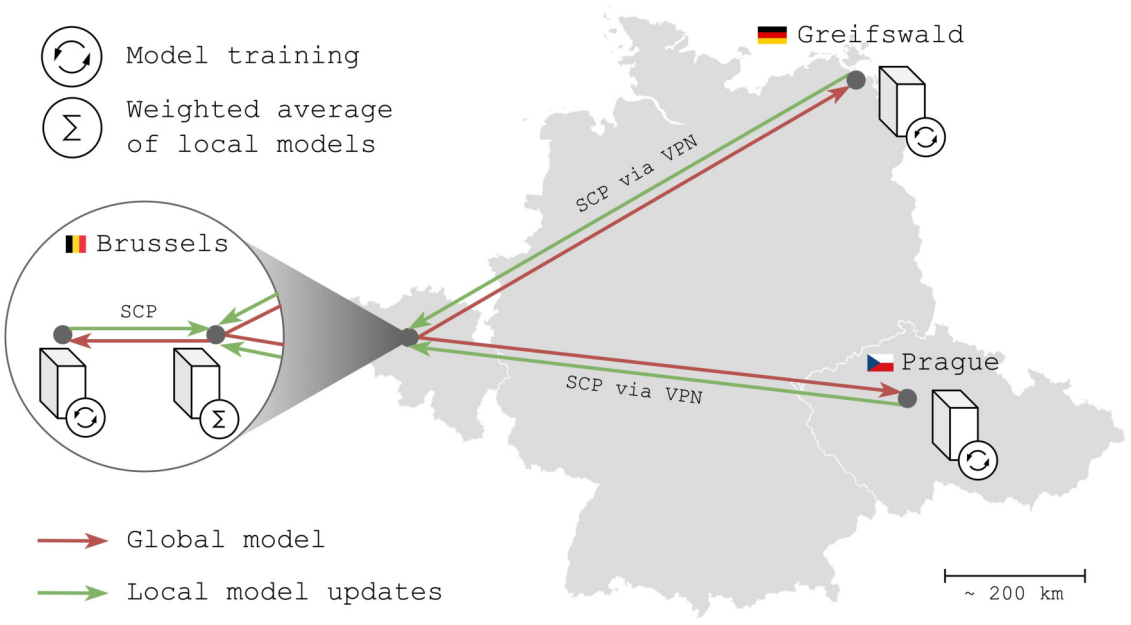

FIGURE 1
The federated learning network for a real-world experiment. The computer with the "sigma" symbol is the server, whereas computers with an "update" symbol are clients. SCP, secure copy protocol; VPN, virtual private network. Belgium, Germany and Czech Republic flags from flagcolorcodes (https://www.flagcolorcodes.com/). Icons from JGraph (https://github.com/jgraph), licensed under CC BY 4.0

was defined by the sample size of that client divided by the total sample size of both clients in the sample.

Ethics

The "Commissie Medische Ethiek" (CME) of UZ Brussel judged this retrospective study to be exempt from ethical approval (B.U.N. 1432022000303). For the MS data at each centre in this study, ethical approval was obtained prior to data acquisition (Brussels: B.U.N. 143201423263, Greifswald: BB159/18, Prague: 113/22 S-IV and 28/17). Healthy control data used to train the brain age model was obtained from open-source datasets.

Data

To train the brain age network (step 1), we used three open-source datasets: the Cambridge Centre for Ageing and Neuroscience (CamCAN) dataset (38), the Information eXtraction from Images (IXI) dataset (39), and the Southwest University Adult Lifespan Dataset (SALD) (40). The CamCAN dataset was stored on the Prague computer, the IXI dataset in Brussels and the SALD dataset in Greifswald. The datasets contained T1-weighted MRI, sex at assessment, and age from healthy subjects. The data are summarised in Table 1.

The cognition prediction network (step 2) was trained on three MS datasets at the Vrije Universiteit Brussel (VUB), the General University Hospital Prague (VFN), and Greifswald University Hospital. The data were organised locally in the BIDS format and contained T1-weighted MR images, demographics, and clinical information. This contained sex at assessment, age, expanded disability status scale [EDSS (41), physical disability], disease duration, MS subtype (relapsing versus progressive onset), and the symbol digit modalities test [SDMT (42)], i.e., the target to predict. We chose the SDMT as it is the most sensitive to cognitive impairment in MS (42, 43) and as we previously found a link between brain age and SDMT in people with MS (19). The popularity of the SDMT in MS research and care is reflected in its presence in all modern cognitive test batteries (42). As a result, choosing the SDMT enabled us to extract the largest possible decentralised cognition-labelled MRI database. In the SDMT, a subject is presented a list of symbols that need to be converted to numbers using a key at the top of the page, matching symbols with numbers. In 90 s, the subject must convert as many symbols to numbers as possible, each time saying the number out loud for the test administrator to write down. The SDMT is a measure of information processing speed.

The T1-weighted MR images were pre-processed using the pipeline described by Wood et al. (35). The pipeline first aligns a volume to the right anterior superior (RAS) orientation, and then applies skull-stripping using the HD-BET brain extraction algorithm (44). The image is then bias field-corrected using the N4 algorithm from Advanced Normalisation Tools [ANTs, (45)]. Also using ANTs, the image is affine registered to Montreal Neurosciences Institute (MNI) 152 space (1 mm isotropic) using all 12 degrees of freedom (rotation, translation, scaling, and shearing). After ensuring that the volume is in RAS orientation, the image is resampled with a voxel spacing of 1.4 and cropped/zero-padded to a window of $130 \times 130 \times 130$ voxels. All steps are defined in the "pre_process.py" script (https://github.com/MIDIconsortium/BrainAge/blob/main/pre_process.py). An updated version of this script was included in our GitHub repository with permission from the authors (35).

The data were randomly split into 80% train and 20% test data on each client. To prevent data leakage, multiple images of a single subject were collected in either the training or test set. During each

TABLE 1 Dataset characteristics.

Variable	Open-source (step 1)			Closed-source (step 2)		
	CamCAN	IXI	SALD	Brussels	Greifswald	Prague
N	653	586	491	96	338	916
N image—SDMT pairs	653	586	491	96	756	2,424
Sex at assessment (m:f)	323:330	258:328	185:306	27:69	110:228	275:641
Age at scan (*M* ± SD)	54.3 ± 18.6	49.4 ± 16.7	45.2 ± 17.5	47.9 ± 9.9	43.5 ± 12.1	42.2 ± 9.5
SDMT (*M* ± SD)	/	/	/	48.0 ± 11.6	51.7 ± 15.3	58.9 ± 12.0
EDSS (Med; IQR)	/	/	/	3; 2	2; 2	2.5; 2
Disease duration (*M* ± SD)	/	/	/	15.5 ± 8.5	9.2 ± 6.9	13.2 ± 8.1
Type MS	/	/	/	CIS: 2 RRMS: 81, SPMS: 6, PPMS: 7	CIS: 7 RRMS: 680, SPMS: 41, PPMS: 23, RIS: 5	CIS: 448, RRMS: 1566, SPMS: 363, PPMS: 38, PRMS: 6
T1w MR images						
Manufacturer	Siemens	Philips, GE	Siemens	Philips	Siemens	Siemens
Model	TrioTim	Philips: Intera, Gyroscan Intera	TrioTim	10. Achieva: 31 Ingenia: 65	Verio	Skyra
Echo time (ms)	2.98	4.603	2.52	2.303; 2.287	2.58	2.96
Repetition time (ms)	2.25	9.600; 9.813	1.9	4.952; 5.189	1.900	2.300
Slice thickness (mm)	1	/	/	1.0	0.90	1
Flip angle (degrees)	9	8	9	8	9	9
Field strength (*T*)	3	3 and 1.5	3	3	3	3

n, sample size; m, male; f, female; *M*, mean; SD, standard deviation; SDMT, symbol digit modalities test; EDSS, expanded disability status scale; GE, general electric. (1) Variable distributions are calculated across all image-SDMT pairs. (2) Missing values: 14 EDSS (Prague), 3 disease course (Prague), and 15 disease duration (Greifswald). (3) The MR image acquisition info was extracted from a single subject per scanner model and might therefore deviate among subjects. For IXI, which was acquired at three different sites, MRI acquisition info was obtained from the website: https://brain-development.org/ixi-dataset/ (accessed March 10, 2025). Detailed scanner info was only available for the Philips scanners.

FL round, the training data were bootstrapped five times, with 75% for training and 25% for validation. This resulted in a train/validation/test split of 60/20/20 (cfr. Supplementary Table 4).

The 3D DenseNet model

The 3D Dense Convolutional Network (DenseNet) (34) was used during both the initial brain age task (step 1) and the transfer learning task to SDMT (step 2). DenseNet has outperformed other network architectures in a medical imaging context (46) and is unique in directly connecting all layers inside the network with each other (34). Each layer therefore takes all previous feature maps—outputs of previous layers—as input, improving the propagation of features throughout the network. The DenseNet architecture moreover reduces the "vanishing gradient" problem, where gradients used to update weights in the network gradually approach zero during backpropagation to earlier layers. Lastly, the model reduces the number of parameters in the network (34). The 3D DenseNet model in this paper has 11,243,649 updatable parameters and is schematically illustrated in Figure 2.

For the brain age prediction task, all layers in the network were unfrozen, whereas for the SDMT prediction task, two transfer learning methods were explored. In the "shallow TL" task, only the fully connected layer, consisting of 1,025 parameters (1,024 weights and one bias), was updated during training (cfr. Figure 2); the other layers—the feature extractor parts of the network—were frozen. In the "deep TL" task, all layers of the network were updated. Figure 2 also summarises the transfer learning methodology in the box on top.

Evaluation

The final performance of all models was evaluated on the test dataset per client using the MAE and Pearson correlation. Overall model performance was calculated using Equation 1:

$$\mathrm{MAE}_{\text{test, overall}} = \sum_{i=0}^{m} \frac{\mathrm{MAE}_{\text{test},i} * n_i}{N}$$

Equation 1 Overall test MAE calculation, where m is the number of clients and N is the total sample size across clients.

Lastly, for the brain age analyses, we also reported the Pearson correlation between the brain age difference [BAD, predicted age (brain age) minus the calendar age] and calendar age to investigate a potential bias (47).

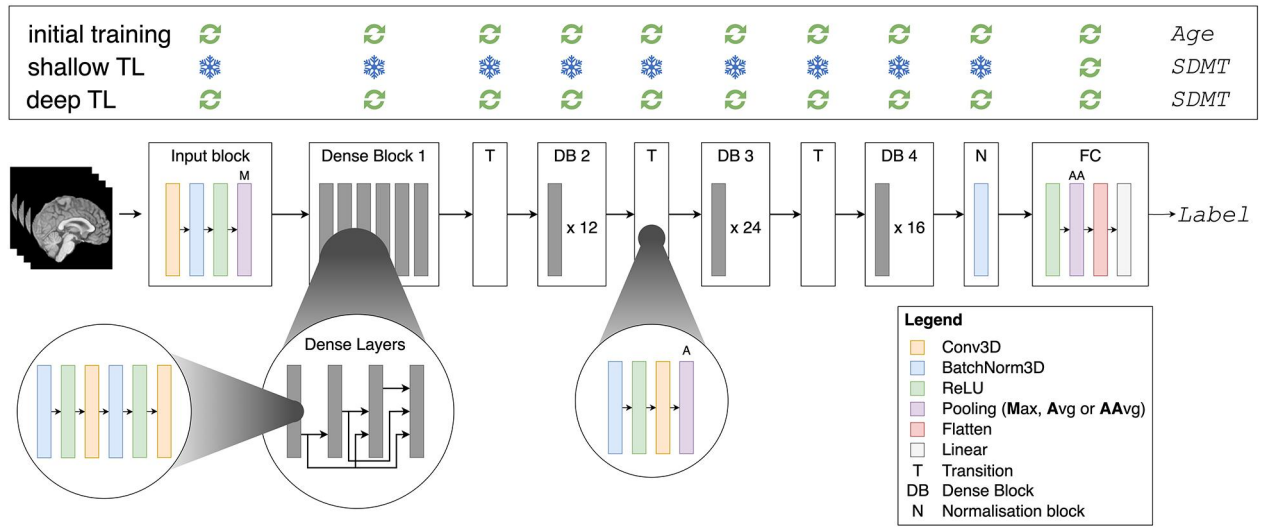

FIGURE 2
Densenet architecture and training methodology (top panel). Avg, average; AAvg, adaptive average. The "Label" is the chronological age for the initial brain age prediction task, and the SDMT for the shallow and deep transfer learning (TL).

Benchmarking: centralised brain age training

As a benchmark model, we additionally updated the brain age model on a centralised version of the three open-source datasets. Here, we mimicked federated training as closely as possible by syncing the hyperparameters with the federated experiment (Supplementary Table 4), aggregating the test datasets to a centralised version, and using an aggregated version of the same train/validation splits used in federated training. As the centralised setting had access to all test results together, the test MAE was calculated across all subjects simultaneously instead of the per-client approach in Equation 1.

Results

Step 1: predicting brain age

Real-world federated training

The total FL process took 1 h, 48 min, and 21 s to complete. The model reached a minimum after 48 rounds and training stopped early after 27 rounds. The final network achieved an overall test MAE of 6.08, and an MAE of 5.91, 6.15, and 6.22 on the test datasets of CamCAN, IXI, and SALD, respectively. Pearson correlations for the respective datasets were 0.93 ($p < 0.001$), 0.88 ($p < 0.001$), and 0.91 ($p < 0.001$). Pearson correlations between age and BAD were −0.59 ($p < 0.001$), −0.37 ($p < 0.001$), and −0.50 ($p < 0.001$), respectively. The training process is visualised in Figure 3, and scatterplots of true versus predicted age (brain age) are displayed in Figure 4.

Centralised benchmark

Centralised training on the three open-source datasets took 3 h, 20 min, and 33 s. The model reached a minimum after 18 rounds and was stopped early after 38 rounds. The model achieved an overall test MAE of 7.02, a Pearson correlation (age and brain age) of 0.87 ($p < 0.001$), and a Pearson correlation (age and BAD) of −0.53 ($p < 0.001$). On CamCAN, IXI, and SALD, respectively, the model predicted brain age with a test MAE of 6.72, 6.78, and 7.70; a Pearson correlation (age and brain age) of 0.91 ($p < 0.001$), 0.86 ($p < 0.001$) and 0.86 ($p < 0.001$); and a Pearson correlation (age and BAD) of −0.71 ($p < 0.001$), −0.35 ($p < 0.001$), and −0.44 ($p < 0.001$). The overall results are displayed in Supplementary Figure 2 (loss figure) and Supplementary Figure 3 (test results scatterplot).

Step 2: predicting SDMT

Figure 5 illustrates the training process for the SDMT models, using shallow TL (left) and deep TL (right), starting from the centralised brain age model. Figure 6 shows the results of applying the final model to the test set of each centre.

Shallow TL took 3 h, 21 min, and 3 s. The model reached a minimum average validation MAE after 18 FL rounds and achieved an overall test MAE of 11.05 SDMT points. The per-client test MAEs were 13.08 (Brussels), 12.71 (Greifswald), and 10.45 (Prague), whereas the Pearson correlations were −0.17 ($p = 0.467$), −0.28 ($p < .001$), and −0.43 ($p < .001$), respectively.

For the deep TL model (right), training took 8 h, 4 min, and 35 s. The model reached a minimum at FL round 49 and achieved an overall test MAE of 9.19 SDMT points. The per-client test MAEs were 10.71 (Brussels), 9.67 (Greifswald), and 8.98 (Prague), whereas the Pearson correlations were 0.25 ($p = 0.282$), 0.40 ($p < 0.001$), and 0.50 ($p < 0.001$), respectively.

Discussion

In this manuscript, we introduced FLightcase, a federated learning toolbox specifically designed to promote real-world,

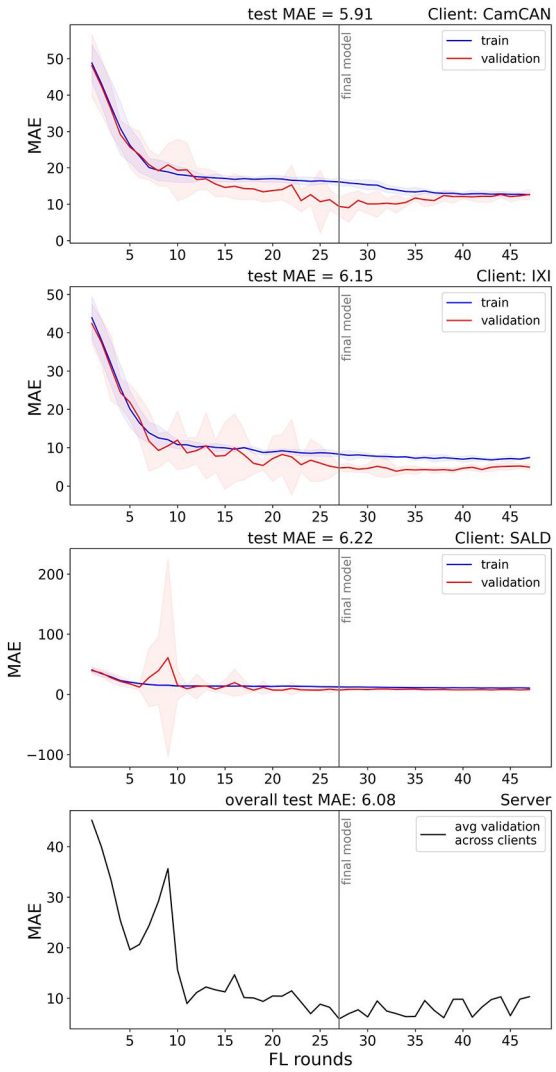

FIGURE 3
Federated training of the brain age network. The red (train) and blue (validation) lines represent the average train and validation MAE, respectively, and the shaded red and blue areas represent the 95% confidence interval, all calculated across five bootstraps per FL round. The right lower panel shows the average validation MAE across the clients in the aggregation, calculated on the server side.

decentralised machine learning for brain scientists. We then demonstrated its real-world readiness by training models to predict SDMT from T1-weighted brain MRI images, using transfer learning from a brain age model. During brain age modelling, federated training outperformed centralised training. During transfer learning, a deep paradigm—where all network weights were trained—outperformed shallow learning—where only the last network layer was updated. The final model predicted SDMT with an average error of 9.19 points and performed especially well on the Greifswald and Prague datasets.

Decentralised brain age modelling

The final federated brain age model predicted age from T1w MRI of healthy controls with an average error of 6.08 years across datasets. Strong correlations were observed on each client test dataset, indicating that federated learning successfully sensitised the model to individual differences in brain structure. Federated learning also outperformed centralised training with a decreased MAE of 0.94 years. Notably, in earlier experiments, where train and validation splits were not harmonised between federated and centralised experiments, the inverse was observed. An FL simulation study by Basodi et al. found similar performances of brain age models trained on decentralised and centralised datasets (12). In their study, besides a larger dataset of over 10,000 images from two sources, data from each source was distributed across six simulated centres, contrasting with our design of one source per centre. Overall, these results indicate that federated training can be a compelling alternative to a centralised training paradigm when data cannot be shared.

Our brain age models may suffer from low generalisability due to the nature of the open-source datasets, which contained both Caucasian (CamCAN and IXI) and Chinese (SALD) brains, known to differ in structure (48, 49). The model indeed had a higher test MAE in SALD (6.22) compared to CamCAN (5.91) and IXI (6.15). Although the test performance per sample was similar, the zigzagging pattern at the end of training (Figure 3) reveals the importance of client sampling; valleys coincided with IXI and SALD being included in the sample. This zigzagging pattern was especially clear in the last rounds (Figure 3). This highlights the importance of considering alternative client sampling schemes such as FedSampling (50). Lastly, the combined size of the decentralised dataset was 1,730 images, while the best model reported in a comprehensive review on brain age model performance (6) was trained by Peng et al. on 14.503 T1-weighted brain images (51). Enlarging our dataset would likely have improved brain age prediction performance and reduced discrepancies between centralised and federated training, such as observed in Basodi et al. (12).

Finally, our brain age models displayed the well-described bias in which brain age is inherently overestimated in younger individuals and underestimated in older individuals. This was evident from the negative correlations between BAD and age. Research often corrects for this bias (52), although the effectiveness of these methods for downstream tasks has been questioned by Zhang et al. (47). Although it is beyond the scope of this manuscript to further discuss brain age correction, we underline its importance when using the models for downstream tasks.

Transfer learning to SDMT

Deep transfer learning from the centralised brain age model outperformed shallow transfer learning, with an overall MAE of 9.19 versus 11.05. This indicates that the latent feature representation for the brain age task was incompatible with the SDMT prediction task, requiring an update of the feature extractor part of the network.

Although deep TL substantially improved performance, the model had difficulty decoding SDMT performance on the Brussels dataset. While training and validation loss, as well as confidence intervals, gradually decreased in Greifswald and Prague, loss increased in the Brussels dataset. Not unlike federated brain age modelling (cfr. supra), this translates into

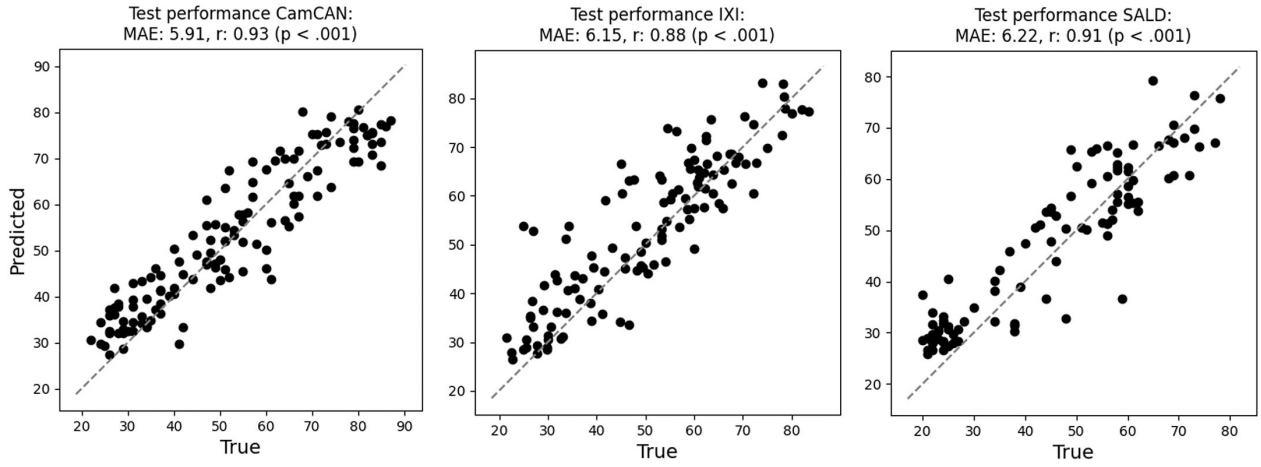

FIGURE 4
Scatterplots between true and predicted age (brain age) on the test datasets of CamCAN (stored in Prague), IXI (stored in Brussels), and SALD (stored in Greifswald). The grey dashed line indicates the perfect prediction line where age equals brain age.

heavy loss fluctuations on the server side, depending on which centre models are in the aggregation sample. Valleys coincided with the local models of Greifswald and Prague in the aggregation sample. We hypothesise two factors underlying this behaviour.

First, the Greifswald and Prague datasets were larger. The contributions of their local models in each round were therefore bigger than the Brussels local model, causing the model to learn less from the Brussels dataset. This is an inherent problem of the conventional FedAvg paradigm, which is why solutions for equality of local model updates have started to emerge. An example of this is q-FedAvg, which "reweights" the loss by a parameter q (53). The result is that client models with higher loss receive a higher weight in the aggregation.

Second, the behaviour nicely illustrates the non-IID problem in federated learning, meaning that we cannot assume that data points are drawn from the same underlying sample (54). This violates an important assumption in machine learning (55). Indeed, MS centres work with different equipment and workflows, which can be expected to increase across country borders. As illustrated in Table 1, different MRI scanners were used across different centres, and MS samples differed, most prominently in the SDMT ground truth distribution. This underscores the importance of data harmonisation and understanding between-centre differences, including test administration practices. It is therefore encouraging to observe the emergence of open-source federated harmonisation toolboxes like FedHarmony (65), which should be considered in future research for optimal model performance.

The road ahead for real-world FL in neuroimaging

Federated learning offers a compelling alternative to centralised machine learning. It distributes computational load and addresses data sharing—one of the major impediments to international scientific collaborations (56). The promises for federated learning in the medical field are substantial, with some of the most notable breakthroughs emerging from cancer research. One of the first large-scale real-world FL demonstrations was reported in 2022 by Pati and colleagues (15). In a global federated learning setting involving 71 institutions, they updated a publicly available algorithm to detect glioblastoma boundaries and obtained a considerable gain in the validation Dice score across clients of about 25%. The authors stated the following: "It is the use of FL that successfully enabled (i) access to such an unprecedented dataset of the most common and fatal adult brain tumour, and (ii) meaningful ML training to ensure the generalisability of models across out-of-sample data." Indeed, although Vo et al. confirmed that comparable results can be obtained in centralised contexts in ideal scenarios (57), the key benefit of FL is that it can still achieve these results when this is not the case. Similarly, technical advances are emerging that enable one to address data heterogeneity across institutions in a federated setting, such as tackling the non-IID issue by a combination of distributed gradient blending and proximity-aware client weighting (58).

Considering these medical FL breakthroughs, it is even more surprising that literature often ignores the very first step to achieving these results: how to build an FL network across clinical institutions. This practical step appears to be an assumption, as the literature focusses on challenges after a network has been established (59). In our experience, however, building a real-world FL network is the key roadblock between the compelling idea of federated learning and its practical realisation, and should be the key focus to usher in an era of real-world decentralised machine learning beyond simulation.

Collaboration

A first prerequisite for building an FL network is a trust among all partners. With malicious intent, source data could, for example, be shared between computers via VPN. The team should moreover be multidisciplinary, containing both medical experts and technical experts in information technology (IT) and data science. Real-world federated learning poses unique IT challenges in terms of communication and encryption, in

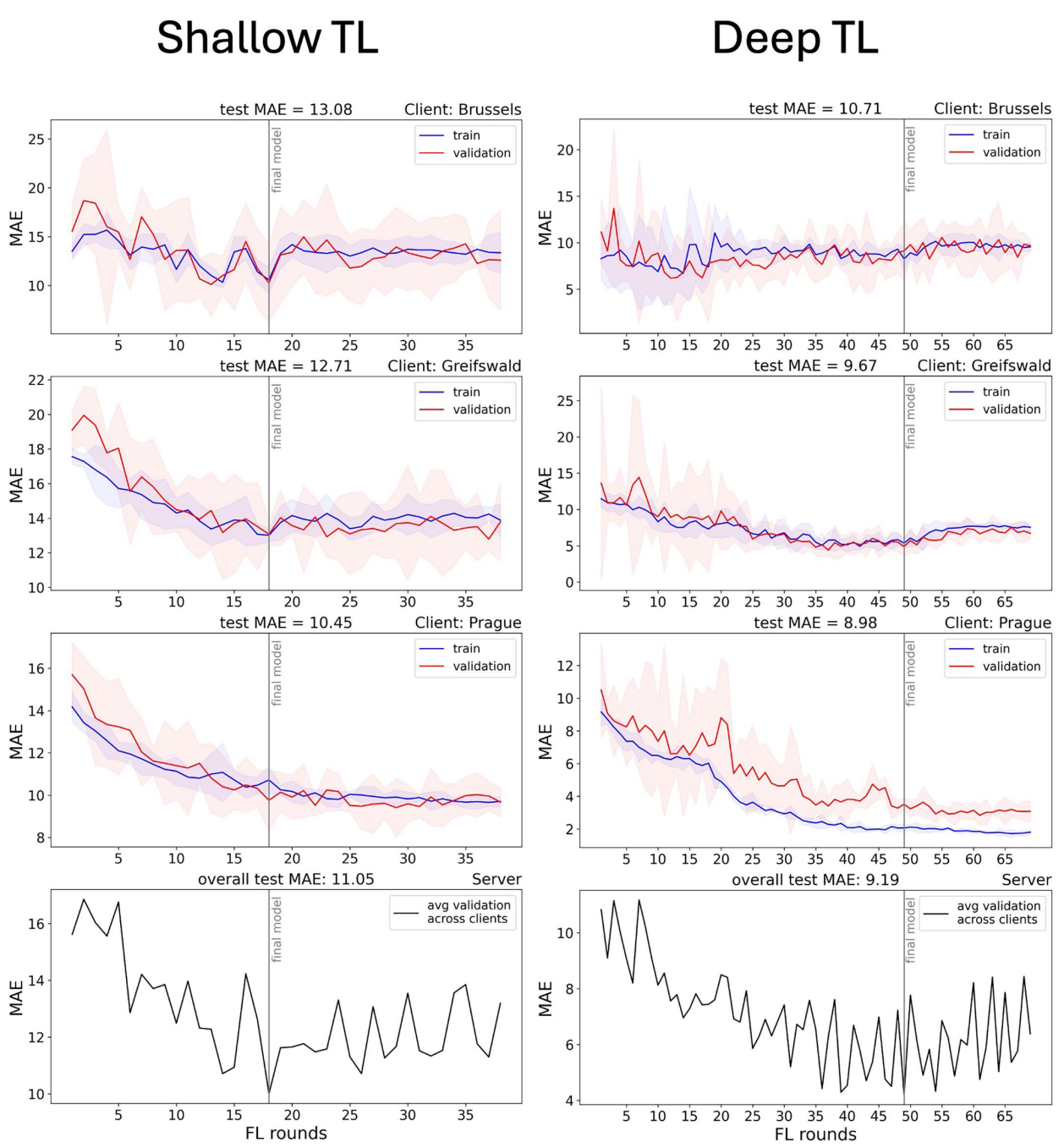

FIGURE 5
Loss figure of fine-tuning the brain age model to predict SDMT using shallow (left) and deep (right) transfer learning. FL, federated learning; MAE, mean absolute error; avg, average. The lower panels show the average validation MAE across the clients in the aggregation, calculated on the server side.

addition to other existing technical challenges in deep learning research. On the other hand, medical expertise is required in pre-processing and interpreting the data, as well as understanding the model and its output.

Hardware

Besides sufficient storage, updating neural networks such as the DenseNet model (over 11 million parameters) requires processing units like a graphical (GPU) or tensor processing unit (TPU) on each client computer. In our network, all client computers were equipped with a GPU with 24GB of random access memory (RAM).

Software

There is a lack of FL toolboxes that are rigorously tested in real-world circumstances, with most designed as general frameworks rather than specifically tailored to medical data needs. We began designing FLightcase in early 2023, when real-

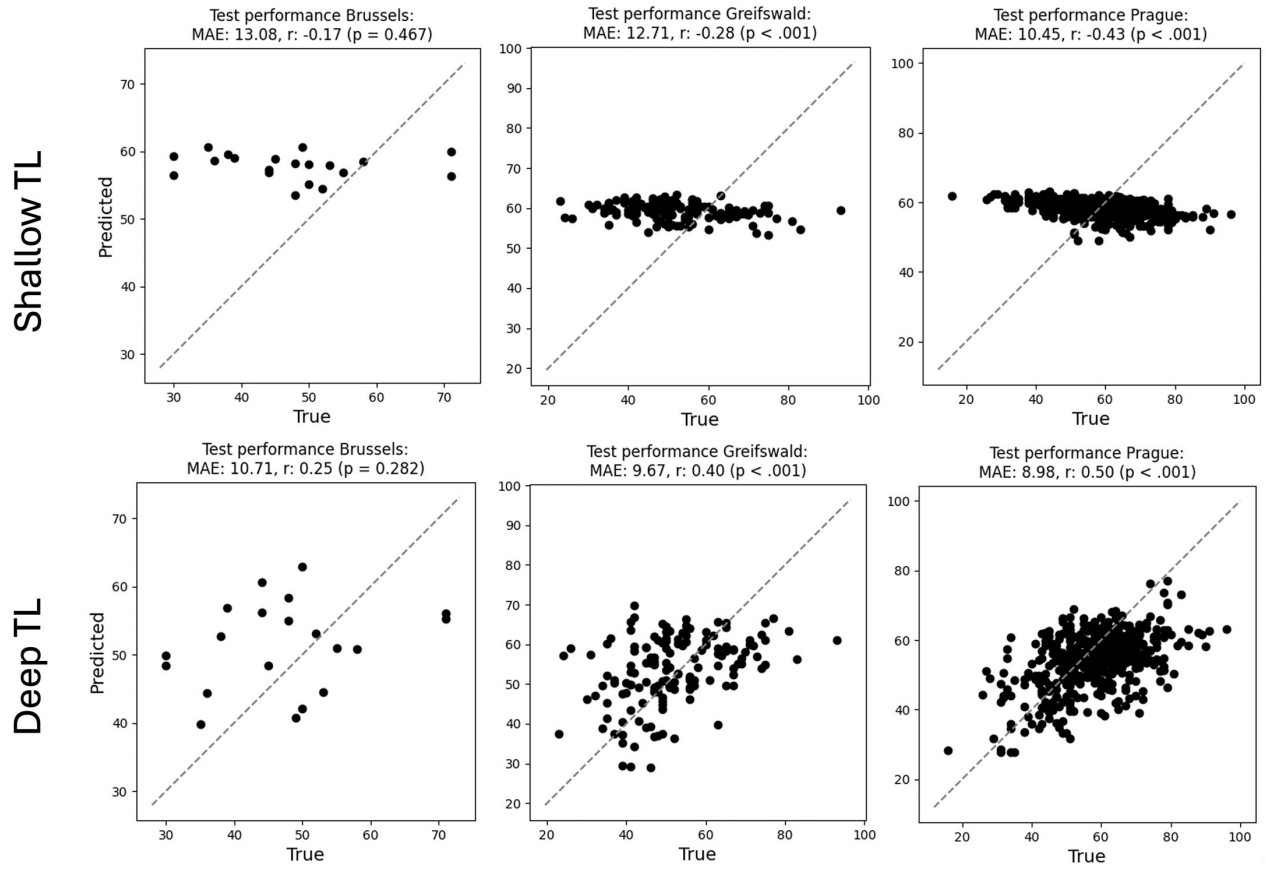

FIGURE 6
Scatterplots visualising test performance of the SDMT model trained with shallow TL (upper row) and deep TL (lower row).

world FL examples in neuroimaging had just started to emerge (15). Although some toolboxes such as Flower (13) and OpenFL (60) were available open source, we experienced difficulties in real-world deployment on our own network. We therefore set out to design a framework relying on a simple but stable connection between computers, focussing on the core needs of our methodology, such as compliance with BIDS, support for loading anatomical brain images, and focussing on the original and most popular FL algorithm, FedAvg. Since then, toolboxes such as Flower have further developed into communities with peer support. Awareness about the difficulty of real-world FL is increasing, reflected by companies facilitating deployment such as ScaleOut, and efforts are underway to tailor toolboxes to the specific needs of different modelling domains. Almost a decade after the pioneering work of Plis and colleagues (61) on standardised analysis of decentralised neuroimaging data, FL software is approaching a technology readiness level where centralised workflows can be seamlessly converted to federated ones. For FLightcase in particular, we have reconsidered the basic communication layer, allowing file sharing by uploading to and downloading from a Flask web server.

Connectivity and remote access

All computers in our network were connected to locally available internet providers. This introduces new centre-specific difficulties pertaining to local security settings. Blocking web addresses such as TeamViewer, for example, complicated remote access. Combined with limited scalability due to a maximum number of remote computers in TeamViewer, we switched to tmate (cfr. methods). The drawback of this approach is the absence of a graphical interface; remote computers are operated via command line.

Financial

Ensuring the aforementioned requirements comes with a significant cost. In particular, the hardware for each client with expensive GPUs imposes additional financial restraints, limiting the feasibility of setting up a network.

Limitations

Federated averaging (FedAvg) was the original aggregation algorithm when federated learning was introduced by H. Brendan McMahan and colleagues in 2016. Ten years after its introduction, the algorithm remains popular and attractive for its simplicity, which resonates with this manuscript's core message of facilitating real-world federated learning. Over the years, however, FedAvg has been criticised for reduced performance under non-IID circumstances (62). Newer

algorithms—such as the "federated learning approach based on a greedy algorithm" [FedGA, (62)]—show promise in overcoming this limitation.

Final considerations

The goal of any medical AI endeavour is to deliver models that can be used in a clinical setting. Simultaneously, we strive for generalisability of models. In a multi-centre federated learning setting, models may perform worse in a single-centre context as they attempt to find the "middle ground." Going back to the core objective of delivering clinically ready AI tools, the question can thus be raised whether overfitting on single-centre data is truly problematic. Moreover, FL risks encoding centre-specific bias pertaining to, for example, test administration and MR scanner properties. The central challenge is therefore how to best use large multi-centre data.

In line with the current trend of building so-called "foundation models" (63), the way forward may be to use a transfer learning paradigm where multi-centre data are first used to create a strong foundation model that encodes as much centre-independent information as possible. Transfer learning can then be employed to fine-tune the model on single-centre datasets. In this manuscript, we attempted to do so by creating a "brain age foundational model." Although we achieved fair performance on the downstream SDMT task, we consider an "SDMT foundational model" possible by investing in real-world FL. By offering scientists a starting point to engage in real-world FL, we hope our manuscript facilitates the creation of strong cognitive screening models across diverse clinical settings.

Data availability statement

The data analysed in this study are subject to the following licenses/restrictions: data underlying the real-world and simulated brain age experiments are open source available. Other datasets are accessible upon reasonable request. Requests to access these datasets should be directed to stijn.denissen@vub.be.

Ethics statement

The studies involving humans were approved by overarching ethical approval: B.U.N. 1432022000303 ("Commissie Medische Ethiek," Brussels). Ethical approvals were obtained at each participating centre: Prague: 113/22 S-IV and 28/17; (2) Greifswald: BB159/18; and (3) Brussels, "Commissie Medische Ethiek": B.U.N. 143201423263. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

SD: Funding acquisition, Writing – original draft, Methodology, Software, Formal analysis, Conceptualization, Visualization, Data curation, Investigation, Validation. JL: Software, Methodology, Conceptualization, Writing – review & editing, Investigation. MG: Investigation, Resources, Writing – review & editing. MVa: Investigation, Resources, Writing – review & editing, Project administration, Funding acquisition. TU: Investigation, Writing – review & editing, Project administration, Resources. MaK: Investigation, Methodology, Writing – review & editing, Resources. DH: Resources, Investigation, Project administration, Writing – review & editing, Funding acquisition. JB: Writing – review & editing, Investigation, Funding acquisition, Methodology. I-KP: Investigation, Writing – review & editing, Resources. MiK: Resources, Investigation, Writing – review & editing. JM: Writing – review & editing, Investigation, Resources. MD: Writing – review & editing, Conceptualization. OC: Project administration, Writing – review & editing, Conceptualization, Methodology. JV: Resources, Project administration, Methodology, Investigation, Supervision, Funding acquisition, Writing – review & editing, Conceptualization. DS: Project administration, Methodology, Supervision, Investigation, Writing – review & editing, Funding acquisition, Conceptualization. GN: Funding acquisition, Resources, Project administration, Conceptualization, Investigation, Writing – review & editing, Supervision, Methodology.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This study was funded by a personal industrial PhD grant (Baekeland, HBC.2019.2579) awarded by Flanders Innovation and Entrepreneurship to SD; a personal travel grant (V412023N) awarded by the "Fonds Wetenschappelijk Onderzoek" (FWO) to SD for his stay in Prague in the context of this study; a grant (SRP85) from the Vrije Universiteit Brussel; and a junior post-doctoral grant (12A6U25N) from the FWO Flanders. This project was furthermore funded by an IOF-POC grant (IOFPOC57) from the Vrije Universiteit Brussel (VUB); an ITEA grant (20030 HeKDisco, HBC.2021.0500) from Flanders Innovation and Entrepreneurship; institutional support from the Czech Ministry of Health (RVO-VFN 64165); and funding from the National Institute for Neurological Research, Czech Republic, Programme EXCELES, ID Project No. LX22NPO5107, supported by the European Union—Next Generation EU. GN is a senior clinical research fellow of the FWO Flanders (1805620N).

Acknowledgments

We would like to thank JL and Robert Malinowski for their help in setting up and maintaining the hardware used in the federated learning network. We also thank Luc Van Kempen for IT support and André Vital Serafim Silva for insights into centralized brain age modelling generated during his master thesis. The Python package GADM v0.0.3 (https://pypi.org/project/gadm/0.0.3/), was used to create Figure 1. This article was submitted as a preprint to medRxiv as Denissen et al. 2023 (64).

Conflict of interest

This work was partly performed during the industrial PhD project of SD in collaboration with icometrix. GN reports a relationship with icometrix that includes equity or stocks. DS reports a relationship with icometrix that includes employment. MVa reports a relationship with Biogen Idec that includes consulting or advisory, speaking and lecture fees, and travel reimbursement; Novartis that includes consulting or advisory, speaking and lecture fees, and travel reimbursement; Roche that includes consulting or advisory, speaking and lecture fees, and travel reimbursement; Merck that includes consulting or advisory, speaking and lecture fees, and travel reimbursement; and Teva that includes consulting or advisory, speaking and lecture fees, and travel reimbursement. TU reports a relationship with Biogen that includes speaking and lecture fees and travel reimbursement, Novartis that includes speaking and lecture fees and travel reimbursement, Sanofi that includes funding grants and travel reimbursement, Roche that includes speaking and lecture fees and travel reimbursement, Merck Serono that includes travel reimbursement, and Biogen Idec that includes funding grants. DH reports a relationship with Biogen Idec that includes consulting or advisory, funding grants, speaking and lecture fees, and travel reimbursement; Novartis that includes consulting or advisory, speaking and lecture fees, and travel reimbursement; Merck that includes consulting or advisory, speaking and lecture fees, and travel reimbursement; Bayer that includes consulting or advisory, speaking and lecture fees, and travel reimbursement; Sanofi Genzyme that includes consulting or advisory, speaking and lecture fees, and travel reimbursement; Roche that includes consulting or advisory, speaking and lecture fees, and travel reimbursement; and Teva that includes consulting or advisory, speaking and lecture fees, and travel reimbursement. JM reports a relationship with Sanofi Genzyme that includes speaking and lecture fees and travel reimbursement, Biogen that includes speaking and lecture fees and travel reimbursement, Novartis that includes speaking and lecture fees and travel reimbursement, and Merck that includes speaking and lecture fees and travel reimbursement.

The remaining author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The authors TU, MG, IK-P declared that they were an editorial board member of Frontiers at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth.2026.1691088/full#supplementary-material

References

1. Hirano R, Asai M, Nakasato N, Kanno A, Uda T, Tsuyuguchi N, et al. Deep learning based automatic detection and dipole estimation of epileptic discharges in MEG: a multi-center study. *Sci Rep*. (2024) 14(1):24574. doi: 10.1038/s41598-024-75370-9

2. Joo Y, Namgung E, Jeong H, Kang I, Kim J, Oh S, et al. Brain age prediction using combined deep convolutional neural network and multi-layer perceptron algorithms. *Sci Rep*. (2023) 13(1):22388. doi: 10.1038/s41598-023-49514-2

3. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. (2015) 521(7553):436–44. doi: 10.1038/nature14539

4. Ranjbarzadeh R, Bagherian Kasgari A, Jafarzadeh Ghoushchi S, Anari S, Naseri M, Bendechache M. Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images. *Sci Rep*. (2021) 11(1):10930. doi: 10.1038/s41598-021-90428-8

5. Simarro J, Meyer MI, Van Eyndhoven S, Phan TV, Billiet T, Sima DM, et al. A deep learning model for brain segmentation across pediatric and adult populations. *Sci Rep*. (2024) 14(1):11735. doi: 10.1038/s41598-024-61798-6

6. Tanveer M, Ganaie MA, Beheshti I, Goel T, Ahmad N, Lai KT, et al. Deep learning for brain age estimation: a systematic review. *Inform Fusion*. (2023) 96:130–43. doi: 10.1016/j.inffus.2023.03.007

7. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. (2015) 12(3):e1001779. doi: 10.1371/journal.pmed.1001779

8. Markiewicz CJ, Gorgolewski KJ, Feingold F, Blair R, Halchenko YO, Miller E, et al. The OpenNeuro resource for sharing of neuroscience data. *Elife*. (2021) 10:e71774. doi: 10.7554/eLife.71774

9. van Panhuis WG, Paul P, Emerson C, Grefenstette J, Wilder R, Herbst AJ, et al. A systematic review of barriers to data sharing in public health. *BMC Public Health*. (2014) 14(1):1144. doi: 10.1186/1471-2458-14-1144

10. McMahan B, Moore E, Ramage D, Hampson S, Arcas BAY. Communication-efficient learning of deep networks from decentralized data. In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. PMLR (2017). p. 1273–82. Available online at: https://proceedings.mlr.press/v54/mcmahan17a.html (Accessed July 4, 2023).

11. Li T, Sahu AK, Zaheer M, Sanjabi M, Talwalkar A, Smith V. Federated Optimization in Heterogeneous Networks. arXiv. (2020). Available online at: http://arxiv.org/abs/1812.06127 (Accessed February 7, 2025).

12. Basodi S, Raja R, Ray B, Gazula H, Sarwate AD, Plis S, et al. Decentralized brain age estimation using MRI data. *Neuroinform*. (2022) 20(4):981–90. doi: 10.1007/s12021-022-09570-x

13. Beutel DJ, Topal T, Mathur A, Qiu X, Fernandez-Marques J, Gao Y, et al. Flower: A Friendly Federated Learning Research Framework. arXiv; (2022). Available online at: http://arxiv.org/abs/2007.14390 (Accessed July 3, 2023).

14. Lee EH, Han M, Wright J, Kuwabara M, Mevorach J, Fu G, et al. An international study presenting a federated learning AI platform for pediatric brain tumors. *Nat Commun*. (2024) 15(1):7615. doi: 10.1038/s41467-024-51172-5

15. Pati S, Baid U, Edwards B, Sheller M, Wang SH, Reina GA, et al. Federated learning enables big data for rare cancer boundary detection. *Nat Commun*. (2022) 13(1):7346. doi: 10.1038/s41467-022-33407-5

16. Gorgolewski KJ, Auer T, Calhoun VD, Craddock RC, Das S, Duff EP, et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci Data*. (2016) 3(1):1–9. doi: 10.1038/sdata.2016.44

17. Poldrack RA, Markiewicz CJ, Appelhoff S, Ashar YK, Auer T, Baillet S, et al. The past, present, and future of the brain imaging data structure (BIDS). *Imaging Neuroscience*. (2024) 2:1–19. doi: 10.1162/imag_a_00103

18. Murman DL. The impact of age on cognition. *Semin Hear*. (2015) 36(3):111–21. doi: 10.1055/s-0035-1555115

19. Denissen S, Engemann DA, De Cock A, Costers L, Baijot J, Laton J, et al. Brain age as a surrogate marker for cognitive performance in multiple sclerosis. *Eur J Neurol*. (2022) 29(10):3039–49. doi: 10.1111/ene.15473

20. Wattjes MP, Rovira À, Miller D, Yousry TA, Sormani MP, de Stefano N, et al. Magnims consensus guidelines on the use of MRI in multiple sclerosis—establishing disease prognosis and monitoring patients. *Nat Rev Neurol*. (2015) 11(10):597–606. doi: 10.1038/nrneurol.2015.157

21. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv; (2019). Available online at: http://arxiv.org/abs/1912.01703 (Accessed July 3, 2023).

22. Team T Pandas Development. *Pandas-Dev/Pandas: Pandas*. Geneva: Zenodo (2024). Available online at: https://zenodo.org/records/13819579 (Accessed February 7, 2025).

23. Cardoso MJ, Li W, Brown R, Ma N, Kerfoot E, Wang Y, et al. MONAI: An open-source framework for deep learning in healthcare. arXiv (2022). Available online at: http://arxiv.org/abs/2211.02701 (Accessed February 7, 2025).

24. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. *J Mach Learn Res*. (2011) 12:2825–30. doi: 10.5555/1953048.2078195

25. da Costa-Luis C, Larroque SK, Altendorf K, Mary H, Sheridan R, Korobov M, et al. tqdm: A fast, Extensible Progress Bar for Python and CLI. Geneva: Zenodo (2024). Available online at: https://zenodo.org/doi/10.5281/zenodo.595120 (Accessed February 7, 2025).

26. Brett M, Markiewicz CJ, Hanke M, Côté MA, Cipollini B, Papadopoulos Orfanos D, et al. *Nipy/Nibabel: 5.3.1*. Geneva: Zenodo (2024). Available online at: https://zenodo.org/records/13936989 (Accessed February 7, 2025).

27. Zadka M. Paramiko. In. (2019). p. 111–9. doi: 10.1007/978-1-4842-4433-3

28. scp: scp module for paramiko. Available online at: https://github.com/jbardin/scp.py (Accessed February 7, 2025).

29. Hunter JD. Matplotlib: a 2D graphics environment. *Comput Sci Eng*. (2007) 9(3):90–5. doi: 10.1109/MCSE.2007.55

30. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nat Methods*. (2020) 17:261–72. doi: 10.1038/s41592-019-0686-2

31. Harris CR, Millman KJ, SJ van der W, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. *Nature*. (2020) 585(7825):357–62. doi: 10.1038/s41586-020-2649-2

32. click: Composable command line interface toolkit. Available online at: https://click.palletsprojects.com/en/stable/

33. twine: Collection of utilities for publishing packages on PyPI. Available online at: https://twine.readthedocs.io/ (Accessed February 7, 2025).

34. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely Connected Convolutional Networks. arXiv (2018). Available online at: http://arxiv.org/abs/1608.06993 (Accessed July 3, 2023).

35. Wood DA, Kafiabadi S, Busaidi AA, Guilhem E, Montvila A, Lynch J, et al. Accurate brain-age models for routine clinical MRI examinations. *Neuroimage*. (2022) 249:118871. doi: 10.1016/j.neuroimage.2022.118871

36. Roca V, Tommasi M, Andrey P, Bellet A, Schirmer MD, Henon H, et al. Federated Learning for MRI-based BrainAGE: a multicenter study on post-stroke functional outcome prediction. arXiv (2025). Available online at: http://arxiv.org/abs/2506.15626 (Accessed January 26, 2026).

37. Mateus P, Yu J, Garst SJF, Harms AGJ, Cats D, Delgado IB, et al. Federated BrainAge estimation from MRI: a proof of concept. *Alzheimer's Dement*. (2023) 19(S16):e076747. doi: 10.1002/alz.076747

38. Shafto MA, Tyler LK, Dixon M, Taylor JR, Rowe JB, Cusack R, et al. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. *BMC Neurol*. (2014) 14:204. doi: 10.1186/s12883-014-0204-1

39. IXI Dataset—Brain Development. Available online at: https://brain-development.org/ixi-dataset/ (Accessed February 21, 2025).

40. Wei D, Zhuang K, Ai L, Chen Q, Yang W, Liu W, et al. Structural and functional brain scans from the cross-sectional southwest university adult lifespan dataset. *Sci Data*. (2018) 5(1):180134. doi: 10.1038/sdata.2018.134

41. Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). *Neurology*. (1983) 33(11):1444–52. doi: 10.1212/WNL.33.11.1444

42. Benedict RH, DeLuca J, Phillips G, LaRocca N, Hudson LD, Rudick R. Validity of the symbol digit modalities test as a cognition performance outcome measure for multiple sclerosis. *Mult Scler*. (2017) 23(5):721–33. doi: 10.1177/1352458517690821

43. Van Schependom J, D'hooghe MB, Cleynhens K, D'hooge M, Haelewyck MC, De Keyser J, et al. Reduced information processing speed as primum movens for cognitive decline in MS. *Mult Scler*. (2015) 21(1):83–91. doi: 10.1177/1352458514537012

44. Isensee F, Schell M, Pflueger I, Brugnara G, Bonekamp D, Neuberger U, et al. Automated brain extraction of multisequence MRI using artificial neural networks. *Hum Brain Mapp*. (2019) 40(17):4952–64. doi: 10.1002/hbm.24750

45. Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage*. (2011) 54(3):2033–44. doi: 10.1016/j.neuroimage.2010.09.025

46. Zhong Z, Zheng M, Mai H, Zhao J, Liu X. Cancer image classification based on DenseNet model. *J Phys: Conf Ser*. (2020) 1651(1):012143. doi: 10.1088/1742-6596/1651/1/012143

47. Zhang B, Zhang S, Feng J, Zhang S. Age-level bias correction in brain age prediction. *Neuroimage Clin*. (2023) 37:103319. doi: 10.1016/j.nicl.2023.103319

48. Lou Y, Zhao L, Yu S, Sun B, Hou Z, Zhang Z, et al. Brain asymmetry differences between Chinese and Caucasian populations: a surface-based morphometric comparison study. *Brain Imaging Behav*. (2020) 14(6):2323–32. doi: 10.1007/s11682-019-00184-7

49. Tang Y, Zhao L, Lou Y, Shi Y, Fang R, Lin X, et al. Brain structure differences between Chinese and Caucasian cohorts: a comprehensive morphometry study. *Hum Brain Mapp*. (2018) 39(5):2147–55. doi: 10.1002/hbm.23994

50. Qi T, Wu F, Lyu L, Huang Y, Xie X. FedSampling: A Better Sampling Strategy for Federated Learning. arXiv (2023). Available online at: http://arxiv.org/abs/2306.14245 (Accessed March 21, 2025).

51. Peng H, Gong W, Beckmann CF, Vedaldi A, Smith SM. Accurate brain age prediction with lightweight deep neural networks. *Med Image Anal*. (2021) 68:101871. doi: 10.1016/j.media.2020.101871

52. La Rosa F, Dos Santos Silva J, Dereskewicz E, Invernizzi A, Cahan N, Galasso J, et al. BrainAgeNeXt: Advancing Brain Age Modeling for Individuals with Multiple Sclerosis. medRxiv. (2024) 2024.08.10.24311686. doi: 10.1162/imag_a_00487

53. Li T, Sanjabi M, Beirami A, Smith V. Fair Resource Allocation in Federated Learning. arXiv; (2020). Available online at: http://arxiv.org/abs/1905.10497 (Accessed August 22, 2025).

54. Iyer VN. A review on different techniques used to combat the non-IID and heterogeneous nature of data in FL. arXiv; (2024). Available online at: http://arxiv.org/abs/2401.00809 (Accessed August 22, 2025).

55. Yoo JH, Jeong H, Lee J, Chung TM. Open problems in medical federated learning. *Int J Web Inform Syst*. (2022) 18(2/3):77–99. doi: 10.1108/IJWIS-04-2022-0080

56. Matthews KRW, Yang E, Lewis SW, Vaidyanathan BR, Gorman M. International scientific collaborative activities and barriers to them in eight societies. *Account Res*. (2020) 27(8):477–95. doi: 10.1080/08989621.2020.1774373

57. Vo VTT, Shin T ho, Yang HJ, Kang SR, Kim SH. A comparison between centralized and asynchronous federated learning approaches for survival outcome prediction using clinical and PET data from non-small cell lung cancer patients. *Comput Methods Programs Biomed*. (2024) 248:108104. doi: 10.1016/j.cmpb.2024.108104

58. Borazjani K, Khosravan N, Ying L, Hosseinalipour S. Multi-Modal federated learning for cancer staging over non-IID datasets with unbalanced modalities. *IEEE Transact Med Imaging*. (2025) 44(1):556–73. doi: 10.1109/TMI.2024.3450855

59. Wen J, Zhang Z, Lan Y, Cui Z, Cai J, Zhang W. A survey on federated learning: challenges and applications. *Int J Mach Learn Cyber*. (2023) 14(2):513–35. doi: 10.1007/s13042-022-01647-y

60. Foley P, Sheller MJ, Edwards B, Pati S, Riviera W, Sharma M, et al. Openfl: the open federated learning library. *Phys Med Biol*. (2022) 67(21):214001. doi: 10.1088/1361-6560/ac97d9

61. Plis SM, Sarwate AD, Wood D, Dieringer C, Landis D, Reed C, et al. Coinstac: a privacy enabled model and prototype for leveraging and processing decentralized brain imaging data. *Front Neurosci*. (2016) 10:365. doi: 10.3389/fnins.2016.00365

62. Cong Y, Zeng Y, Qiu J, Fang Z, Zhang L, Cheng D, et al. Fedga: a greedy approach to enhance federated learning with non-IID data. *Knowl Based Syst*. (2024) 301:112201. doi: 10.1016/j.knosys.2024.112201

63. Khan W, Leem S, See KB, Wong JK, Zhang S, Fang R. A comprehensive survey of foundation models in medicine. *IEEE Rev Biomed Eng*. (2025) 19. doi: 10.1109/RBME.2025.3531360

64. Denissen S, Grothe M, Vaněčková M, Uher T, Laton J, Kudrna M, et al. Transfer learning on structural brain age models to decode cognition in MS: a federated learning approach. medRxiv (2023). p. 2023.04.22.23288741. Available online at: https://www.medrxiv.org/content/10.1101/2023.04.22.23288741v1 (Accessed September 3, 2023).

65. Dinsdale NK, Jenkinson M, Namburete AIL. Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal. *NeuroImage*. (2021) 228:117689. doi: 10.1016/j.neuroimage.2020.117689