Computers in Biology and Medicine 192 (2025) 110160

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

FedSynthCT-Brain: A federated learning framework for multi-institutional brain MRI-to-CT synthesis

Ciro Benito Raggio [a],[*], Mathias Krohmer Zabaleta [a], Nils Skupien [a], Oliver Blanck [b], Francesco Cicone [c], Giuseppe Lucio Cascini [c], Paolo Zaffino [c], Lucia Migliorelli [d], Maria Francesca Spadea [a]

[a] Institute of Biomedical Engineering, Karlsruhe Institute of Technology, Fritz-Haber-Weg 1, Karlsruhe, 76131, Baden-Württemberg, Germany
[b] Department of Radiation Oncology, University Medical Center Schleswig-Holstein, Feldstrasse 21, Kiel, 24105, Schleswig-Holstein, Germany
[c] Department of Experimental and Clinical Medicine, Magna Graecia University, Viale Europa, Catanzaro, 88100, Calabria, Italy
[d] Department of Information Engineering, Università Politecnica delle Marche, Via Brecce Bianche 12, Ancona, 60131, Marche, Italy

ARTICLE INFO

Dataset link: https://synthrad2023.grand-challenge.org/

Keywords:
Federated learning
Medical image synthesis
Synthetic computed tomography
Deep learning
Image-to-image translation
Data sharing
Data privacy

ABSTRACT

The generation of Synthetic Computed Tomography (sCT) images has become a pivotal methodology in modern clinical practice, particularly in the context of Radiotherapy (RT) treatment planning. The use of sCT enables the calculation of doses, pushing towards Magnetic Resonance Imaging (MRI) guided radiotherapy treatments. Moreover, with the introduction of MRI-Positron Emission Tomography (PET) hybrid scanners, the derivation of sCT from MRI can improve the attenuation correction of PET images.

Deep learning methods for MRI-to-sCT have shown promising results, but their reliance on single-centre training dataset limits generalisation capabilities to diverse clinical settings. Moreover, creating centralised multi-centre datasets may pose privacy concerns. To address the aforementioned issues, we introduced FedSynthCT-Brain, an approach based on the Federated Learning (FL) paradigm for MRI-to-sCT in brain imaging. This is among the first applications of FL for MRI-to-sCT, employing a cross-silo horizontal FL approach that allows multiple centres to collaboratively train a U-Net-based deep learning model. We validated our method using real multicentre data from four European and American centres, simulating heterogeneous scanner types and acquisition modalities, and tested its performance on an independent dataset from a centre outside the federation.

In the case of the unseen centre, the federated model achieved a median Mean Absolute Error (MAE) of 102.0 HU across 23 patients, with an interquartile range of 96.7–110.5 HU. The median (interquartile range) for the Structural Similarity Index (SSIM) and the Peak Signal to Noise Ratio (PNSR) were 0.89 (0.86–0.89) and 26.58 (25.52–27.42), respectively.

The analysis of the results showed acceptable performances of the federated approach, thus highlighting the potential of FL to enhance MRI-to-sCT to improve generalisability and advancing safe and equitable clinical applications while fostering collaboration and preserving data privacy.

1. Introduction

Intermodality medical Image-To-Image (I2I) translation refers to the process of mapping images from one modality (e.g., Magnetic Resonance Imaging — MRI, Computed Tomography — CT, Positron Emission Tomography — PET) to another, while preserving essential structural and anatomical information. Deep Learning (DL) based techniques have been shown to outperform traditional methods based on image registration and/or voxel-wise analysis [1] and they have gained attention in several clinical fields, the most important being the derivation of Synthetic CT (sCT) from MRI or Cone Beam CT (CBCT), to facilitate radiotherapy planning/adaptation or to improve attenuation correction in PET/MRI acquisition [2–4].

Despite the advancements in DL-based I2I, several critical limitations remain. Indeed, current DL approaches are trained on datasets that represent specific patient populations, which hampers their generalisability across different clinical settings [5–8]. For instance, the training datasets often exclude patients with implants or atypical anatomy, resulting in synthetic CTs that are less reliable for these unique cases [9]. Additionally, single-site models tend to underperform

[*] Corresponding author.
E-mail address: ciro.raggio@kit.edu (C.B. Raggio).

https://doi.org/10.1016/j.compbiomed.2025.110160
Received 12 December 2024; Received in revised form 5 March 2025; Accepted 4 April 2025
Available online 22 April 2025

when applied to images from different medical centres, as they are sensitive to domain shifts in imaging protocols, scanner types, and patient demographics [10].

To improve generalisability in MRI-to-sCT and foster innovation in the field, several studies have been conducted using cross-centre and multi-centre datasets to create models that generalise to prostate and pelvic images from different sources [11–13]. The study by Han et al. [13] used data from multiple centres also to address the impact of pre-processing to harmonise images and reduce the variability between images using normalisation, clipping and intensity correction techniques. Additionally, it investigated the impact of various combinations of data augmentation techniques for the MRI-to-sCT task focusing on deformation and rotation techniques in the case of multiple centres.

Furthermore, initiatives like the SynthRAD2023 [14] challenge have been proposed by the scientific community. These efforts allow DL models — such as the MSEP framework, the challenge winner [15,16] — to be trained on multi-centre datasets using both brain and pelvic regions, providing access to more diverse patient data. While directly pooling medical images from multiple sites could intuitively address the issue of small sample sizes, this approach faces significant barriers due to stringent privacy protection policies such as the Health Insurance Portability and Accountability Act (HIPAA)[1] in the United States and the General Data Protection Regulation (GDPR)[2] in Europe. These regulations impose limitations on the exchange of personal health data, posing practical challenging to sharing and centralising medical images across sites [17,18].

The presence of bias from limited diversity in images from a single institution, along with concerns about data centralisation, underscores the need for a collaborative approach that enables access to substantial amounts of data without centralising it [19].

Federated Learning (FL) can overcome data-sharing barriers by allowing multiple institutions to train a DL model collaboratively without sharing their own data on a central server. This approach both facilitates secure data collaboration and also leverages distributed computing resources to strengthen model generalisability across diverse, unseen datasets [20].

Given these premises, the aim of the proposed study was to establish a FL framework for MRI-to-sCT, which we call FedSynthCT, and to prove its validity in the context of brain imaging (henceforth we will refer to our approach as FedSynthCT-Brain).

A benchmark analysis was conducted across multiple architectures and different aggregation strategies in order to assess the trade-offs between efficiency and effectivenesses for realistic clinical applications.

To the best of the authors' knowledge, this is among the first applications in the literature to employ a FL paradigm in MRI-to-sCT. The main contributions and innovative aspects of our work are:

- The proposal of a cross-silo horizontal FL approach for MRI-to-sCT in brain imaging, where data of the same type (horizontal FL) from various centres (or "silos") is collaboratively used to train a deep learning model. This innovation addresses key privacy concerns by avoiding the need to pool sensitive patient data.
- The validation of the proposed approach on real multi-centre data, simulating an authentic scenario where each centre handles heterogeneous data due to differences in scanners and acquisition modalities.
- The testing of the proposed FedSynthCT-Brain on an independent dataset from a centre outside the federation to assess the federated model's generalisation capability.

The structure of this paper is as follows: Section 2 reviews existing research on computer-aided MRI-to-sCT and the use of FL for the task of medical-image analysis. Section 3 outlines the dataset and the proposed FedSynthCT-Brain approach. The experimental setup is detailed in Section 4. Section 5 presents the results, with further discussion in Section 6. Finally, Section 7 summarises the main findings and the impact of the proposed research while proposing future directions.

2. Related work

2.1. Centralised MRI-to-sCT

Over the past years, numerous DL methods have been implemented for MRI-to-sCT. The methods and models employed differed depending on the target anatomical regions and the computational resources available. Models trained on single image planes (axial, sagittal, or coronal) were termed 2D models. The combination of independently trained 2D models on different planes during or after inference led to improvements in accuracy and was referred to as 2D+ models [1,2,4].

Multi-plane models (Multi-2D) employed images from multiple planes to train a single model, leveraging inter-planar information to outperform 2D and 2D+ configurations. Consecutive slices also formed multi-channel input during 2D training, an approach known as 2.5D, which enhanced spatial coherence without requiring full 3D volumes. Smaller image segments were employed across 2D, 2D+, Multi-2D, and 2.5D methods using 2D patching [1,2,4].

More recently, the employment of entire 3D volumes or sub-volumes (also named 3D patches) as input provided the most comprehensive spatial information and mitigated inter-slice artefacts observed in 2D-based methods [1]. However, the primary limitation of 3D models lies in their higher computational demand, which requires splitting volumetric data into smaller 3D patches for processing [2]. This patch-based processing may lead to the loss of global context and structural coherence, especially when patches are analysed in isolation [21,22]. Moreover, the scalability of 3D approaches in FL settings is further constrained due to the increased communication overhead and memory requirements associated with volumetric data. These factors make 3D-DL approaches less suitable for FL real-world scenarios where limited computational resources requirements must be accounted for [23].

The proposed model architectures and methodologies have become increasingly complex in order to achieve accurate results. Among the most widely used models are: U-Nets, Generative Adversarial Networks (GANs), Transformers, and Diffusion-based architectures [1,2,4]. In addition to fundamental architectures, recent advancements in the field, such as I2I-Mamba, have demonstrated strong potential for multi-modal medical image synthesis and medical image reconstruction by leveraging state-space operators to enhance contextual interactions across different spatial orientations [24,25]. Additionally, novel diffusion-based techniques, such as SelfRDB, have been proposed to improve reliability in multi-modal medical image translation, integrating self-consistent recursive sampling and conditional guidance to enhance information transfer between modalities [26].

A further macro-classification of methodologies can be made by dividing supervised and unsupervised methods for synthetic CTs generation. Supervised techniques require accurate MRI and CT registration, whereas unsupervised techniques do not require registration, but have often demonstrated instability in convergence [27,28]. The study by Texier et al. [28] addressed and solved this problem related to unsupervised learning, training a patch-based 3D cGAN using the "Content and Style Representation for Enhanced Perceptual synthesis" (CREPs) loss function and multi-centre prostate data.

Several works employed three different 2-dimensional (2D) fully Convolutional Neural Networks (CNN), to retain structural information in sCTs while simultaneously reducing computational costs, especially when using U-Nets [1,29–31]. These 2D methods, by processing slices independently or in parallel, inherently simplify training in distributed environments, such as FL, where data sharing is constrained by computational heterogeneity [23].

[1] https://www.hhs.gov/hipaa/index.html.

[2] https://gdpr-info.eu/.

2.2. Federated learning in medical imaging

The growing number of studies applying FL in healthcare, particularly in medical imaging [18,32–35], reflects its increasing importance. By keeping sensitive information, such as patient data, within local environments (e.g., clinical centres), FL addresses the crucial need for confidentiality in medical applications, enabling collaborative large-scale model development without data sharing.

In the medical context, the focus is on cross-silo FL [36], characterised by a limited number of participating entities (often ranging from 2 to 100), where each entity is recognised as a "silo" of data. During a single round –a single communication cycle between the aggregation server and a set of clients (or silos)–, clients perform local training of their models using private data on the device itself, and then share updates with the server for aggregation.

The partitioning of data is of pivotal importance in this context, particularly in the domain of medical imaging, where datasets are frequently scarce, heterogeneous, and non-Independent and Identically Distributed (non-IID). Data imbalance has the potential to significantly impact model generalisation and performance across institutions, as medical images from disparate sources are susceptible to bias due to differences in equipment, protocols, and labelling processes. FL addresses these challenges by aggregating models from various institutions, thereby mitigating bias and improving generalisability [37].

Several studies investigated the use of FL on medical images, especially for classification and segmentation tasks.

In the field of classification, FL has been applied X-ray and CT images for the diagnosis of COVID-19 [38,39]. Furthermore, numerous studies have focused also on the classification of skin diseases and lesions [40,41], prostate classification [40,42], as well as in the ophthalmology field for the diabetic retinopathy classification [43].

More recently, FL approaches were proposed for segmentation tasks, such as for brain tumour [44,45], prostate [46], optic cup and disc segmentation [47,48]. This led to the development of frameworks that integrate the aforementioned works and are capable of performing segmentations of multiple anatomical structures or different image modalities using FL and foundational models, such as the FedFMS [49], which was created to integrate the Segment Anything Model (SAM) [50] into the medical FL context.

Recent advancements in medical image synthesis have introduced innovative methods, such as FedMed-GAN, which specifically addresses the challenge of synthesising cross-modality brain MRI images in an FL setting [51]. It employs Federated Averaging (FedAvg), the *de facto* aggregation strategy algorithm for FL [36], enabling the combination of locally trained models from multiple institutions by computing a weighted average of their parameters.

Inspired by the FL research, this study investigates the feasibility of FL for MRI-to-CT synthesis, addressing a critical gap in the literature. Indeed, to the best of the authors' knowledge, this is among the first work to apply a cross-silo horizontal FL approach in this context, leveraging multi-centre real heterogeneous datasets.

3. Methods

3.1. Datasets and preprocessing

For each centre (named A, B, C, D and E) a dataset containing paired, co-registered MRI T1-weighted and CT images was available. Image acquisition was performed in accordance with the ethical standards of the 1964 Declaration of Helsinki and later amendments. Written informed consent to use data for research purposes was obtained from all patients.

The image acquisition protocol and scanners differed among the various centres, as it is reported in Table 1. This led to image variability among the federated dataset and the most relevant difference can be appreciated in Fig. 1.

The field of view (FOV) of images from Centre A started at the mouth region with a straight cut off and spans towards the top of the skull. A fraction of the ears was also cut off.

MRIs from Centre B were acquired using a PET-MRI hybrid scanner. Differently from the other centres, this acquisitions were not performed for radiotherapy planning purposes. Thus, the tube energy in the CT was 120 kV and the slice thickness was 3.75 mm which introduced partial volume artefacts in the upper part of the skull bone. In light of the substantial difference in the acquisition protocol and scanners, Centre B dataset was used to increase the image variability within the federation (i.e. to introduce noise) to test the robustness of the federated model and simulate a real context.

Images belonging to Centre C presented higher spatial resolution in the cranial caudal direction compared to the other datasets. However, the Centre C's dataset presented a number of additional challenges, including: the absence of the top of the skull in some patients, the presence of the scanner bed and pillow for each CT, metal artefacts, and high bias field artefact.

The Centre D's and Centre E's datasets derived exclusively from two different institutions of the SynthRAD2023 multi-centre dataset (SynthRAD Centre C and SynthRAD Centre B respectively), which was created for the SynthRAD challenge within different MRI-to-sCT algorithms [14]. Before publication, the datasets were cropped to remove the facial region. The FOV covered by both imaging modalities was similar, except for a cut-off at the lower end with a slight angle in addition to a straight horizontal line.

In general, Centre D and E presented less bias field artefact when compared to Centre C. However, between Centre D and E, it was more pronounced in the Centre E dataset. Furthermore, both Centre MRIs were obtained using gadolinium-based contrast agents, thus introducing further heterogeneity within the federation.

To address the image variability, a pre-processing strategy was implemented at each client site with the aim of harmonising data within the federation, without the exchange of data. Following [13], N4 Bias Field Correction [52] was applied to the MRI data to correct intensity inhomogeneity, which can result from factors such as magnetic field variations or patient positioning. Additionally, Min–Max Normalization was applied within each client to reduce variability in pixel intensity values.

Due to the differences in acquisition protocols, the datasets included images of varying dimensions (See Table 1). To standardise input for the federated model, all image volumes were cropped, resized, and padded to a uniform size of 256 × 256 × 256 voxels (See Section 4). This process enabled each client to preprocess its data independently while maintaining a consistent input format across the federation, thereby allowing seamless aggregation of the DL model updated from each client.

The elements outside the patient body (couch, immobilisation system, pillow) were masked out on the CT volume using 3D Slicer image computing platform [53](Fig. 2). The obtained mask was applied also to the corresponding MRI to remove background noise. Voxels outside the mask were set to −1000 and 0 for CT and MRI respectively.

3.2. Federation setup

As shown in Fig. 3, we created a FL environment to simulate a realistic collaboration involving 4 distinct clients, each representing a different clinical centre as outlined. In this setup, each client trained a local DL model on its own data, while a central server coordinated the creation of a federated model. This federated model was built by aggregating key parameters (e.g., weights, biases, and batch normalisation parameters) from each local model after every local training round, effectively combining weights from all clients without sharing raw data. The aggregated model parameters were then redistributed to each client, enhancing their local models with information learned across all sites.

To evaluate generalisation, the federated model was tested on an independent dataset from a fifth centre, which none of the clients had access to during training.

Table 1
Summary of the most relevant characteristics of the image datasets used in the federated learning (FL) framework. Data from Centre A, B, C, and D were used to train the federated model; thus Centre A, B, C, D acted as clients. The data from the Centre E (SynthRAD Centre B) served as an independent dataset for testing the generalisability of the federated model on unseen data. For each dataset we list the number of patients, image sizes, magnetic resonance imaging (MRI) and computed tomography (CT) scanners used, and voxel spacing in millimetres.

Parameter		Centre A	Centre B	Centre C	Centre D	Centre E
Patients #		15	14	21	29	23
Voxel size [mm^3]	MR	1,1,1	1,1,1	0.78,0.78,1	0.98–1.12, 0.98–1.12, 0.98–1.12	0.98,0.98,0.98
	CT	0.49–0.67, 0.49–0.67, 2.5	0.98, 0.98, 3.27	0.78, 0.78, 1	0.69–0.78, 0.69–0.79, 1–3	0.59–1.27, 0.59–1.27, 1–2
Scanners	MR [T]	MAGNETOM Trio [3]	Biograph mMR [3]	Vantage Titan [1.5]	MAGNETOM Avanto_fit, Skyra, Vida_fit, Prisma_fit [1.5–3]	MAGNETOM Aera, Avanto_fit [1.5–3]
	CT [kVp]	LightSpeed QX/i [140]	Discovery ST [120]	Brilliance Sensation Open [120]	Brilliance Big Bore [120]	SOMATOM Definition AS [120]
Size	Axial	256	256	226–357	167–213	167–262
	Sagittal	176	176	512	216–262	200–225
	Coronal	256	248	512	250–277	225–248

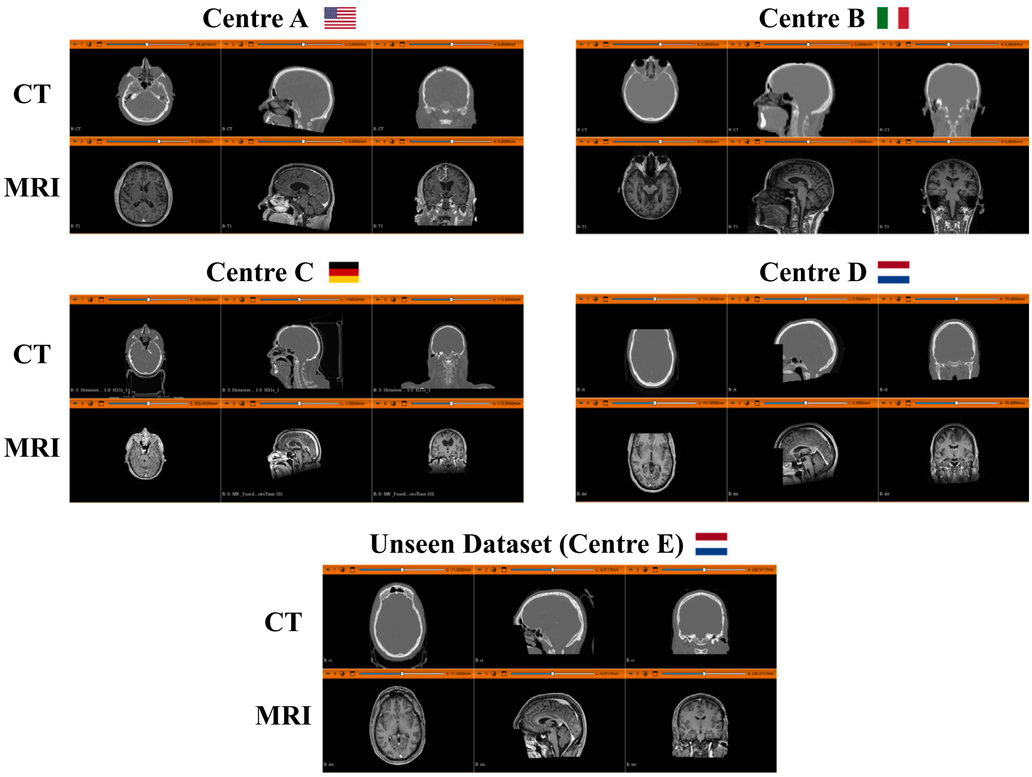

Fig. 1. Visual comparison of image appearance across datasets from multiple Centres. The FOV of the images from Centre A was limited to the level of the mouth and included partial cut-off of the ears. Images from centre B were acquired using a PET-MRI hybrid scanner. Furthermore the higher CT slice thickness introduced partial volume artefacts in the skull bone reconstruction. Images coming from centre C exhibited higher spatial resolution in the cranial caudal direction, but presented other problems, including partial cropping at the top of the skull and metal artefacts. Centres D and E–derived from SynthRAD2023 challenge–shared similar imaging regions with less bias fields compared to Centre C.

3.3. Deep learning model

Following preliminary experiments (see Section 4.4.1), the DL architecture employed in the proposed FL framework was the U-Net model proposed by Li et al. [54]. The network is a residual learning based U-Net, designed to predict 256 × 256 sCT 2D slices from 256 × 256 MR inputs. The architecture comprises 34 convolutional layers, starting with a 7 × 7 convolutional layer for initial feature extraction, followed by multiple 3 × 3 convolutional layers.

A key feature of the architecture dealt with the inclusion of residual blocks, which consisted of two 3 × 3 convolutional layers, each followed by batch normalisation and rectified linear unit (ReLU) activation, and a short skip connection that added the input of the block to its output to improve gradient propagation and enable efficient training. In

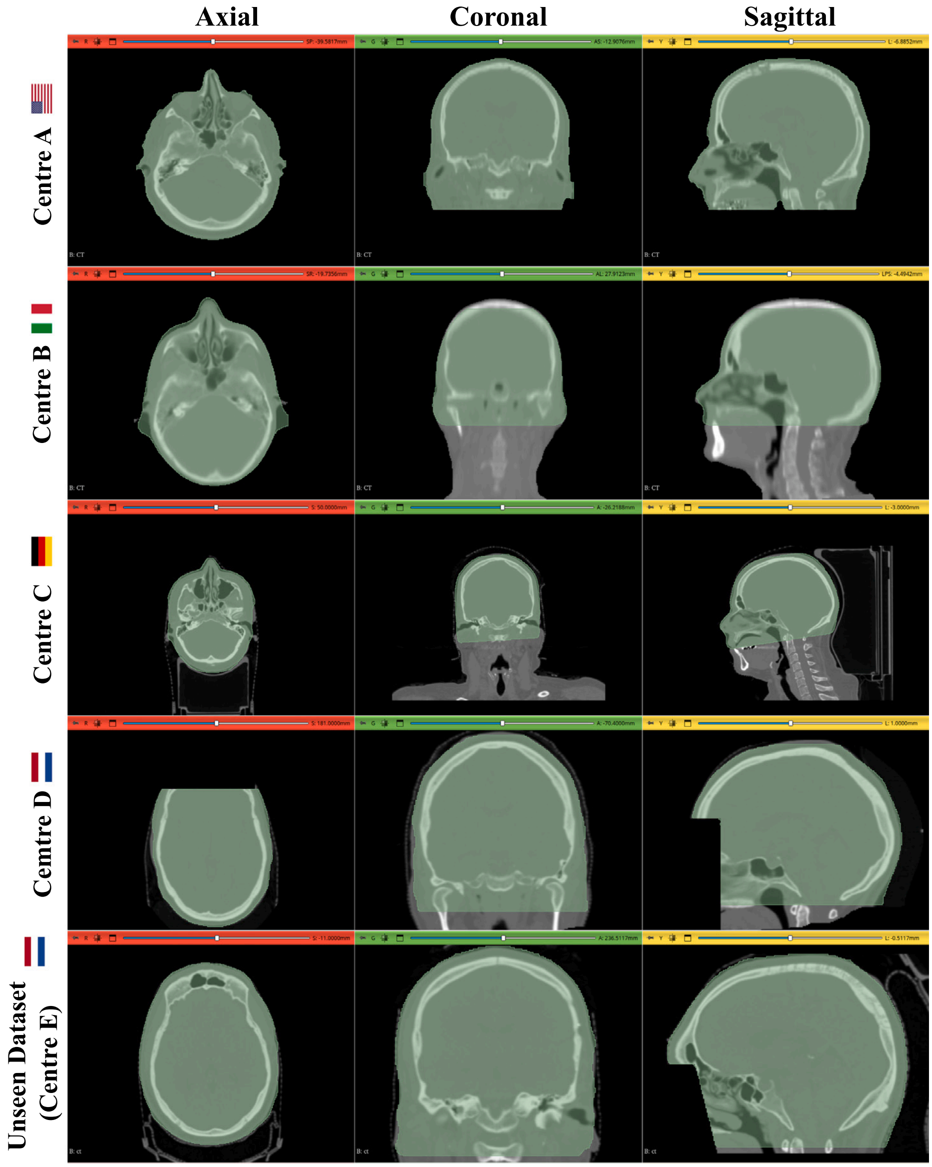

Fig. 2. The figure illustrates the background masking for each institution. By eliminating potential misleading features for the model, the masking approach limits the prediction of the model to areas of interest and facilitates the harmonisation of the model's knowledge within the federation. The masking process was conducted within the individual Centre. The parts of interest were established beforehand, as would be done in a real-world context.

the encoder path, max-pooling operations were used for downsampling the dimensions of the feature maps, doubling the number of filters at each step. In the decoder path, upsampling operations were performed before convolutional layers to restore feature maps to their original size, halving the number of filters at each step. Skip connections between the encoder and decoder paths facilitated the reuse of spatial information.

To train the aforementioned architecture, each client input 2D slices from each MRI in 3 anatomical planes, axial, coronal, and sagittal, as it is shown in Fig. 4. Slices were presented to the network in a randomised order. After extracting all slices from each anatomical plane, they were shuffled to ensure that the DL model received non-sequential input from more than one anatomical plane at a time. This minimised the risk of the model becoming overly dependent on any single plane or orientation, encouraging comprehensive and plane-invariant feature learning across all planes [55]. This method was adapted from the Multi-2D approach [2], modified to enhance generalisation. As a result, it was renamed Random Multi-2D.

3.4. Voting

As shown in Fig. 5, in order to reconstruct the sCT, the predictions generated by the three anatomical planes were combined using a median voting approach to obtain the final voxel value, as proposed in [30]. This post-processing procedure ensured a more uniform and reliable output and was applied on both the server-side and the client-side predictions before the evaluation process. Specifically, once the predictions for each slice had been generated for each anatomical plane, the median result across the three axes was used to generate the final sCT voxel according to the formula [30]:

$$sHU_{V_n} = \text{median}(sHU_{\text{Cor}_n}; sHU_{\text{Ax}_n}; sHU_{\text{Sag}_n}), \quad n \in \{1, \dots, N\} \tag{1}$$

where V_n (with $n \in \{1, \dots, N\}$) refers to the pixel number, sHU refers to the synthetic Hounsfield Unit value of the n pixel for the coronal, axial, and sagittal sCT.

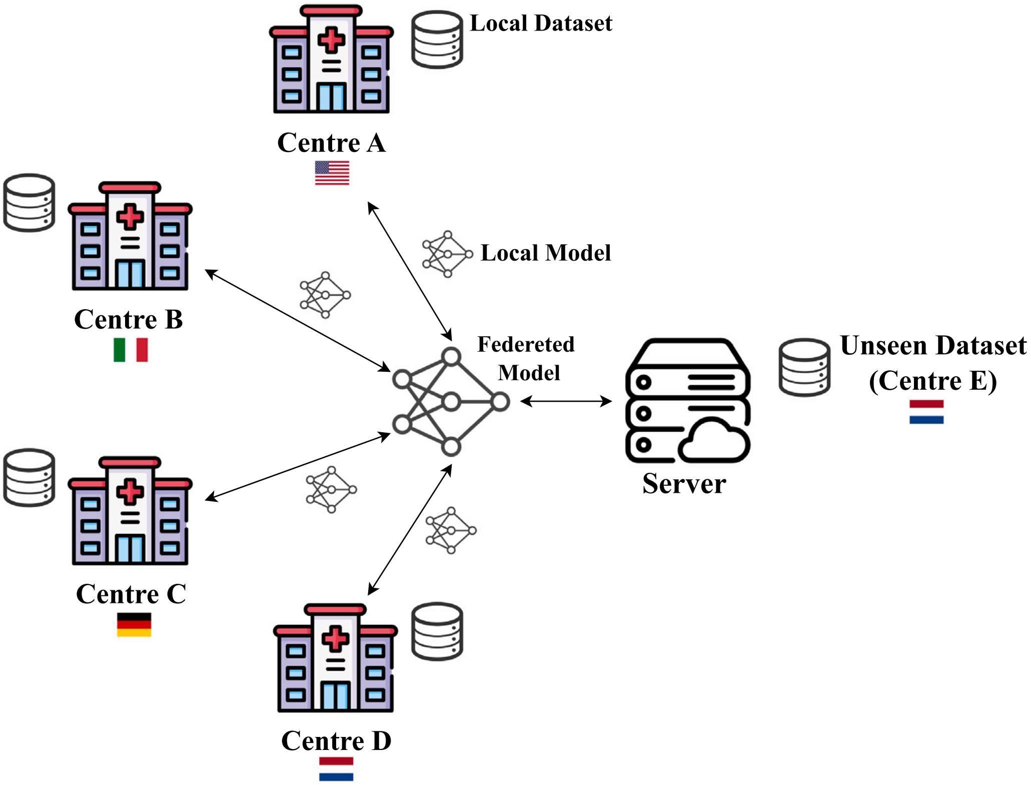

Fig. 3. Federated setup. Each centre (also called client) has its own local data and trains a model on the local dataset. The weights of the local model are exchanged with the server, which, after the aggregation process, returns a federated model to the clients. The server also hosts a dataset that is not used during training, but acts as a benchmark to determine the generalisation ability of the model.

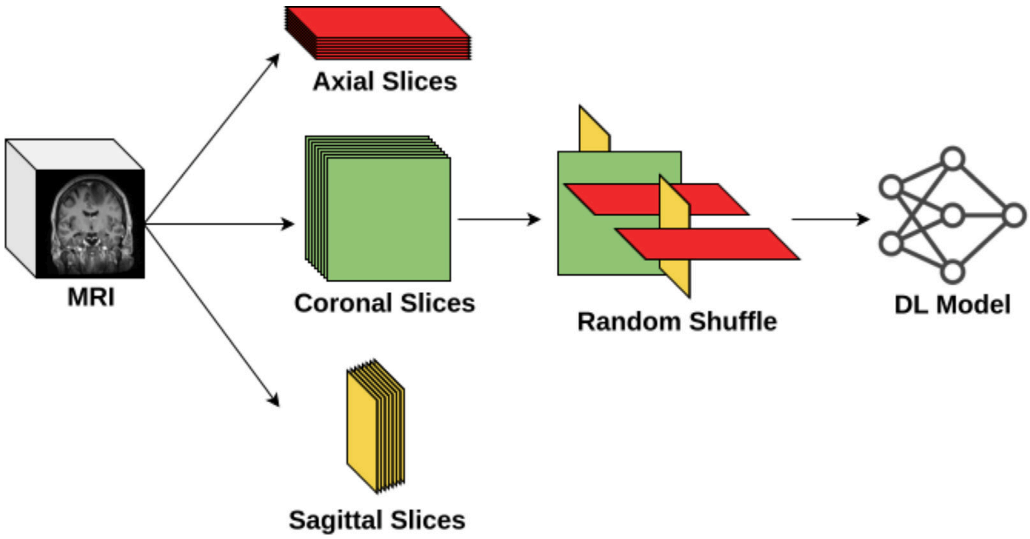

Fig. 4. Proposed training methodology. The pre-processed and augmented 3D (MRI) volume is subdivided into 2D slices, extracted along the three anatomical planes (sagittal, axial, coronal). These slices are then subjected to a random shuffling process, ensuring that the model is agnostic with respect to both the anatomical plane and the imaging sequence. The shuffled batch of images is then provided as input to the model.

3.5. *Aggregation strategy*

After testing different aggregation strategies (see Section 4.4.2), FedAvg in combination with FedProx to address the challenges associated with non-IID data across clients [56], was adopted for FedSynthCT-Brain.

More in detail, the weights from each client are aggregated round-by-round using a weighted average (thus, FedAvg), with each client's contribution proportional to the size of its dataset.

$$w_{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} w_{t+1}^{k} \tag{2}$$

where K represents the number of clients, n_k denotes the dataset size for client k, n is the total dataset size across all clients, and w_{t+1}^{k} are the weights of client k post-training.

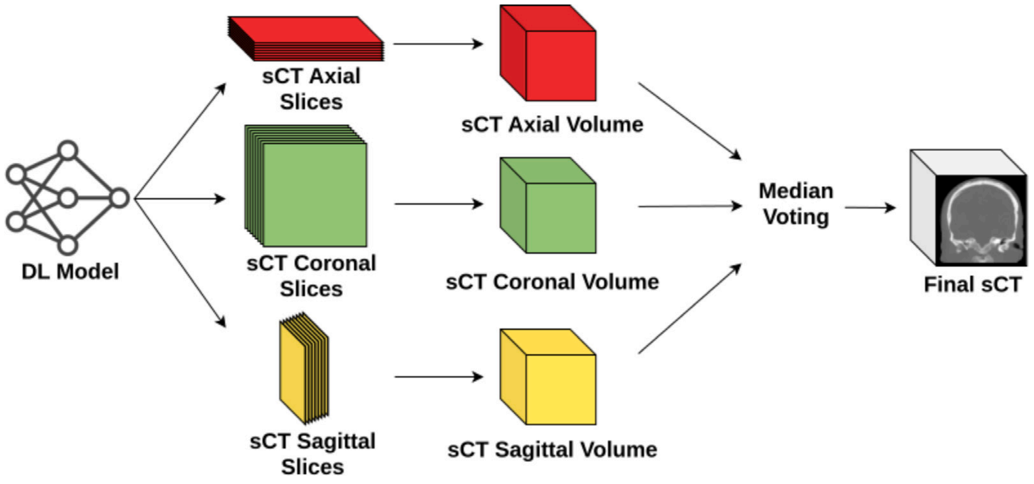

Fig. 5. Reconstruction of the sCT. The axial, coronal, and sagittal slice predictions are rearranged to construct three distinct volumes. Subsequently, a median voting approach is adopted using the three volumes in order to enhance the final result.

As FedAvg underperform with non-IID data [57], FedProx was incorporated to address this limitation. The FedProx method, specifically designed for heterogeneous federated environment, introduces a proximal term $\frac{\mu}{2}\|w - w_t\|^2$ to the local objective function, penalising large deviations from the global model weights to control local update variations:

$$h_k(w, w_t) = F_k(w) + \frac{\mu}{2}\|w - w_t\|^2 \tag{3}$$

where $F_k(w)$ is the original local objective, w_t is the pre-training global model weight, w represents weights during local training, and μ is the proximal coefficient, within the range $[0, 1]$. Therefore, FedProx reduces update divergence and approximates FedAvg when $\mu = 0$ [56].

4. Experimental protocol

4.1. Data preprocessing details

The pre-processing pipeline used for each dataset can be divided into different steps, as shown in pseudo-code (Algorithm 1). To correct the bias field, the N4ITK algorithm [52] was employed. The algorithm was applied to each MRI with identical parameters.

Algorithm 1 Pre-processing pipeline for each Centre in pseudo-code

```
if Volume is MRI then
    Apply N4 Bias Field Correction
end if
if Volume Size > $256 \times 256 \times 256$ then
    Apply Crop to centre the volume
    Resize to $256 \times 256 \times 256$
end if
if Volume Size < $256 \times 256 \times 256$ then
    Pad to $256 \times 256 \times 256$
end if
if Volume is MRI then
    Apply Min-Max Normalization
end if
```

The second step was the standardisation of the anatomical axes orientation, which facilitates the application of the subsequent transforms and the extraction of slices. Then, all volumes that had a larger dimension than the target dimension were centred cropped to a size of $328 \times 256 \times 328$. Finally, all volumes were reshaped to a voxel size of $1\ \text{mm} \times 1\ \text{mm} \times 1\ \text{mm}$.

Furthermore, the resizing and padding process was executed by establishing a maximum dimension of 256 for all directions. The padding values were −1000 HU and 0 for CTs and MRIs, respectively.

Finally, the intensity values of each MRI were normalised to the $[0, 1]$ range using the Min–Max Normalisation.

All experiments proposed in the subsequent sections were conducted with 4 clients, where each client used 2 patients for validation, 2 patients for testing, and the remaining for training.

4.2. Implementation and training settings

The implementation of the proposed work, was carried out in Python, using the Flower framework [58] for the FL infrastructure and the PyTorch [59] library for the DL implementation. To reproduce the federated environment and train the models, we used an NVIDIA A100 80 GB GPU, 16 CPU cores, and 64 GB of RAM from a Supermicro AS-4124GS-TNR 4U Rackmount GPU SuperServer.

To maintain consistency and prevent overfitting, a custom-built data augmentation pipeline was implemented for each client. The AugmentedDataLoader library[3] was used to perform optimised on-the-fly data augmentation during training. In order to address the considerable computational demands associated with this data augmentation approach, two distinct data augmentation pipelines were employed: (i) a minimal pipeline, comprising a flip and a fixed-angle single-axis rotation; (ii) an extended pipeline, comprising a flip, a random-angle three-axis rotation and a fixed spatial translation; based on the amount of data available for the individual site.

A preliminary study was conducted to establish the optimal number of local epochs each client, thereby ensuring that the global model converged efficiently, rather than focusing exclusively on local optimisation. This involved an investigation into the impact of varying the number of local clients training epochs on the overall performance of the global model.

As reported in [56], it is beneficial to minimise the size of local updates in the context of FL. With this view, experiments were conducted using $1, 2, 3$, and 5 epochs per round. It has been experimentally observed that an increasing number of epochs per round resulted in a proportional increase in computation time, with a degradation in image similarity metrics for the federated model. Consequently, the optimal number of local epochs was experimentally set to 1.

Each client model was trained using the Adam optimiser and the L1 loss. After an initial experimental analysis, the learning rate (LR) was set to 10^{-4}. The batch size was set to 32.

[3] https://github.com/ciroraggio/AugmentedDataLoader.

Each experiment ran for 30 rounds. This setup allowed for a rigorous and fair evaluation of both the model's ability to generalise across rounds and its performance on unseen data.

4.3. Performance metrics

In order to evaluate the efficacy of the federated model both on the clients test datasets (centres A, B, C, and D), and unseen test dataset (centre E) standard image similarity metrics in the MRI to CT translation task were employed [2,4]. These included:

- Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|CT_i - sCT_i\right|$$

where n is the total number of voxels in the region of interest.

- Structural Similarity Index Measure (SSIM) [2]:

$$\text{SSIM} = \frac{(2\mu_{\text{sCT}}\mu_{\text{CT}} + C_1)(2\sigma_{\text{sCT,CT}} + C_2)}{(\mu_{\text{sCT}}^2 + \mu_{\text{CT}}^2 + C_1)(\sigma_{\text{sCT}}^2 + \sigma_{\text{CT}}^2 + C_2)}$$

$$C_1 = (k_1 L)^2, \quad C_2 = (k_2 L)^2$$

$$\mu = \text{Mean}, \quad \sigma = \text{Variance/Covariance}$$

$$L = \text{Dynamic range}, \quad k_1 = 0.01, \quad k_2 = 0.03$$

- Peak Signal-to-Noise Ratio (PSNR) [2]:

$$\text{PSNR} = 10 \cdot \log\left(\frac{MAX_{CT}^2}{\text{MSE}}\right)$$

We further evaluated the efficiency using the time elapsed to complete a fixed number of rounds. This was aimed at identifying the most suitable architecture for the FL framework, balancing generalisation capability and training time. Indeed, minimising training time is crucial, as extended training may exacerbate resource constraints, increase communication overhead among clients, and compromise the scalability of the system, particularly in multi-institutional collaborations [60].

4.4. Deep learning architecture and aggregation strategy selection

4.4.1. Deep learning architecture selection

We explored different DL architectures used for MRI-to-sCT conversion, including Conditional GAN, CycleGAN and Transformers (see the supplementary Appendix A for further details). Among these, we further investigated several variants of the U-Net architecture which has been used for centralised MRI-to-sCT-Simple U-Net [61], Spadea et al. U-Net [30], Fu et al. U-Net [62], Li et al. U-Nets [63]. These variants included variations in implementations, such as changes in network depth, feature map sizes, and the integration of attention mechanisms. Our analysis focused on evaluating the balance between computational complexity and generalisation capability across these architectures.

An additional investigation is presented in Appendix B which involved a comparison of the adopted training methodology (Random Multi-2D) with other methodologies (such as 2D+, 2D patches and classic Multi-2D) using the U-Net presented in Section 3.3.

4.4.2. Aggregation strategy selection

Once the most appropriate architecture was identified, a comparative analysis was conducted between the chosen aggregation strategy (i.e., FedAvg+FedProx), simple aggregation methods, such as FedAvg, and optimised and more complex techniques, such as FedYogi and FedAvgM with the incorporation of the momentum coefficient and the server-side learning rate.

Further optimisation was pursued through the incorporation of additional techniques, including FedBN [64], and the combination of FedProx and FedBN, with the aim of refining and optimising the federated training process. In our study, the FedBN method was specifically designed by the exclusion of batch normalisation layers from communication between the server and client. This was done with a particular focus on ensuring that the global model could converge more rapidly and improve the robustness.

To determine the range of variability of the convergence trend within 25 rounds, the experiments were repeated 5 times for each strategy.

Regarding the hyper parameters used for each strategy, FedAvgM required the tuning of the momentum coefficient β and the server-side learning rate η_s. It was noted experimentally that the use of $\beta > 0.4$ and $\eta_s < 0.1$ leads to instability due to the inability of the model to reach the convergence. In light of this, the parameters used for FedAvgM were $\beta = 0.3$ and $\eta_s = 0.2$, respectively.

The FedYogi strategy necessitates the configuration and optimisation of different hyper parameters. These have been empirically optimised following a series of experiments with the objective of stabilising convergence. The best result for FedYogi was achieved with the following parameters: $\eta = 0.03$, $\eta_l = 10^{-4}$, $\beta_1 = 0.6$, $\beta_2 = 0.6$, and $\tau = 0.01$.

To add FedProx, different values of the proximal coefficient μ were tested. Higher μ corresponds to a greater penalty for significant deviations from the global model during local training, which may impede the model's convergence. On the other hand, setting $\mu = 0$ would be equivalent to FedAvg. The optimal results for FedProx were obtained with a $\mu = 3$.

The FedBN approach was also tested with FedAvg and combined to FedAvg and FedProx. While no parameter tuning was required for FedBN, the conventional approach was modified.

Batch normalisation layers were excluded solely when transmitting the global model to clients, thereby enabling the aggregation of all layers, including batch normalisation, at the server site. This alteration guarantees that the global model can leverage client-specific batch normalisation updates, enhancing its resilience when evaluated with previously unseen data.

5. Results

5.1. Benchmarking deep learning architectures

The performances of different state-of-the-art U-Nets, using a FL setup and the chosen aggregation strategy (FedAvg+FedProx), are presented in Table 2, in terms of median and inter-quartile range values of MAE, SSIM, and PSNR. The results are provided the test dataset of each centre participating to the federation (A, B, C, and D) and for the external centre E.

Image similarity assessment did not reveal any relevant difference across the tested U-Net architectures. However, when examining the total time required to complete 30 rounds, the U-Net architecture proposed by [54] required 12 h and 34 min.

Additional results obtained with different architectures, which proved to be less suitable than U-Nets for the proposed federated task, are presented in Table A.4. Nevertheless, the results obtained from the supplementary investigation on the performance obtained of the U-Net by Li et al. [54] when using different DL training methodologies are reported in Table B.5.

5.2. Benchmarking aggregation strategies

Table 3 presents the best result obtained for each aggregation strategy, tested on the unseen dataset (centre E). Results are reported in terms of MAE, which resulted the most sensitive image similarity metric in the benchmarking of DL architectures. The experiments (Experiment 1, 2, 3, 4, 5), were repeated 5 times for each strategy to determine the range of variability of the convergence trend within 25 rounds.

Table 2
Comparison of the results obtained with different state-of-the-art U-Nets on the unseen centre dataset and federated centres (A, B, C, D) test data based on image similarity metrics and training time, considered as the total time to complete a fixed number of rounds.

	Method	MAE [HU]	SSIM	PSNR [dB]	Total time
Centre A	Simple U-Net [61]	85.6 (84.8–86.4)	0.96 (0.96–0.96)	33.91 (33.65–34.18)	–
	Spadea et al.[30]	88.6 (88.0–89.3)	0.95 (0.95-0.95)	33.72 (33.45–33.99)	–
	Fu et al. [62]	91.3 (90.7–92.0)	0.95 (0.95–0.95)	33.43 (33.15–33.71)	–
	Li et al. (2020) [65]	93.0 (92.6–93.4)	0.95 (0.95–0.95)	33.34 (33.07–33.61)	–
	Li et al. (2019) [54]	89.7 (89.4–90.1)	0.95 (0.95–0.95)	33.64 (33.46–33.81)	–
Centre B	Simple U-Net [61]	126.1 (125.6–126.6)	0.86 (0.86–0.86)	21.83 (21.06–22.61)	–
	Spadea et al.[30]	128.6 (128.1–129.0)	0.86 (0.85–0.86)	21.81 (21.04–22.59)	–
	Fu et al. [62]	125.7 (124.8–126.7)	0.86 (0.86–0.86)	21.82 (21.04–22.61)	–
	Li et al. (2020) [65]	131.0 (130.5–131.5)	0.86 (0.85–0.86)	21.77 (21.01–22.54)	–
	Li et al. (2019) [54]	124.7 (124.4–125.1)	0.86 (0.86–0.86)	21.85 (21.08–22.63)	–
Centre C	Simple U-Net [61]	97.0 (94.1–99.9)	0.74 (0.71–0.77)	20.03 (19.36–20.71)	–
	Spadea et al.[30]	95.3 (92.0–98.6)	0.74 (0.71–0.77)	20.06 (19.37–20.73)	–
	Fu et al. [62]	89.1 (83.8–94.3)	0.74 (0.72–0.77)	20.05 (19.38–20.75)	–
	Li et al. (2020) [65]	97.1 (95.6–98.6)	0.74 (0.72–0.77)	20.03 (19.36–20.70)	–
	Li et al. (2019) [54]	92.9 (89.8–96.0)	0.74 (0.72–0.77)	20.05 (19.37–20.73)	–
Centre D	Simple U-Net [61]	70.8 (66.9–74.7)	0.93 (0.93–0.93)	27.05 (27.03–27.08)	–
	Spadea et al.[30]	71.7 (68.8–74.7)	0.93 (0.93–0.93)	27.05 (27.03–27.06)	–
	Fu et al. [62]	71.7 (68.4–75.0)	0.93 (0.93–0.93)	27.03 (27.01–27.05)	–
	Li et al. (2020) [65]	71.6 (66.5–76.7)	0.93 (0.93–0.93)	27.04 (27.00–27.07)	–
	Li et al. (2019) [54]	72.6 (63.4–76.8)	0.93 (0.93–0.93)	27.04 (27.02–27.06)	–
Centre E (Unseen)	Simple U-Net [61]	100.3 (94.5–108.2)	0.89 (0.86–0.89)	26.56 (25.53–27.52)	16 h 22 min
	Spadea et al.[30]	99.8 (92.6–108.9)	0.89 (0.86–0.89)	26.56 (25.53-27.52)	22 h 41 min
	Fu et al. [62]	99.8 (94.9–106.1)	0.89 (0.86–0.89)	26.61 (25.53–27.49)	20 h 22 min
	Li et al. (2020) [65]	98.9 (93.3–108.0)	0.89 (0.86–0.89)	26.55 (25.57–27.45)	15 h 57 min
	Li et al. (2019) [54]	102.0 (96.7–110.5)	0.89 (0.86–0.89)	26.58 (25.52–27.42)	**12 h 34 min**

Table 3
Comparison, in terms of image similarity (MAE) and number of rounds to achieve the best results among different aggregation strategies. Experiments were repeated 5 times and the mean and standard deviation (Std) are presented for both number of rounds and the MAE.

Strategy	Metric	Experiment					Mean±Std
		1	2	3	4	5	
FedAvg	Round	10	16	20	10	18	14.8 ± 4.1
	MAE [HU]	99.5	99.3	100.8	102.6	101.1	100.7 ± 1.4
FedAvgM	Round	28	27	28	29	27	27.8 ± 0.7
	MAE [HU]	102	101.8	103	101	101.4	101.8 ± 0.7
FedYogi	Round	13	13	17	4	12	11.8 ± 4.3
	MAE [HU]	207.7	230.6	225.9	236.1	234.4	226.9 ± 10.2
FedAvg + FedBN	Round	20	6	15	21	21	16.6 ± 5.7
	MAE [HU]	101	100.9	99.6	100.7	98.3	100.1 ± 1.1
FedAvg + FedProx + FedBN	Round	15	8	15	21	19	15.6 ± 4.4
	MAE [HU]	99.2	99.9	97.5	100.1	100.1	99.36 ± 1
FedAvg + FedProx (Our)	Round	18	8	14	14	9	**12.6 ± 3.7**
	MAE [HU]	99.8	99.2	100.1	101.6	98.3	99.8 ± 1.1

This comparison helps to understand which aggregation strategy gives the best result in a stable and efficient way. For each experiment, the number of rounds performed to obtain the best result is also reported. Fig. D.10 reports the convergence plots for all evaluated strategies. Furthermore, a visual comparison of the results obtained with different strategies were reported in Fig. D.11.

5.3. Federated sCTs

As reported in Table 2, using the U-Net by Li et al. [54] and FedAvg combined with FedProx as aggregation strategy, the median MAE (inter-quartile range) for the unseen dataset was 102.0 HU (96.7–110.5 HU).

Fig. 6 shows results for the best and worst cases of Centre E.

A view of the central slice in the three anatomical places is provided, demonstrating the advantages and limitations of the framework. Furthermore, the difference between the ground truth CT and the sCT is shown to assess critical pattern in the sCT generation.

The results obtained by the federated model on one random patient of each client test dataset are also presented for each centre in Figs. 7 and 8. The central slice of the three anatomical planes is provided for the clients' data, along with the ground truth CT and sCT difference. The figures include rectangles of different colours to identify errors or artefacts observed in the images during the comparative analysis, that are not evident when considering only the similarity metrics.

6. Discussion

This study addressed the challenges of bias and limited diversity in medical imaging datasets caused by single-institution reliance in the context of MRI-to-sCT. Our proposal was to introduce a FL framework designed for MRI-to-CT synthesis in brain imaging, FedSyntCT-Brain. We reproduced a realistic FL environment, adhering to all FL principles, using a cross-silo horizontal FL paradigm to enable the collaborative training of DL models as if multiple institutions were involved. To further validate the effectiveness of our approach, we tested the performance of the federated model on an unseen dataset, demonstrating its generalisability to external heterogeneous data sources.

Regarding the evaluation of U-Net architectures, the results obtained on Centre E — the unseen dataset — as well as across the various Centres in the federation, reveal no relevant differences in image similarity metrics when comparing different architectures. Based

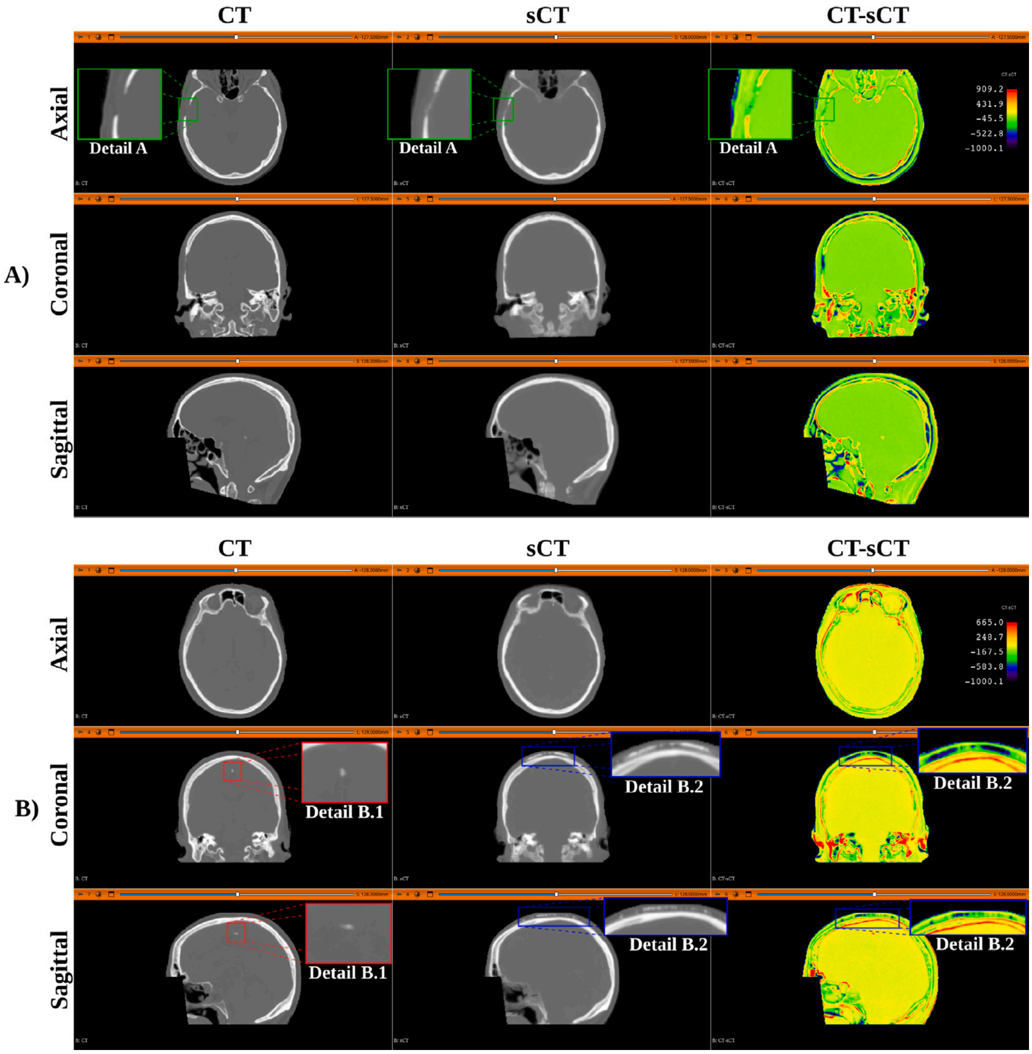

Fig. 6. The figure shows the central slice for the axial, coronal and sagittal planes, which is used for an evaluation of the worst and best – A and B respectively – predictions based on image similarity metrics of the federated model on the unseen dataset (Centre E). The Detail A (in green), shows the ability of the model to predict a missing portion of bone, although the case A was from the unseen Centre. In the case B, the Detail B.1 (in red) highlights a calcification in the ground truth CT, which is not correctly predicted by the model. In the case B, the Detail B.2 (in blue) also highlights a thin duplication of the upper skull bone. This issue has been identified exclusively within the predictions generated by the federated model on the Centre E dataset. The colour map on the left illustrates the differences between the CT and the sCT, highlighting regions where the model exhibits a higher or lower tendency to over or underestimate the actual outcome.

on the MAE metric, the minimal variation in HU does not allow for the identification of a clearly superior U-Net design. This could imply that, within the context of the FedSynthCT-Brain framework, the architectures are sufficiently robust to handle the data variability across the different centres. Furthermore, SSIM and PSNR, being less sensitive to small variations than MAE, remained stable when comparing different architectures within the same centre. Consequently, the focus shifted to evaluating the disparity in efficiency among the different implementations, as this becomes a critical factor in determining the most suitable model for FL applications.

Referring to results about the time taken to complete the experiment for a fixed number of rounds (Total Time, see Table 2), the U-Net proposed by Li et al. [54] was the most efficient. These findings are relevant in federated environments, where limited computational resources and the time spent on computation become major factors. Consequently, a model that requires less time to be trained and federated, might mitigate interruptions in the communication during training and minimise synchronisation times.

The employment of U-Net architectures was further justified by supplementary experiments comparing various models, including Conditional GAN, CycleGAN and Transformers. The results of these experiments, which are discussed in Appendix A, demonstrated that the U-Net achieved the optimal overall performance, balancing image similarity metrics and computational efficiency, thus rendering them more suitable for FL MRI-to-sCT applications.

Moving to the aggregation strategies, the analysis of FedAvg results (see Table 3, revealed that the optimal performance were obtained in 25 rounds, exhibiting minimal variation – 1.4 HU – in terms of MAE. However, when compared to the proposed aggregation method (FedAvg+FedProx), FedAvg required an average higher number of rounds, indicating a longer time to reach its optimal result. Referring to the Table 3, using FedProx enhances the performance of FedAvg and accelerates the convergence process, achieving its optimal result in an average of 12.6 ± 3.7 rounds, compared to 14.8 ± 4.1 rounds for the FedAvg alone. This finding aligns with previous studies in closer fields of research, which highlight that the introduction of a proximal term can mitigate the impact of client heterogeneity and improve convergence stability in FL settings.

Despite the introduction of the momentum coefficient and the server-side learning rate, the FedAvgM strategy was unable to improve the performance in terms of image similarity. Also in this case, a minimal stochastic variation in terms of MAE was observed. As illustrated in both Fig. D.10 and Table 3, while the trend of convergence remains

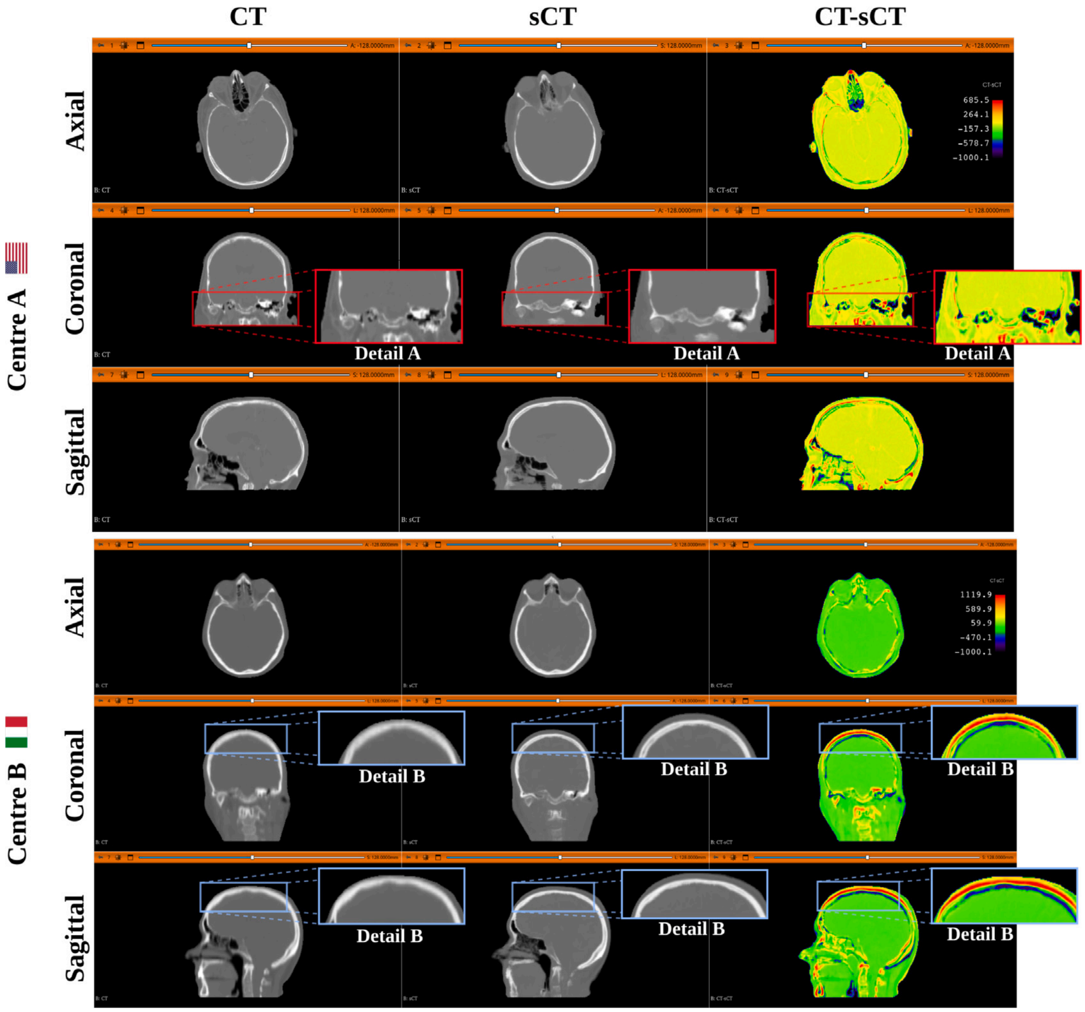

Fig. 7. The figure shows the central slice for the axial, coronal, and sagittal planes, which is used for an assessment of one of the two test cases of Centres A and B, respectively. The Detail A (in red) comparison highlight the approximation of the small structure performed by the model during the sCT generation for Centre A. Detail B comparison (in blue) shows the presence of artefact in Centre B's ground-truth CTs, whereby the upper skull is larger than expected. Despite this artefact, the federated model predicts a smaller skull structure, and thus a more realistic outcome.

consistent across experiments, FedAvgM requires 27.8 ± 0.7 (mean ± std) rounds, thereby doubling the number of rounds to achieve the optimal result in comparison to the proposed aggregation strategy. Consequently, it can be deduced that the momentum coefficient and the server-side learning rate do not offer any advantages in terms of performance or image quality when compared to the proposed method. Similar challenges with FedAvgM have been reported in [66], where the additional hyperparameters introduced by momentum strategies were found to complicate tuning without consistently improving performance.

The more complex and optimised strategy FedYogi, presented several stability problems in the performed experiments. Despite the efforts made to tune the hyper-parameters of the strategy, it did not yield comparable or acceptable results compared to simpler strategies, as evidenced by the average MAE of 226.9 ± 10.2.

From Table 3, a delay in the convergence of the simpler FedAvg has been noted when introducing FedBN, in contrast to the approach of combining only FedProx.

To confirm the benefit of the proximal term in this specific federated task, a combination of FedProx with FedAvg and FedBN (FedAvg + FedProx + FedBN) was investigated. It is evident from Table 3 that, despite the lack of improvement, the introduction of the proximal term (FedProx) resulted in a slight reduction in the average number of rounds required for convergence. Consequently, we concluded that the chosen strategy, which involves the simple combination of FedAvg with FedProx, with a proximal coefficient $\mu = 3$, offers a comparable performance to more complex and optimised strategies, while requiring fewer rounds.

As in the case of the U-Nets evaluation, our findings reaffirm previous insights into the balance between model efficiency and efficacy [56] while highlighting opportunities for future exploration, particularly in aggregation strategies that improve robustness and resource efficiency in FL scenarios.

Regarding the generalisation ability of the federated model, the evaluation of CT and sCT similarity on the unseen dataset (see Fig. 6) revealed that for the predominant portion of the image, specifically the soft tissue, the discrepancy ranges between −100 HU and 100 HU, with peak errors occurring in specific and limited regions, such as the nasal and oral cavities.

Some inconsistencies were identified in the upper skull of a few patients, indicated by the Detail B.2 in Fig. 6, case B). The wrong I2I translation produced a subtle bone mirroring. In other cases the federated model was not able to predict the presence of calcifications in the brain tissue. This failure can be attributed to the limited number of examples that exhibited this brain calcification in the federation, which precluded the model from identifying abnormal cases such as the one highlighted.

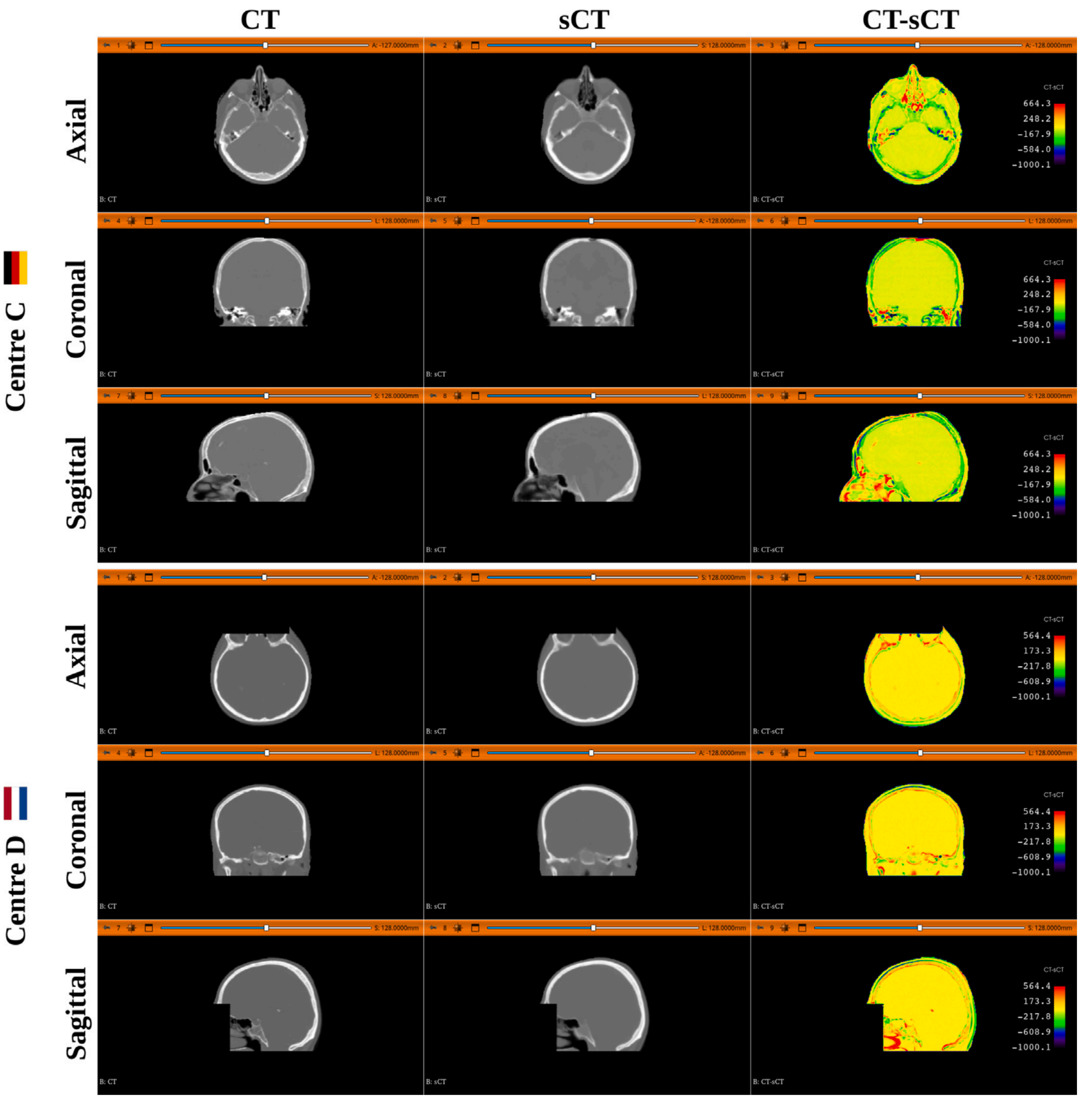

Fig. 8. The results obtained from the federated model for one of the two test cases of Centres C and D. As for the other Centres of the federation, conversion problems emerge in small and detailed areas, while for the majority of the image composed of soft tissue the error remains contained.

Despite the Centre E's dataset being external to the training, the federated model demonstrated its capacity to predict a missing portion of bone in the axial view (see Fig. 6, case A, Detail A).

The results of the individual centres presented in Figs. 7 and 8, exhibited various advantages and disadvantages, given the differing characteristics of the data from each Centre.

In Fig. 7 the aforementioned issue identified in the upper part of the skull for some cases of the Centre E does not occur. However, the smaller and more detailed areas tended to be approximated, resulting in a loss of detail, particularly in the proximity of the nasal cavities, as can be note within the red box provided in the coronal view for the test case of the centre A.

As described in Section 3.1, Centre B was introduced in the federation to increase data heterogeneity, as the MRIs were acquired with a hybrid PET-MRI scanner, with larger CT slice thickness and different X-ray tube energy (120 kV vs. 140 kV). From Table 2, it is evident that the error for the Centre B is greater than for the other clients, emphasising the difference from the other images in the federation. However, it can be observed that the error is partly due to the lower spatial resolution of the ground truth CT, introducing partial volume artefacts in the upper part of the skull (see Fig. 7, Detail B). The result of this artefact is a noticeable, unrealistic thickening of the bone, as presented in Fig. C.9. Despite this problem, the skull bone of the predicted federated sCT appears more realistic and consistent with other examples. This establishes an advantage of the federation, as the influence given by the weights of the other Centres allows the federated model to remain robust in front of optimisation errors during training.

In Fig. 8, results obtained for Centres C and D test cases are presented. The limitations of the model in reconstructing small, detailed parts are also evident in these cases. However, from the difference shown on the left, it can be observed that the error for the largest part of the images is always in the [−100,100] HU range, while the highest errors are consistently present in the proximity of the air and detailed zones.

Furthermore, in order to support the validity of the proposed training strategy, the results obtained through alternative well-established training methodologies using the Li et al. model were presented and discussed in detail in Appendix B, demonstrating the effectiveness of the chosen Random Multi-2D training approach.

This work established the foundations for future developments. Several limitations of the employed architecture could be addressed by adopting more advanced medical image synthesis methodologies [24–26], optimising their efficiency, and ensuring compatibility with the FL environment. A possible improvement could also involve optimising the implementation of 3D architectures for federated scenarios by leveraging more efficient training strategies, such as model pruning, knowledge distillation, or dimensionality reduction [67]. Additionally, exploring hybrid federated approaches, such as cross-silo horizontal FL with asynchronous updates or client clustering [68,69], could reduce synchronisation times.

A further limitation of the present study was the available computational resources, which imposed a limit on the number of clients that could be incorporated into the federation. In future work, a real-world deployment could be considered to mitigate the constraints imposed by a single, although powerful, machine. In addition, the evaluation of the generated sCTs using the emerging sCT quality control algorithms [70] has the potential to enhance the comprehension of the limitations of the federated framework and to introduce enhancements in future studies.

Moreover, the implementation of specific pre-processing or data harmonisation techniques can be considered to enhance the homogeneity of the data and to mitigate potential hidden artefacts [13]. This, consequently, will improve the quality of the sCTs for both the Centres within the federation and in the generalisation task.

7. Conclusion

We proposed, for the first time to best of our knowledge, a FL approach for MRI to sCT, with a specific focus on the brain region. Despite the intrinsic complexity of the task, which was evaluated including real clinical datasets. The results illustrate the potential of this methodology to safeguard patient confidentiality while sustaining robust model performance. The incorporation of additional clinics and institutions into the federation will enhance the accessibility of diverse and heterogeneous data, thereby optimising the model's generalisability and performance in a secure and scalable manner. This approach allows for the creation of a model with the capacity to generalise to previously unseen datasets, which is a crucial aspect for wider clinical applicability. Furthermore, it underscores the viability of federated learning as a sustainable solution for medical imaging tasks.

Apart from the advantages that have been highlighted, the objective of this work was to establish the foundations for future research. Multiple approaches can be followed to enhance the final outcome and progressively reduce the necessity for data centralisation and the training of individual models for each clinic. Future developments will include dosimetric studies to establish whether the results obtained can be used also in the clinical environment for external beam radiotherapy and targeted radionuclide treatments, as well as the benchmarking of aggregation methods that extend beyond classical methodologies but necessitate additional effort for implementation, such as FedFisher [71] and FedDG [72]. Furthermore, efforts will be made to integrate and optimise recent sCT methods employing complex 3D architectures.

CRediT authorship contribution statement

Ciro Benito Raggio: Writing – original draft, Writing – review & editing, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Mathias Krohmer Zabaleta:** Validation, Software, Methodology, Investigation, Data curation. **Nils Skupien:** Validation, Software, Methodology, Investigation, Data curation. **Oliver Blanck:** Writing – review & editing, Resources, Data curation. **Francesco Cicone:** Writing – review & editing, Resources. **Giuseppe Lucio Cascini:** Resources. **Paolo Zaffino:** Writing – review & editing, Resources, Data curation. **Lucia Migliorelli:** Writing – original draft, Writing – review & editing, Visualization, Formal analysis. **Maria Francesca Spadea:** Writing – original draft, Writing – review & editing, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Formal analysis, Conceptualization.

Ethics statement

All data obtained from institutions A, B, and C were collected with proper approvals, and written informed consent was obtained from all participants. The data were handled in accordance with the ethical standards of the 1964 Declaration of Helsinki and its subsequent amendments. For the datasets from centres D and E, no consent was required as the data are publicly available.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We would like to acknowledge Dr. C. Catana (Massachusetts General Hospital) for contributing to data sharing and support during this research.

Appendix A. Investigation of deep learning architectures

This preliminary investigation was conducted using data from centres A and C, with Centre E used as the unseen dataset. This choice reduced computational time and allowed evaluation of multiple architectures. The hyper-parameters used are detailed in Section 4.2. The minimal augmentation pipeline was used to further reduce the computational time. Among the selected architectures, were: U-Net (described in Section 3.3), Conditional GAN (Pix2Pix GAN, with a U-Net-based generator) [73], and CycleGAN (with a ResNet-based generator) [74], which are commonly used in image-to-image translation or CT synthesis tasks [1,2,4]. Despite its inherent complexity, we also evaluated the winning deep learning (DL) model of the SynthRAD challenge, which used a patch-based approach for centralised MRI-to-sCT [14], namely the MSEP architecture [16].

Pix2Pix Conditional GAN. The Pix2Pix conditional GAN (cGAN) model employed a U-Net-based generator and a PatchGAN discriminator, used to have a 30 × 30 probability patch in output rather than a single value. For this model, an hybrid loss that integrated the pixel-wise error of the prediction and the discriminator's assessment was employed.

CycleGAN. The CycleGAN used in this study was composed of two generators and two discriminators. The generator networks were implemented as instances of a ResNet, while PatchGANs with an output size of 30 × 30 were employed for the discriminators. The generator loss was calculated as a weighted average of three components: paired pixel-wise loss, cycle-consistency loss, and adversarial loss.

MSEP. To enable its use in a FL context, we adapted the original implementation accordingly. To train the MSEP, patches with dimensions 48 × 160 × 160 were extracted from each sample. The location of each patch was chosen randomly every epoch to increase the amount of possible training input data needed by intrinsic nature of the architecture. Since with this 3D DL approach, the model predicted overlapping patches and the number of predictions was not constant across the volume, the average value was calculated for each voxel using the median voting approach to reach a consensus.

In this scenario, the hardware needed to lead the experiment was an NVIDIA A100 80 GB GPU, 32 CPU cores, and 256 GB of RAM from a Supermicro AS-4124GS-TNR 4U Rackmount GPU SuperServer.

The results reported in A.4 demonstrated that the U-Net obtained the optimal overall performance, taking into account both image similarity metrics and computational efficiency, with a training duration of 6 h employing only Centre A and C.

Table A.4
The median and interquartile range (IR) for the mean absolute error (MAE), the structural similarity index (SSIM), and the peak signal-to-noise ratio (PSNR) obtained using different architectures, as well as the total time required for the experiments.

Model	Centre E (Unseen)			
	MAE [HU]	SSIM	PSNR [dB]	Time
U-Net	**110.4 (91.2–126.1)**	**0.88 (0.85–0.89)**	**26.44 (25.42–27.27)**	6 h 00 min
Pix2Pix	122.0 (104.6–139.1)	0.87 (0.85–0.87)	26.31 (25.27–26.98)	**4 h 00 min**
CycleGAN	121.4 (115.7–130.9)	0.87 (0.85–0.88)	26.30 (25.32–27.07)	10 h 28 min
MSEP	240.5 (220.7–250.7)	0.79 (0.77–0.80)	23.70 (23.09–24.29)	85 h 37 min[a]

[a] Time required to perform a single round.

Table B.5
The median and interquartile range (IR) for the mean absolute error (MAE), structural similarity index (SSIM), and peak signal-to-noise ratio (PSNR) obtained for each training paradigm. The total training time required to complete 25 rounds is also reported.

Method	Centre E (Unseen)			
	MAE [HU]	SSIM	PSNR [dB]	Time
2D+	104.0 (99.5–115.6)	0.89 (0.86–0.89)	26.49 (25.43–27.35)	16 h 00 min
2D Patches	162.4 (155.3–173.0)	0.84 (0.82-0.85)	25.52 (24.65–26.12)	**8 h 39 min**
Multi-2D	111.7 (107.7–122.3)	0.88 (0.85–0.89)	26.33 (25.32–27.20)	9 h 35 min
Random Multi-2D (Our)	**102.0 (96.7–110.5)**	**0.89 (0.86–0.89)**	**26.58 (25.52–27.42)**	12 h 34 min

Although the timing of Pix2Pix GAN was favourable, its conversion performance was inferior to that of U-Net under the same conditions. This difference was especially evident from the median and interquartile range of the MAE presented in Table A.4.

The CycleGAN model demonstrated inferior performance in terms of image similarity when employing equivalent settings, with a training duration of 10 h and 28 min that was nevertheless less efficient than the U-Net.

Regarding the MSEP transformer-based architecture, despite efforts to reduce computational overhead — such as limiting the number of participating clients and reducing the size of the 3D patches — the time required for a single federated round remained prohibitively high. Specifically, one round took approximately 85 h and 37 min. Moreover, the federated MSEP architecture's performance fell short when compared to its centralised counterpart. This decline can be primarily attributed to: (i) the adjustments made to the training pipeline to accommodate the federated framework, and (ii) the reduced number of rounds and clients necessitated by the computational complexity of the 3D approach.

Appendix B. Training methods investigation

The primary objective of this investigation was to further evaluate the effectiveness of the proposed FL training approach compared to well-established training methods described in the literature for the MRI-to-sCT task in a centralised setting, described in Section 2.1. This investigation was applied to the selected model, U-Net by Li et al. [54], which had previously demonstrated better performance in terms of efficiency when compared with different U-Nets. The hyper-parameters used are detailed in Section 4.2. The model was trained across Centres A, B, C, D, with Centre E serving as an unseen test dataset. The performance of each training methodology was assessed using the metrics described in Section 4.3, computed on the unseen dataset. In the patch-based method, we extracted randomly cropped patches of 160 × 160 pixels to maximise contextual information. Overlapping regions were merged using pixel-wise averaging.

The results presented in Table B.5 further supported the superiority of the proposed Random Multi-2D approach over the alternative 2D methods evaluated. Specifically, the Random Multi-2D method achieved performance in terms of image similarity metrics comparable to the 2D+ approach, where each model was specifically trained on a single plane (axial, coronal and sagittal). Our training method required only a single model for all anatomical planes, thereby simplifying the federated training and inference processes. Additionally, the training time was significantly reduced with the Random Multi-2D model completing training in 12 h and 34 min, compared to 16 h for the 2D+ models. In contrast, the 2D patch-based method requiring less training time but exhibited inferior performance across all image similarity metrics.

In comparison with the classical Multi-2D [2] method, the proposed Random Multi-2D approach yielded superior image similarity metrics, albeit at a higher computational cost, which can be attributed to the shuffling operation conducted by local clients during federated training. Moreover, the utilisation of our method enabled the federated model to attain superior outcomes in comparison to alternative training methods after only 5 rounds.

The findings of this additional study demonstrated that the Random Multi-2D paradigm not only exhibited competitive performance in comparison to the 2D+ approach, but also demonstrated improved computational efficiency and substituting a more generic model for three specific models. Furthermore, the proposed training method yielded superior generalisation performance compared to the classical Multi-2D method, thus justifying the suitability of the Random Multi-2D paradigm for FL MRI-to-sCT scenarios.

Appendix C. Centre B artefact measurement

See Fig. C.9.

Appendix D. Aggregation strategies investigation

Convergence trends

See Fig. D.10.

Generated sCTs

As shown in Table 3, the evaluated strategies produced qualitatively comparable results, with similar mean absolute errors (MAE). This observation is further supported by Fig. D.11, which provides a visual comparison of the results obtained using different aggregation strategies on an axial slice of an example sCT from the same patient of the unseen Centre E. The discrepancy between the ground truth CT and the sCT is also illustrated. Notably, the sCT corresponding to FedYogi demonstrated a more pronounced discrepancy from the ground truth CT, in line with the highest MAE value reported in Table 3. This finding confirmed the limitations of FedYogi for the MRI-to-sCT task compared to the other aggregation strategies, as well as the pivotal

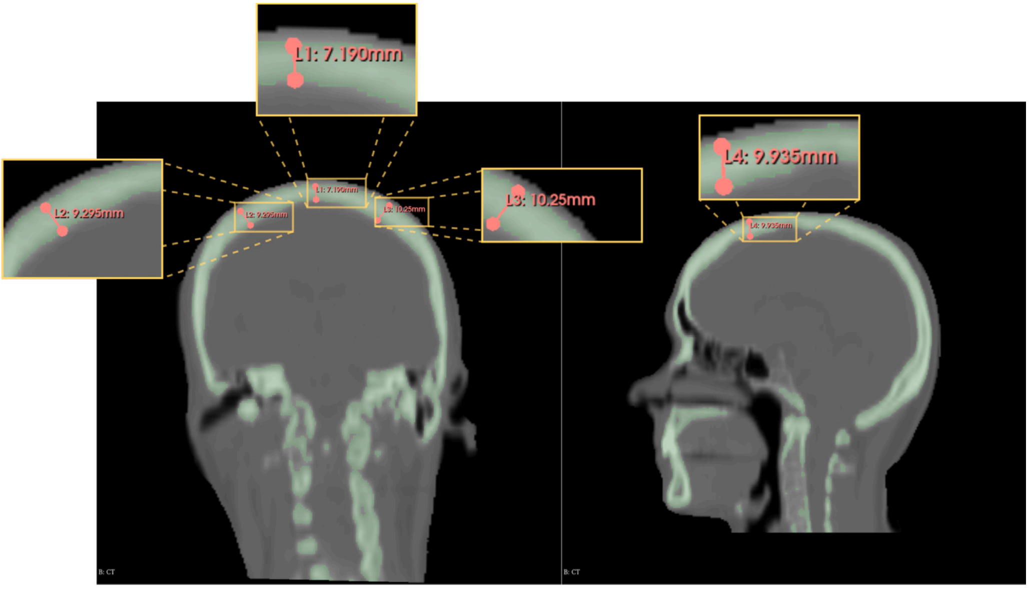

Fig. C.9. Measurement of a portion of the skull to highlight the partial volume artefact due to the higher slice thickness in Centre B CTs.

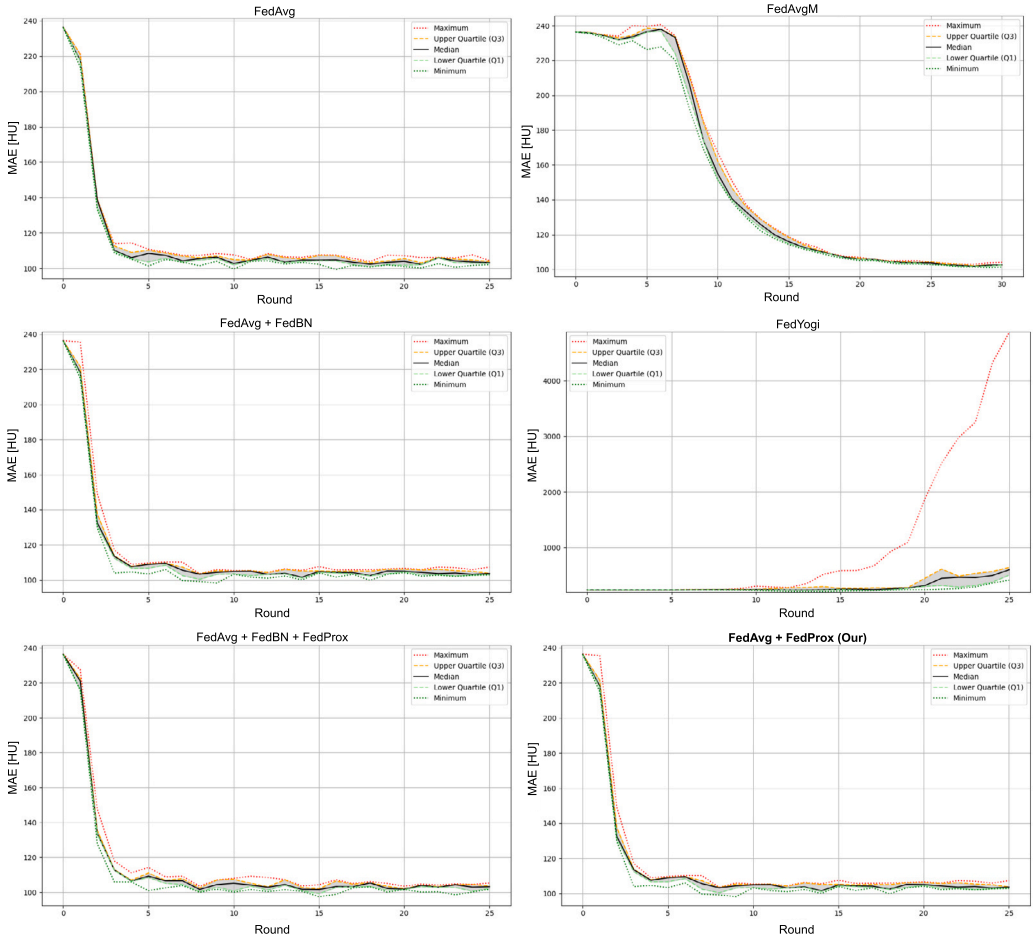

Fig. D.10. Comparison of convergence trends of aggregation strategies.

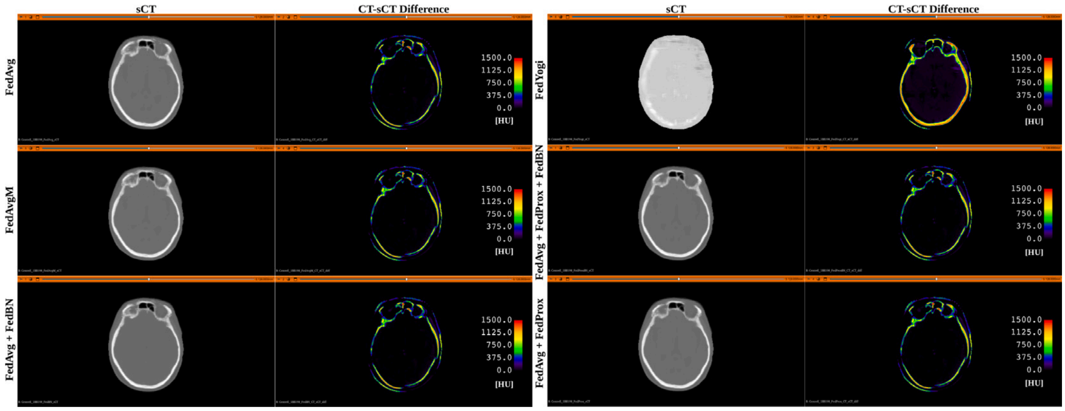

Fig. D.11. Visual comparison of sCT generated with different aggregation strategies on an axial slice from a patient of Centre E. The discrepancy between the ground truth CT and the sCT is highlighted on the right.

role of the aggregation strategy in the federated context. Indeed, as well as determining the quality of the generated sCT, the strategy had a decisive influence on model convergence.

Data availability

The data used for the study acquired from centre A, B and C are confidential. The data from Centre D and Centre E were extracted from the public SynthRAD2023 Grand Challenge dataset and are available at https://synthrad2023.grand-challenge.org/.

References

[1] S. Dayarathna, K.T. Islam, S. Uribe, G. Yang, M. Hayat, Z. Chen, Deep learning based synthesis of mri, ct and pet: Review and analysis, Med. Image Anal. 92 (2024) 103046, http://dx.doi.org/10.1016/j.media.2023.103046.

[2] M.F. Spadea, M. Maspero, P. Zaffino, J. Seco, Deep learning based synthetic-ct generation in radiotherapy and pet: A review, Med. Phys. 48 (2021) 6537–6566, http://dx.doi.org/10.1002/mp.15150.

[3] A. Thummerer, P. Zaffino, M.F. Spadea, A. Knopf, M. Maspero, Artificial intelligence to generate synthetic ct for adaptive particle therapy, in: Imaging in Particle Therapy, IOP Publishing, 2024, pp. 8–1 to 8–16, http://dx.doi.org/10.1088/978-0-7503-5117-1ch8, 2053-2563.

[4] M. Boulanger, J.-C. Nunes, H. Chourak, A. Largent, S. Tahri, O. Acosta, R. De Crevoisier, C. Lafond, A. Barateau, Deep learning methods to generate synthetic ct from mri in radiotherapy: A literature review, Phys. Medica 89 (2021) 265–281, URL http://dx.doi.org/10.1016/j.ejmp.2021.07.027.

[5] Y. Liu, A. Chen, H. Shi, S. Huang, W. Zheng, Z. Liu, Q. Zhang, X. Yang, Ct synthesis from mri using multi-cycle gan for head-and-neck radiation therapy, Comput. Med. Imaging Graph. 91 (2021) 101953, http://dx.doi.org/10.1016/j.compmedimag.2021.101953.

[6] R. Graf, J. Schmitt, S. Schlaeger, H.K. Möller, V. Sideri-Lampretsa, A. Sekuboyina, S.M. Krieg, B. Wiestler, B. Menze, D. Rueckert, J.S. Kirschke, Denoising diffusion-based mri to ct image translation enables automated spinal segmentation, Eur. Radiol. Exp. 7 (2023) 70, http://dx.doi.org/10.1186/s41747-023-00385-2.

[7] M. Özbey, O. Dalmaz, S.U. Dar, H.A. Bedel, Ş. Özturk, A. Güngör, T. Çukur, Unsupervised medical image translation with adversarial diffusion models, IEEE Trans. Med. Imaging 42 (2023) 3524–3539, http://dx.doi.org/10.1109/TMI.2023.3290149.

[8] M.A. Bahloul, S. Jabeen, S. Benoumhani, H.A. Alsaleh, Z. Belkhatir, A. Al-Wabil, Advancements in synthetic ct generation from mri: A review of techniques, and trends in radiation therapy planning, J. Appl. Clin. Med. Phys. 25 (2024) http://dx.doi.org/10.1002/acm2.14499.

[9] J. Grigo, J. Szkitsak, D. Höfler, R. Fietkau, F. Putz, C. Bert, Sct-feasibility - a feasibility study for deep learning-based mri-only brain radiotherapy, Radiat. Oncol. 19 (2024) 33, http://dx.doi.org/10.1186/s13014-024-02428-3.

[10] O. Dalmaz, M.U. Mirza, G. Elmas, M. Özbey, S.U. Dar, E. Ceyani, K.K. Oguz, S. Avestimehr, T. Çukur, One model to unite them all: Personalized federated learning of multi-contrast mri synthesis, Med. Image Anal. 94 (2024) 103121, http://dx.doi.org/10.1016/j.media.2024.103121.

[11] S. Tahri, B. Texier, J.-C. Nunes, C. Hemon, P. Lekieffre, E. Collot, H. Chourak, J. Le Guevelou, P. Greer, J. Dowling, O. Acosta, I. Bessieres, L. Marage, A. Boue-Rafle, R. De Crevoisier, C. Lafond, A. Barateau, A deep learning model to generate synthetic ct for prostate mr-only radiotherapy dose planning: a multicenter study, Front. Oncol. 13 (2023) URL http://dx.doi.org/10.3389/fonc.2023.1279750.

[12] B. Texier, C. Hémon, P. Lekieffre, E. Collot, S. Tahri, H. Chourak, J. Dowling, P. Greer, I. Bessieres, O. Acosta, A. Boue-Rafle, J.L. Guevelou, R. de Crevoisier, C. Lafond, J. Castelli, A. Barateau, J.-C. Nunes, Computed tomography synthesis from magnetic resonance imaging using cycle generative adversarial networks with multicenter learning, Phys. Imaging Radiat. Oncol. 28 (2023) 100511, URL http://dx.doi.org/10.1016/j.phro.2023.100511.

[13] S. Han, B. Texier, C. Hémon, Y. Kortli, A. Queffelec, R. de Crevoisier, J. Dowling, P. Greer, I. Bessières, A. Barateau, C. Lafond, J.-C. Nunes, Balancing data consistency and diversity: Preprocessing and online data augmentation for multi-center deep learning-based mr-to-ct synthesis, Pattern Recognit. Lett. (2025) URL http://dx.doi.org/10.1016/j.patrec.2025.01.002.

[14] E.M. Huijben, M.L. Terpstra, S. Pai, A. Thummerer, P. Koopmans, M. Afonso, M. van Eijnatten, O. Gurney-Champion, Z. Chen, Y. Zhang, et al., Generating synthetic computed tomography for radiotherapy: Synthrad2023 challenge report, Med. Image Anal. 97 (2024) 103276.

[15] E.M. Huijben, M.L. Terpstra, A.J. Galapon, S. Pai, A. Thummerer, P. Koopmans, M. Afonso, M. van Eijnatten, O. Gurney-Champion, Z. Chen, Y. Zhang, K. Zheng, C. Li, H. Pang, C. Ye, R. Wang, T. Song, F. Fan, J. Qiu, Y. Huang, J. Ha, J. Sung Park, A. Alain-Beaudoin, S. Bériault, P. Yu, H. Guo, Z. Huang, G. Li, X. Zhang, Y. Fan, H. Liu, B. Xin, A. Nicolson, L. Zhong, Z. Deng, G. Müller-Franzes, F. Khader, X. Li, Y. Zhang, C. Hémon, V. Boussot, Z. Zhang, L. Wang, L. Bai, S. Wang, D. Mus, B. Kooiman, C.A. Sargeant, E.G. Henderson, S. Kondo, S. Kasai, R. Karimzadeh, B. Ibragimov, T. Helfer, J. Dafflon, Z. Chen, E. Wang, Z. Perko, M. Maspero, Generating synthetic computed tomography for radiotherapy: Synthrad2023 challenge report, Med. Image Anal. 97 (2024) 103276, http://dx.doi.org/10.1016/j.media.2024.103276, URL https://www.sciencedirect.com/science/article/pii/S1361841524002019.

[16] Z. Chen, K. Zheng, C. Li, Z. Yiwen, A hybrid network with multi-scale structure extraction and preservation for mr-to-ct synthesis in synthrad2023, 2023, SynthRAD2023.

[17] G.A. Kaissis, M.R. Makowski, D. Rückert, R.F. Braren, Secure, privacy-preserving and federated machine learning in medical imaging, Nat. Mach. Intell. 2 (2020) 305–311.

[18] H. Guan, P.-T. Yap, A. Bozoki, M. Liu, Federated learning for medical image analysis: A survey, Pattern Recognit. 151 (2024) 110424, http://dx.doi.org/10.1016/j.patcog.2024.110424, URL https://www.sciencedirect.com/science/article/pii/S0031320324001754.

[19] M. Adnan, S. Kalra, J.C. Cresswell, G.W. Taylor, H.R. Tizhoosh, Federated learning and differential privacy for medical image analysis, Sci. Rep. 12 (2022) 1953.

[20] N. Hernandez-Cruz, P. Saha, M.M.K. Sarker, J.A. Noble, Review of federated learning and machine learning-based methods for medical image analysis, Big Data Cogn. Comput. 8 (2024) http://dx.doi.org/10.3390/bdcc8090099.

[21] Y. Lei, J. Harms, T. Wang, Y. Liu, H.-K. Shu, A.B. Jani, W.J. Curran, H. Mao, T. Liu, X. Yang, MRI-only based synthetic CT generation using dense cycle consistent generative adversarial networks, Med. Phys. 46 (2019) 3565–3581.

[22] H. Koh, T.Y. Park, Y.A. Chung, J.-H. Lee, H. Kim, Acoustic simulation for transcranial focused ultrasound using gan-based synthetic ct, IEEE J. Biomed. Heal. Inform. 26 (2022) 161–171, http://dx.doi.org/10.1109/JBHI.2021.3103387.

[23] R.S. Antunes, C. André da Costa, A. Küderle, I.A. Yari, B. Eskofier, Federated learning for healthcare: Systematic review and architecture proposal, ACM Trans. Intell. Syst. Technol. (TIST) 13 (2022) 1–23.

[24] O.F. Atli, B. Kabas, F. Arslan, A.C. Demirtas, M. Yurt, O. Dalmaz, T. Çukur, I2i-mamba: Multi-modal medical image synthesis via selective state space modeling, 2024, URL https://arxiv.org/abs/2405.14022. arXiv:2405.14022.

[25] B. Kabas, F. Arslan, V.A. Nezhad, S. Ozturk, E.U. Saritas, T. Çukur, Physics-driven autoregressive state space models for medical image reconstruction, 2024, URL https://arxiv.org/abs/2412.09331. arXiv:2412.09331.

[26] F. Arslan, B. Kabas, O. Dalmaz, M. Ozbey, T. Çukur, Self-consistent recursive diffusion bridge for medical image translation, 2024, URL https://arxiv.org/abs/2405.06789. arXiv:2405.06789.

[27] D. Saxena, J. Cao, Generative adversarial networks (gans): Challenges, solutions, and future directions, ACM Comput. Surv. 54 (2021) 1–42, URL http://dx.doi.org/10.1145/3446374.

[28] B. Texier, C. Hémon, A. Queffélec, J. Dowling, I. Bessieres, P. Greer, O. Acosta, A. Boue-Rafle, R. de Crevoisier, C. Lafond, J. Castelli, A. Barateau, J.-C. Nunes, 3D unsupervised deep learning method for magnetic resonance imaging-to-computed tomography synthesis in prostate radiotherapy, Phys. Imaging Radiat. Oncol. 31 (2024) 100612, http://dx.doi.org/10.1016/j.phro.2024.100612, URL https://www.sciencedirect.com/science/article/pii/S2405631624000824.

[29] G. Li, L. Bai, C. Zhu, E. Wu, R. Ma, A novel method of synthetic ct generation from mr images based on convolutional neural networks, in: 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics, CISP-BMEI, 2018, pp. 1–5, http://dx.doi.org/10.1109/CISP-BMEI.2018.8633142.

[30] M.F. Spadea, G. Pileggi, P. Zaffino, P. Salome, C. Catana, D. Izquierdo-Garcia, F. Amato, J. Seco, Deep convolution neural network (dcnn) multiplane approach to synthetic ct generation from mr images—application in brain proton therapy, Int. J. Radiat. Oncol. Biol. Phys. 105 (2019) 495–503, http://dx.doi.org/10.1016/j.ijrobp.2019.06.2535.

[31] S.-H. Hsu, Z. Han, J.E. Leeman, Y.-H. Hu, R.H. Mak, A. Sudhyadhom, Synthetic ct generation for mri-guided adaptive radiotherapy in prostate cancer, Front. Oncol. 12 (2022) http://dx.doi.org/10.3389/fonc.2022.969463.

[32] M.J. Sheller, G.A. Reina, B. Edwards, J. Martin, S. Bakas, Multi-Institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation, Brainlesion 11383 (2019) 92–104.

[33] W. Li, F. Milletarì, D. Xu, N. Rieke, J. Hancox, W. Zhu, M. Baust, Y. Cheng, S. Ourselin, M.J. Cardoso, A. Feng, Privacy-preserving federated brain tumour segmentation, 2019, arXiv:1910.00962.

[34] D. Li, A. Kar, N. Ravikumar, A.F. Frangi, S. Fidler, Fed-sim: Federated simulation for medical imaging, 2020, arXiv:2009.00668.

[35] Q. Chang, H. Qu, Y. Zhang, M. Sabuncu, C. Chen, T. Zhang, D. Metaxas, Synthetic learning: Learn from distributed asynchronized discriminator gan without sharing medical image data, 2020, arXiv:2006.00080.

[36] A. Xu, W. Li, P. Guo, D. Yang, H.R. Roth, A. Hatamizadeh, C. Zhao, D. Xu, H. Huang, Z. Xu, Closing the generalization gap of cross-silo federated medical image segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 20866–20875.

[37] S.S. Sandhu, H.T. Gorji, P. Tavakolian, K. Tavakolian, A. Akhbardeh, Medical imaging applications of federated learning, Diagnostics 13 (2023) http://dx.doi.org/10.3390/diagnostics13193140.

[38] I. Feki, S. Ammar, Y. Kessentini, K. Muhammad, Federated learning for covid-19 screening from chest x-ray images, Appl. Soft Comput. 106 (2021) 107330, http://dx.doi.org/10.1016/j.asoc.2021.107330, URL https://www.sciencedirect.com/science/article/pii/S1568494621002532.

[39] D. Yang, Z. Xu, W. Li, A. Myronenko, H.R. Roth, S. Harmon, S. Xu, B. Turkbey, E. Turkbey, X. Wang, W. Zhu, G. Carrafiello, F. Patella, M. Cariati, H. Obinata, H. Mori, K. Tamura, P. An, B.J. Wood, D. Xu, Federated semi-supervised learning for covid region segmentation in chest ct using multi-national data from China, Italy, Japan, Med. Image Anal. 70 (2021) 101992, http://dx.doi.org/10.1016/j.media.2021.101992, URL https://www.sciencedirect.com/science/article/pii/S1361841521000384.

[40] J. Wicaksana, Z. Yan, X. Yang, Y. Liu, L. Fan, K.-T. Cheng, Customized federated learning for multi-source decentralized medical image classification, IEEE J. Biomed. Heal. Inform. 26 (2022) 5596–5607, http://dx.doi.org/10.1109/JBHI.2022.3198440.

[41] M.N. Hossen, V. Panneerselvam, D. Koundal, K. Ahmed, F.M. Bui, S.M. Ibrahim, Federated machine learning for detection of skin diseases and enhancement of internet of medical things (iomt) security, IEEE J. Biomed. Heal. Inform. 27 (2023) 835–841, http://dx.doi.org/10.1109/JBHI.2022.3149288.

[42] Z. Yan, J. Wicaksana, Z. Wang, X. Yang, K.-T. Cheng, Variation-aware federated learning with multi-source decentralized medical image data, IEEE J. Biomed. Heal. Inform. 25 (2021) 2615–2628, http://dx.doi.org/10.1109/JBHI.2020.3040015.

[43] J. Lo, T.T. Yu, D. Ma, P. Zang, J.P. Owen, Q. Zhang, R.K. Wang, M.F. Beg, A.Y. Lee, Y. Jia, M.V. Sarunic, Federated learning for microvasculature segmentation and diabetic retinopathy classification of oct data, Ophthalmol. Sci. 1 (2021) 100069, http://dx.doi.org/10.1016/j.xops.2021.100069, URL https://www.sciencedirect.com/science/article/pii/S2666914521000671.

[44] A. Tuladhar, L. Tyagi, R. Souza, N.D. Forkert, Federated learning using variable local training for brain tumor segmentation, in: A. Crimi, S. Bakas (Eds.), Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Springer International Publishing, Cham, 2022, pp. 392–404.

[45] E. Albalawi, M. T.R., V.V. Thakur, M. Kumar, S.B. Gupta, A. Khan, A. Almusharraf, Integrated approach of federated learning with transfer learning for classification and diagnosis of brain tumor, BMC Med. Imaging (2024) http://dx.doi.org/10.1186/s12880-024-01261-0.

[46] M. Zhu, Z. Chen, Y. Yuan, Feddm: Federated weakly supervised segmentation via annotation calibration and gradient de-conflicting, IEEE Trans. Med. Imaging 42 (2023) 1632–1643, http://dx.doi.org/10.1109/TMI.2023.3235757.

[47] L. Qiu, J. Cheng, H. Gao, W. Xiong, H. Ren, Federated semi-supervised learning for medical image segmentation via pseudo-label denoising, IEEE J. Biomed. Heal. Inform. 27 (2023) 4672–4683, http://dx.doi.org/10.1109/JBHI.2023.3274498.

[48] J. Wang, Y. Jin, D. Stoyanov, L. Wang, Feddp: Dual personalization in federated medical image segmentation, IEEE Trans. Med. Imaging 43 (2024) 297–308, http://dx.doi.org/10.1109/TMI.2023.3299206.

[49] Y. Liu, G. Luo, Y. Zhu, Fedfms: Exploring federated foundation models for medical image segmentation, in: M.G. Linguraru, Q. Dou, A. Feragen, S. Giannarou, B. Glocker, K. Lekadir, J.A. Schnabel (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2024, Springer Nature Switzerland, Cham, 2024, pp. 283–293.

[50] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A.C. Berg, W.-Y. Lo, P. Dollar, R. Girshick, Segment anything, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2023, pp. 4015–4026.

[51] J. Wang, G. Xie, Y. Huang, J. Lyu, F. Zheng, Y. Zheng, Y. Jin, Fedmedgan: Federated domain translation on unsupervised cross-modality brain image synthesis, Neurocomputing 546 (2023) 126282, http://dx.doi.org/10.1016/j.neucom.2023.126282.

[52] N.J. Tustison, B.B. Avants, P.A. Cook, Y. Zheng, A. Egan, P.A. Yushkevich, J.C. Gee, N4ITK: improved N3 bias correction, IEEE Trans. Med. Imaging 29 (2010) 1310–1320.

[53] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, J. Finet, J.-C. Fillion-Robin, S. Pujol, C. Bauer, D. Jennings, F. Fennessy, M. Sonka, J. Buatti, S. Aylward, J.V. Miller, S. Pieper, R. Kikinis, 3D slicer as an image computing platform for the quantitative imaging network, Magn. Reson. Imaging 30 (2012) 1323–1341, http://dx.doi.org/10.1016/j.mri.2012.05.001, quantitative Imaging in Cancer.

[54] Y. Li, W. Li, P. He, J. Xiong, J. Xia, Y. Xie, Ct synthesis from mri images based on deep learning methods for mri-only radiotherapy, in: 2019 International Conference on Medical Imaging Physics and Engineering, ICMIPE, 2019, pp. 1–6, http://dx.doi.org/10.1109/ICMIPE47306.2019.9098190.

[55] K. Kamnitsas, C. Ledig, V.F. Newcombe, J.P. Simpson, A.D. Kane, D.K. Menon, D. Rueckert, B. Glocker, Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation, Med. Image Anal. 36 (2017) 61–78.

[56] T. Li, A.K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, V. Smith, Federated optimization in heterogeneous networks, Proc. Mach. Learn. Syst. 2 (2020) 429–450.

[57] X. Li, K. Huang, W. Yang, S. Wang, Z. Zhang, On the convergence of fedavg on non-iid data, 2019, arXiv preprint arXiv:1907.02189.

[58] D.J. Beutel, T. Topal, A. Mathur, X. Qiu, J. Fernandez-Marques, Y. Gao, L. Sani, K.H. Li, T. Parcollet, P.P.B. de Gusmão, N.D. Lane, Flower: A friendly federated learning research framework, 2022, arXiv:2007.14390.

[59] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An Imperative Style, High-Performance Deep Learning Library, Curran Associates Inc., 2019.

[60] X. Chen, X. Zhou, H. Zhang, M. Sun, T. Zhao, Cost-effective federated learning: A unified approach to device and training scheduling, in: ICC 2024-IEEE International Conference on Communications, IEEE, 2024, pp. 3488–3493.

[61] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W.M. Wells, A.F. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, Cham, 2015, pp. 234–241.

[62] J. Fu, Y. Yang, K. Singhrao, D. Ruan, F.-I. Chu, D.A. Low, J.H. Lewis, Deep learning approaches using 2d and 3d convolutional neural networks for generating male pelvic synthetic computed tomography from magnetic resonance imaging, Med. Phys. 46 (2019) 3788–3798, http://dx.doi.org/10.1002/mp.13672.

[63] T. Li, M. Sanjabi, A. Beirami, V. Smith, Fair resource allocation in federated learning, 2020, arXiv:1905.10497.

[64] X. Li, M. Jiang, X. Zhang, M. Kamp, Q. Dou, FedBN: Federated learning on non-iid features via local batch normalization, 2021, arXiv preprint arXiv:2102.07623.

[65] Y. Li, W. Li, J. Xiong, J. Xia, Y. Xie, Comparison of supervised and unsupervised deep learning methods for medical image synthesis between computed tomography and magnetic resonance images, BioMed Res. Int. 2020 (2020) 5193707, http://dx.doi.org/10.1155/2020/5193707.

[66] S. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečnỳ, S. Kumar, H.B. McMahan, Adaptive federated optimization, 2020, arXiv preprint arXiv:2003.00295.

[67] S. Soni, A. Saeed, Y.M. Asano, Federated learning with a single shared image, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 7782–7790.

[68] J. Nguyen, K. Malik, H. Zhan, A. Yousefpour, M. Rabbat, M. Malek, D. Huba, Federated learning with buffered asynchronous aggregation, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2022, pp. 3581–3607.

[69] Z. Wang, H. Xu, J. Liu, Y. Xu, H. Huang, Y. Zhao, Accelerating federated learning with cluster construction and hierarchical aggregation, IEEE Trans. Mob. Comput. 22 (2022) 3805–3822.

[70] P. Zaffino, C.B. Raggio, A. Thummerer, G.G. Marmitt, J.A. Langendijk, A. Procopio, C. Cosentino, J. Seco, A.C. Knopf, S. Both, M.F. Spadea, Toward closing the loop in image-to-image conversion in radiotherapy: A quality control tool to predict synthetic computed tomography hounsfield unit accuracy, J. Imaging 10 (2024) http://dx.doi.org/10.3390/jimaging10120316, URL https://www.mdpi.com/2313-433X/10/12/316.

[71] D. Jhunjhunwala, S. Wang, G. Joshi, Fedfisher: Leveraging fisher information for one-shot federated learning, 2024, URL https://arxiv.org/abs/2403.12329. arXiv:2403.12329.

[72] Q. Liu, C. Chen, J. Qin, Q. Dou, P.-A. Heng, Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space, 2021, arXiv:2103.06030.

[73] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, 2018, URL https://arxiv.org/abs/1611.07004. arXiv:1611.07004.

[74] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: 2017 IEEE International Conference on Computer Vision, ICCV, 2017, pp. 2242–2251, http://dx.doi.org/10.1109/ICCV.2017.244.