Multi-institutional PET/CT image segmentation using federated deep transformer learning

Shiri, Isaac; Razeghi, Behrooz; Vafaei Sadr, Alireza; Amini, Mehdi; Salimi, Yazdan; Ferdowsi, Sohrab; Boor, Peter; Gündüz, Deniz; Voloshynovskiy, Slava; Zaidi, Habib

Published in:
Computer Methods and Programs in Biomedicine

DOI:
10.1016/j.cmpb.2023.107706

Publication date:
2023

Computer Methods and Programs in Biomedicine 240 (2023) 107706

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

Multi-institutional PET/CT image segmentation using federated deep transformer learning

Isaac Shiri [a], Behrooz Razeghi [b], Alireza Vafaei Sadr [c,d], Mehdi Amini [a], Yazdan Salimi [a], Sohrab Ferdowsi [b], Peter Boor [c], Deniz Gündüz [e], Slava Voloshynovskiy [b], Habib Zaidi [a,f,g,h,*]

[a] Division of Nuclear Medicine and Molecular Imaging, Geneva University Hospital, Geneva, Switzerland
[b] Department of Computer Science, University of Geneva, Geneva, Switzerland
[c] Institute of Pathology, RWTH Aachen University Hospital, Aachen, Germany
[d] Department of Public Health Sciences, College of Medicine, The Pennsylvania State University, Hershey, PA 17033, USA
[e] Department of Electrical and Electronic Engineering, Imperial College London, UK
[f] Geneva University Neurocenter, University of Geneva, Geneva, Switzerland
[g] Department of Nuclear Medicine and Molecular Imaging, University of Groningen, Groningen, The Netherlands
[h] Department of Nuclear Medicine, University of Southern Denmark, Odense, Denmark

ARTICLE INFO

Keywords:
PET/CT
Segmentation
Federated learning
Deep transformers
Privacy

ABSTRACT

Background and Objective: Generalizable and trustworthy deep learning models for PET/CT image segmentation necessitates large diverse multi-institutional datasets. However, legal, ethical, and patient privacy issues challenge sharing of datasets between different centers. To overcome these challenges, we developed a federated learning (FL) framework for multi-institutional PET/CT image segmentation.

Methods: A dataset consisting of 328 FL (HN) cancer patients who underwent clinical PET/CT examinations gathered from six different centers was enrolled. A pure transformer network was implemented as fully core segmentation algorithms using dual channel PET/CT images. We evaluated different frameworks (single center-based, centralized baseline, as well as seven different FL algorithms) using 68 PET/CT images (20% of each center data). In particular, the implemented FL algorithms include clipping with the quantile estimator (ClQu), zeroing with the quantile estimator (ZeQu), federated averaging (FedAvg), lossy compression (LoCo), robust aggregation (RoAg), secure aggregation (SeAg), and Gaussian differentially private FedAvg with adaptive quantile clipping (GDP-AQuCl).

Results: The Dice coefficient was 0.80±0.11 for both centralized and SeAg FL algorithms. All FL approaches achieved centralized learning model performance with no statistically significant differences. Among the FL algorithms, SeAg and GDP-AQuCl performed better than the other techniques. However, there was no statistically significant difference. All algorithms, except the center-based approach, resulted in relative errors less than 5% for SUV_{max} and SUV_{mean} for all FL and centralized methods. Centralized and FL algorithms significantly outperformed the single center-based baseline.

Conclusions: The developed FL-based (with centralized method performance) algorithms exhibited promising performance for HN tumor segmentation from PET/CT images.

1. Introduction

1.1. PET/CT-based management of head and neck cancer

Positron emission tomography (PET) and computed tomography (CT) are widely used imaging modalities in cancer diagnosis, staging and restaging, monitoring of treatment response, and radiation treatment planning [1]. Complementary metabolic and anatomical information captured by multimodality PET and CT images, respectively, is commonly used for malignant disease detection, gross tumor volume (GTV), and biological tumor volume (BTV) delineation for radiation therapy planning (RT) [1]. RT plays a major role in the treatment of head and neck (HN) cancer patients, which requires GTV delineation. In addition, tumor delineation is an essential step toward semi-quantitative

* Corresponding author.
E-mail address: habib.zaidi@hcuge.ch (H. Zaidi).

https://doi.org/10.1016/j.cmpb.2023.107706
Received 9 January 2023; Accepted 2 July 2023
Available online 12 July 2023

and quantitative analysis of PET images for staging and response assessment of cancer patients. However, the delineation of GTV on PET/CT images is labor intensive, prone to inter/intra-observer variability, and remains a time-consuming process which involves switching between PET and CT images [2].

The low resolution and noisy nature of PET images and partial volume effects on one hand, and the diverse anatomical variability in the HN region and the presence of highly active lymph nodes and lumen of the airway, on the other hand, challenge the deployment of semi-automated and fully automated PET segmentation algorithms in the clinic [2]. More recently, deep learning (DL) algorithms have been developed for medical image segmentation and, specifically, PET image segmentation [3]. While PET signal is essential for developing DL auto segmenting models for HN patients, anatomical modalities such as CT and MRI can be beneficial for their high resolution and helps the models to better identify the subtle details and accurately delineate boundaries of the tumor.

1.2. HN tumor segmentation from PET/CT images

Andrearczyk et al. [4] proposed fully convolutional 2D and 3D V-Net models for automatically delineating the hepatocellular carcinoma (HCC) tumors and nodal metastases on single- and multi-modality 18F-FDG-PET and CT images. Manually segmented ROI of 202 HCC patients were used as ground truth. They used two approaches for multi-modality modeling; They fed PET and CT images as multiple input channels, or alternatively, in a late fusion approach, averaged the voxel-wise probability outcomes of individual PET and CT models. They achieved a Dice score of 0.48, 0.58, and 0.60 for CT, PET, and late fusion PET/CT models. Also, their model performed better on a 2D basis compared to a similar 3D design. Zhao et al. [5] presented a fully convolutional network with auxiliary paths for automatic segmentation of Nasopharyngeal Carcinoma (NPC) from PET/CT images. They applied their proposed model on 30 patients enrolled from two centers and, with threefold cross-validation, achieved a mean dice score of 0.87. In a study by Guo et al. [6], proposed a DL GTV segmentation framework based on 3D convolution with dense connections based on multi-modality PET/CT images. They split a dataset of 250 HN patients into 140 patients for training, 35 for validation, and 75 patients for testing the proposed model. They compared their proposed model with a 3D U-Net network as the reference model. Their proposed PET/CT Dense-Net showed superior outcomes compared to the 3D U-Net network (Dice 0.73 vs. 0.71) while having fewer parameters to train. The HECKTOR (HEad and neCK TumOR) segmentation challenges are being held in 2020–2022 and continuing in 2023 to address segmentation challenge ([7,8]) in HN patients using PET/CT images. In the second edition (2021) of the HECKTOR challenge, 22 eligible teams participated ([7, 8]). A total number of 325 PET/CT images of HN cancer patients from six centers were split into 224 patients for training and 101 for testing. Models developed by the participants were with the Dice score ranging from 0.63 to 0.78 and the median Hausdorff Distance95% from 6.37 to 3.09. The winner of the challenge [9] achieved an average DSC of 0.78 and a median HD95 of 3.09.

In CNN-based models including Encoder-Decoder explicit long-range and global relation modeling is a major challenge because of the locality of convolution operations [10]. These challenges lead to suboptimal accuracy because of large inter/intra-patient variabilities in HN tumor segmentation using PET/CT images. Transformers that have been successfully used in natural language processing (NLP) and machine translation tasks [11] have recently been shown to outperform CNNs in some image processing tasks ([10,12]). A number of studies implemented the transformer architecture in a variety of learning tasks. For instance, the vision transformer, data-efficient image transformer, and hierarchical Swin transformer have all been successfully used in image classification, image-to-image translation, and image segmentation, respectively ([10,12]). More recently, Swin-U-Net [12], a U-Net-like pure transformer, has been proposed for medical image segmentation and was shown to outperform CNN-based or combination of CNN and transformer (Trans-U-Net) counterparts [10]. The main challenges in transformers are needing large data sets for training and many training parameters in their architecture.

1.3. FL in medical imaging

DL models developed based on single-center datasets face the challenge of model generalizability and result in poor performance for unseen data with the different acquisition, reconstruction, and scanner settings from different centers ([13,14]). In centralized model training, data owners are mandated to pool their data to third-party servers. However, this approach causes ethical and legal concerns as medical data contains highly sensitive private personal information. Federated learning (FL) has been proposed for distributed training without sharing data between different institutions ([13,14]). FL algorithms have been applied to medical image analysis for different tasks, including classification, prognostication, and segmentation ([13,14]). Dayan *et al.* [15] built a predictive model called EXAM (electronic medical record (EMR) chest X-ray AI model) in COVID-19 patients using chest X-ray images and FL across 20 centers. The comparison between FL and center-based model revealed 16% and 38% enhancement in the mean area under the curve (AUC) and generlizablity of the FL model, respectively.

Sheller et al. [16] studied the feasibility of brain tumor segmentation using MR images using FL. They reported identical results for FL with centralized training in multi-modal brain tumor segmentation (Dice score of 0.85 vs 0.86). They implemented two collaborative learning approaches, institutional incremental learning (IIL) and cyclic institutional incremental learning (CIIL), which failed to reach FL performance, and reported that FL outperformed existing collaborative learning approaches. Bercea et al. [17] proposed unsupervised brain pathology (multiple sclerosis and Glioblastoma) segmentation using disentangled FL. They proposed a method that disentangles model parameter spaces into a shape space as they assumed that the brain's anatomical structure is similar across centers. They used open source and in-house datasets for model training and a reported Dice score of 0.38, thus outperforming auto-encoder (42%) and state-of-the-art (SOTA) FL method (11%). In Sarma *et al.* [18], implemented multi-center whole prostate T2-weighted MR image segmentation using 3D anisotropic hybrid network. They reported that FL-based models result in superior and generalizable performance with respect to single center-based models. Dice scores of 0.81, 0.83, and 0.87 were reported for three single center-based models. However, for FL models, they achieved a Dice score of 0.88. Li *et al.* [19] developed a privacy-preserving FL for brain tumor segmentation, which identified a trade-off between performance and cost of privacy. Yang et al. [20] presented a semi-supervised learning-based segmentation for COVID-19 pneumonia using multinational chest CT data from three countries. They reported the effectiveness of the proposed method compared to supervised methods with data sharing. In a recent study, [21], an image-to-image translation task for PET image attenuation correction and scatter compensation was performed using deep FL. Dataset of six different centers (50 patients per center) enrolled and two sequential and parallelled FL algorithms compared with CeBa and CeZe algorithms. Moreover, they reported higher and comparable performance compared to FL algorithms compared to CeBa and CenZe learning algorithms, respectively.

Most recently, Shiri et al. [22] evaluated PET-only image segmentation using FL. Their study enrolled 405 HN cancer patient images from nine different centers. The models were built on cropped PET images using an R2U-Net network. They reported identical performance with a Dice score of 0.84±0.06 vs 0.84±0.05 for FL and centralized approaches, respectively, with no statistically significant differences. In terms of PET parameters, almost zero% relative error (RE%) was reported for both algorithms in SUV_{max}, SUV_{peak}, and in SUV_{mean} RE% of

Table 1
Summary of data description including patient demographics, PET and CT image acquisition, and reconstruction setting for the different centers.

Information	Center 1	Center 2	Center 3	Center 4	Center 5	Center 6
Number of patients	23	32	34	59	81	99
Sex(M/F/NA)	18/5	30/2	25/9	44/15	59/16/6	70/29
Age (mean ± sd)	61 ± 10	55 ± 8	66 ± 9	64 ± 9	61 ± 10	64 ± 10
Weight (Kg)	80.3 ± 15	52.78 ± 14.54	74.57 ± 22.53	76.85 ± 12.10	77.15 ± 17.18	75.88 ± 19.01
N-status (N0/N1/N2/N3/NA)	2/6/13/2	2/7/23/0	3/2/25/4	4/8/40/7	12/13/47/3/6	37/9/50/3
TNM (I, II, III, IV, NA)	2/4/3/14/0	0/0/8/24/0	0/3/2/29/0	0/2/6/50/1	1/5/20/49/0/6	3/18/20/58/0
Chemotherapy (True/False/NA)	23/0/0	0/0/32	17/1/16	52/4/3	51/4/26	54/18/27
Locoregional invasion (True/NA)	0/0/23	4/28/0	6/26/2	4/13/12	9/54/18	14/81/4
PET/CT Scanner	Siemens Biograph	GE-Discovery	GE-Discovery	GE-Discovery ST, Phillips Guardian Body	GE-Discovery ST	Phillips Guardian Body
kVp	120	120	120,140	120	120,140	120,140
Average Tube Current	234.1 ± 31.9	253.2 ± 80.1	191.2 ± 121.6	178.2 ± 96.1	203.2 ± 104.5	384.3 ± 36.2
Matrix Size	512×512	512×512	512×512	512×512	512×512	512×512
Injected Activity (MBq)	405.25±87.02	595.74 ± 119.81	478.01 ± 167.11	372.31 ± 364.76	573.77 ± 71.06	324.33 ± 78.22
Time to Scan (min)	63.35 ± 20.08	93.35 ± 22.08	90.46 ± 18.46	118.25 ± 23.79	102.64 ± 16.46	102.42 ± 15.09
Time Per Bed (min)	2.5 ± 0.15	3.03 ± 0.17	4.87 ± 1.41	4.77 ± 0.72	5.43 ± 1.37	2.49 ± 0.05
Reconstruction	OSEM	OSEM	OSEM	OSEM	OSEM	LOR-RAMLA
Matrix Size	168×168	128×128	128×128	128×128, 144 × 144	128×128	144×144
Slice Thickness (mm)	3	3.27	3.27	3.75 ± 0.35	3.27	4

6.43±4.72 vs 6.61±5.42 were reported for centralized, and FL approaches, respectively. Isik-Polat et al. [23] evaluated different aggregation techniques and hyperparameter values for FL in brain tumor segmentation. They reported higher performance for FedAvgM (federated averaging with server momentum) compared to FedAvg and FedNov (normalized averaging method). In addition, adaptive epochs resulted in faster convergence and higher performance. They concluded that different combinations of hyperparameters may result in lower performance as one parameter may decrease the effectiveness of others. Recently, the Federated Tumor Segmentation (FeTS) challenge, which uses MR images from the BraTS challenge [24], was introduced. FeTS aims to identify optimal weight aggregation and build generalizable models. In a more recent study [25] FL implemented to build a model using 71 site images for detection of the rare disease of glioblastoma, and they reported 33 and 23% improvement in the delineation of surgical targetable tumor and the complete tumor extent, respectively [25].

Although various approaches to FL have been developed to address different issues, including data partition, communication bottleneck, data heterogeneity, and privacy. There is no one-fit-all FL solution that can address all FL challenges ([13,14]). In the current study, we employ different FL approaches for PET/CT image segmentation that have been designed to address different issues and compare them with the centralized benchmark. Considering the pros and cons of each method and client's preferences, one of these approaches can be implemented to train a generalizable model using multicentric data.

The contributions of this research are summarized in the following:

- We provided the integration of purely attention-based transformers and FL algorithms for PET/CT Image segmentation in HN Cancer patients.
- We applied the FL framework for the PET/CT image segmentation, which provides a more generalizable model development in multicenter settings.
- Different FL frameworks are implemented in which each algorithm addresses different challenges of a FL model, including different learning paradigms, aggregation, robustness, privacy, and communication efficiency.
- A comprehensive comparison is performed between center-based, centralized, and FL frameworks.
- A comprehensive quantitative analysis is performed in PET images toward clinical evaluation of segmentation algorithms.

2. Methods

2.1. PET/CT data acquisition and description

In the current study, we enrolled PET/CT images of 328 histologically proven HN cancer patients from six different centers. The number of included patients (after reviewing all patient's PET and CT images in terms of noise and artifacts in all centers) was 23, 32, 34, 59, 81, and 99 from centers 1 to 6, respectively. Different centers acquired and reconstructed 18F-FDG PET/CT images using different scanners and protocols. Detailed information about each center's data (demographic, PET, and CT image acquisition and reconstruction) is provided in Table 1, and more information could be found in ([3,22,26-32]). Ethics approval and consent to participate were unnecessary since the study was performed on open access online dataset. We split the data from each center into a train/validation set (70/10% patients, in total 234/26 patients), and a test set (20% patients, in total 68 patients) with stratification based on centers.

2.2. Manual image segmentation and pre-processing

Manual segmentation of primary tumors performed separately for each center on PET/CT images was used as standard of reference for evaluation. An experienced nuclear medicine physician evaluated and checked all PET/CT segmentations and edited/modified them to offset plausible errors (i.e., missing slices, including lymph nodes, and including the lumen of the airway). PET and CT images were converted to standardized uptake value (SUV) maps and Hounsfield Unit (HU) values. Metal artifacts in CT images were corrected using the iterative metal artifact reduction (iMAR) algorithm [33]. In order to render the computations tractable and to preserve the image resolution, all images were cropped to the HN region with the aid of an automatic CT lung segmentation and body contour extractor [34]. Cropped images were subsequently resized to 200 × 200 with an isotropic voxel size of 1 × 1 × 1mm^3. CT images were clipped to the range [− 1024, 1200] HU to include all HN tissues, and along intact SUV maps were normalized to the range [0,1] for model development. All pre- and post-processing steps were fully automated to ensure fully automated PET/CT image segmentation in a clinical setting.

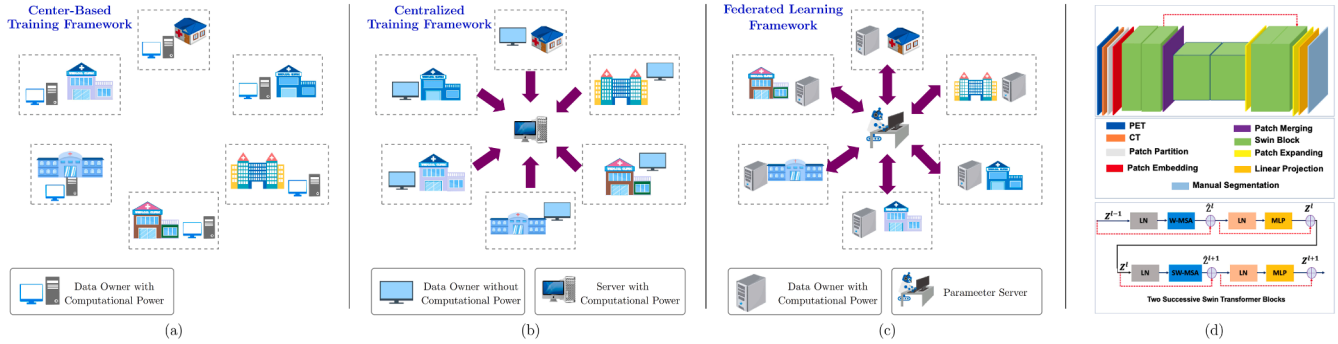

Fig. 1. Visual illustration of (a) center-based training, (b): centralized training, (c): federated learning frameworks; (d) our network architecture.

2.3. FL framework

In general, neural network training methods can be categorized into (i) center-based training framework, (ii) centralized training framework, (iii) distributed training framework, (iv) decentralized training framework, and (v) FL framework. The main difference between these learning frameworks is the way the training data is distributed among the various nodes in the network. Below, we briefly review these training frameworks.

Center-Based (CeBa) Training Framework. In the *center-based training framework*, each party (node) trains its own ML model using its local training dataset, independently of the other centers, and holds the entire control over the functionality of the model. This training framework faces the inability to adapt properly to unseen data.

Centralized (CeZe) Training Framework. In a *centralized training framework*, the participating parties (nodes) send their local data to a centralized server to build and train a global ML model. That is, in a centralized learning framework, all of the training data are *stored* on a *single* node (centralized server), and the other nodes in the network must access these data to train their models. This training framework is a traditional data science pipeline, however, it cannot ensure the privacy and security of the participating data owners.

Distributed Training Framework. In the *distributed training framework*, participating parties independently train ML models using their local datasets and share their local model updates with a server to build the global model. In this learning framework, the training data are divided among multiple nodes, and each node trains its own model using the data it has access to.

Decentralized Training Framework. In a *decentralized training framework*, there is no central node and each node trains its *own* model using the data it has *locally*. Therefore, in this learning framework, there is no server to train a model (like a centralized training framework) or to aggregate the local model updates (like distributed training framework). Instead, the computation process is distributed across all the participating parties.

Federated Learning Framework. In a *FL framework*, the training data remains decentralized and is not shared among the nodes, but the nodes can still collaborate and share their model updates with each other in order to improve the overall performance of the network. In other words, the FL framework is introduced based on a centralized model which uses decentralized model training. That is, the participating parties have their own data, and the ML models are trained independently on the local datasets. Once the local model is trained, each party sends model updates to a central server. Finally, the central server aggregates the model updates to build a global model. Note that, in a distributed training framework [35], we have centralized data and distribute it to computing servers (i.e., workers) for efficient and fast training, while in the FL framework, we have decentralized data and aim to train a global model with the help of a parameter server. In this research, we implement and compare single center training, centralized training, and different FL frameworks (Fig. 1).

2.3.1. Federated deep learning framework

Let $\boldsymbol{\theta} \in \mathbb{R}^d$ denote the parameters of a DL model. Consider $F(\boldsymbol{\theta})$ as an overall loss function. Typically, $F(\boldsymbol{\theta})$ is a non-negative real-valued function computed empirically using available data samples with respect to the model parameters $\boldsymbol{\theta}$. Suppose we have K data centers (owners) that are eager to participate in training a global DL model. Let each of these data centers have a collection of N_k data samples, $k \in \{1, 2, \ldots, K\}$. The local data samples at the k-th center are denoted by $\mathscr{D}_k = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N_k}$, where $\mathbf{x}_i$ and $\mathbf{y}_i$ are the feature vector and the ground-truth label vector, respectively. Let $F_k(\boldsymbol{\theta})$ denote the local aggregated loss corresponding to $\boldsymbol{\theta}$ and all the data samples at the k-th data center (owner). Typically, we take $F_k(\boldsymbol{\theta})$ as follows:

$$F_k(\boldsymbol{\theta}) = \frac{1}{N_k} \sum_{i \in \mathscr{D}_k} \mathscr{L}(\boldsymbol{\theta}; (\mathbf{x}_i, \mathbf{y}_i)), \tag{1}$$

where $\mathscr{L}(\boldsymbol{\theta}; (\mathbf{x}_i, \mathbf{y}_i))$ is the loss of the model parameters $\boldsymbol{\theta}$ for sample $(\mathbf{x}_i, \mathbf{y}_i)$. The distributed learning model objective can then be formulated as the following minimization problem:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} F(\boldsymbol{\theta}) \overset{\Delta}{=} \sum_{k=1}^{K} \frac{N_k}{N} F_k(\boldsymbol{\theta}), \tag{2}$$

where $N = \sum_{k=1}^{K} N_k$ denotes the total number of data samples across K centers. Once the parameter server (possibly a trusted data center) collects the local gradients from the data centers, it updates the global model parameters using the iterative stochastic gradient descent (SGD) algorithm given as:

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \eta \sum_{k=1}^{K} \frac{N_k}{N} \nabla f_k(\boldsymbol{\theta}^t), \tag{3}$$

where η is the learning rate, and $\nabla f_k(\boldsymbol{\theta}^t)$ is the average gradient at center k, computed using the local data samples $\mathscr{D}_k$ and the current model parameter $\boldsymbol{\theta}^t$. The above iterative *distributed SGD* approach is also known as weighted averaging in literature.

In FL, the aggregation method refers to the way the models trained on individual nodes are combined to produce a global model. There are different ways to aggregate the models, and the choice of a method can affect the accuracy and convergence of the global model. In [36], federated averaging (FedAvg) was proposed to optimize the communication-efficiency compared with the naïve distributed SGD method. In FedAvg, firstly, the server initializes a global model parameter and then shares it with a subset of participating data owners, chosen randomly and independently. Next, each data owner performs several epochs of SGD using its local data samples and sends the updated model back to the server. Finally, the server updates the global model parameters as the weighted average of the received local model parameters. This process is repeated for a number of iterations, and the global model

is updated with each iteration. Similarly to Eq. (3), the weighting coefficients are proportional to the size of the data samples of each data owner. The difference between weighted averaging and federated averaging approaches is that in the latter multiple SGD iterations are performed locally before sending the model differences to the server.

Practical FL systems face several challenges, most prominently i) robustness, (ii) privacy preservation, and (iii) communication-efficiency. Below, we briefly review these fundamental challenges related to our research. In previous sections, we introduced CeBa, CeZe, and FedAvg approaches. In what follows, we introduce the techniques we explor in this work, namely, robust aggregation (RoAg), secure aggregation (SeAg), clipping with the quantile estimator (ClQu), zeroing with adaptive quantile estimator (ZeQu), Gaussian differentially private federated averaging with adaptive quantile clipping (GDP-AQuCl), and lossy compression (LoCo).

2.3.2. Robustness in fl

The aggregation of the updates from the participating centers in the training phase significantly impacts the learned model's performance. It is desirable to reduce the model's sensitivity to corrupted updates caused by a failure in hardware or manipulated by potential adversaries. Robustness in FL refers to the ability of the learning system to perform well despite the various challenges, such as malicious attacks, non-i.i.d. data, and communication constraints. Robustness is also important in FL because it helps to protect the system against malicious attacks. In FL, the training data is distributed among multiple nodes, and each node trains its own model using the data it has locally. This can make the system vulnerable to attacks in which malicious nodes try to manipulate the training data or the model parameters in order to cause the global model to perform poorly. Several techniques can be used to improve the robustness of FL systems, including:

- **Robust Aggregation:** Robust aggregation is a variant of federated averaging that is designed to be more resistant to malicious attacks.
- **Federated Transfer Learning**: Federated transfer learning is a technique that involves pre-training a model on a centralized dataset and then fine-tuning the model on decentralized data from multiple nodes. This can help improving the performance of the model in non-iid settings.
- **Outlier Detection and Removal:** Outlier detection and removal is a technique that involves identifying and removing data points that are significantly different from the majority of the data in order to improve the performance and robustness of the model.
- **Data Perturbation:** Data perturbation is a technique that involves adding noise to the training data at each node in order to protect the privacy of the data and improve the robustness of the model.

The standard aggregation scheme in FL, i.e., *arithmetic mean* aggregation, is not robust to data corruption. One possible solution is to use an approximate *geometric median* instead of the weighted arithmetic mean to increase the robustness to update corruption [37]. Alternative popular solutions include *zeroing* and *clipping* techniques. Zeroing (Ze) refers to replacing the components larger than a predefined threshold with zeros. The main objective of the zeroing approach is to increase the robustness of the whole learning model towards data corruption by faulty clients. The most popular zeroing approach in the literature is adaptive zeroing with the quantile estimator. In the clipping (Cl) approach [38], we bound the L_2 norm of client updates by projecting larger updates onto the L_2 ball of radius C centered at the origin. The clipping function Clip : $\mathbb{R}^d \times \mathbb{R} \rightarrow \mathbb{R}^d$ is defined as follows:

$$\text{Clip}(\boldsymbol{\theta}, C) = \boldsymbol{\theta} \Big/ \max\left(1, \frac{\|\boldsymbol{\theta}\|_2}{C}\right). \tag{4}$$

The hyper-parameter C has a significant role in the utility of the DL algorithm. If C is set too high, it entails the addition of more noise. If C is set too small, it can cause high bias in the gradient estimation since we lose the information on the magnitude of the original gradient, which may cause non-accurate training and worse generalization performance.

2.3.3. Privacy preservation in FL

Privacy preservation in FL refers to the ability of the learning model to protect the privacy of the training data while still allowing for effective model training. In FL, since the data centers (owners) avoid transmitting their local data to an external party, it was initially promoted as a private distributed learning algorithm. However, it has been shown that participant's training data may leak via the communicated model updates or the final shared model ([39,40]). To avoid information leakage, the typical solution is to use secure aggregating methods [41] such as *homomorphic encryption*, and *differential privacy (DP)* mechanisms in FL. Furthermore, the aggregation schemes based on averaging are vulnerable to adversarial attacks, e.g., a malicious participant may impose undesired behavior into the global model. Robust aggregation approaches try to address model integrity attacks ([37,42]). The two popular aggregation approaches are *federated averaging* [36] and *secure aggregation* [43]. In this research, we compare our results considering both of these approaches.

- **Secure Aggregation:** Secure aggregation (SeAg) is a method for aggregating models in FL designed to protect the privacy of the training data. In secure aggregation, each node trains its own model using local data and then sends encrypted model parameters to a central server. The server uses a secure aggregation protocol to combine the encrypted model parameters from the nodes and produce a global model. This global model is then sent back to the nodes for further training. Although SeAg is primarily aimed to protect the privacy of the training data, it can also improve the robustness of FL model. The goal of secure aggregation is to prevent the server from observing the individual local updates while being able to compute their aggregate. It also protects the final model from a possible integrity attack. It is mainly inspired by *secure multi-party computation (SMC)* protocols ([44,45]). In the secure aggregation approach, each participant masks its local model update using pairwise random keys and sends it to the parameter server. Two scenarios can be considered for the parameter server: (i) honest-but-curious (passive) model, and (ii) active adversary model. In our experiments, we consider the former. In [44], the authors addressed two masking schemes: (i) masking with one-time-pads and (ii) double-masking approaches. The masking approach with a one-time pad has two shortcomings: (i) it requires quadratic communication overhead, and (ii) there is no tolerance for a participant (data owner) failing to complete the protocol [43]. In our experiments, we use the double-masking approach as described in the following. Let each pair of data owners k, $k' \in \{1, \cdots, K\}$, $k \neq k'$, agree on some random seed (vector) $\mathbf{s}^t_{k,k'}$ at global iteration t. The pairwise random key $\mathbf{s}^t_{k,k'}$ can be generated using a key exchange protocol [46]. In addition, simultaneously, each data center $k \in \{1, \cdots, K\}$ samples (generates) a random seed $\mathbf{s}_k$. Next, the data owner k computes a masked version of its local model parameters as follows:

$$\mathbf{w}^t_k = \boldsymbol{\theta}^t_k + \text{PRG}(\mathbf{s}_k) + \sum_{k:k<k'} \text{PRG}\left(\mathbf{s}^t_{k,k'}\right) - \sum_{k:k>k'} \text{PRG}\left(\mathbf{s}^t_{k,k'}\right), \tag{5}$$

Table 2
Summary of the Different FL Algorithm Properties.

Algorithm	Overcome Client Drifting	Adaptive Learning rate	Cross-device Compatible	Robust to Outliers	Communication Efficient	Address Client Heterogeneity	Ensure Privacy
ClQu	×	✓	✓	✓	✓	✓	×
ZeQu	×	✓	✓	✓	✓	✓	×
FedAvg	×	✓	✓	×	×	×	×
LoCo	×	✓	✓	✓	✓	✓	×
RoAg	✓	✓	✓	✓	✓	✓	×
SeAg	×	✓	✓	✓	✓	×	✓
GDP-AQuCl	×	✓	✓	✓	✓	✓	✓

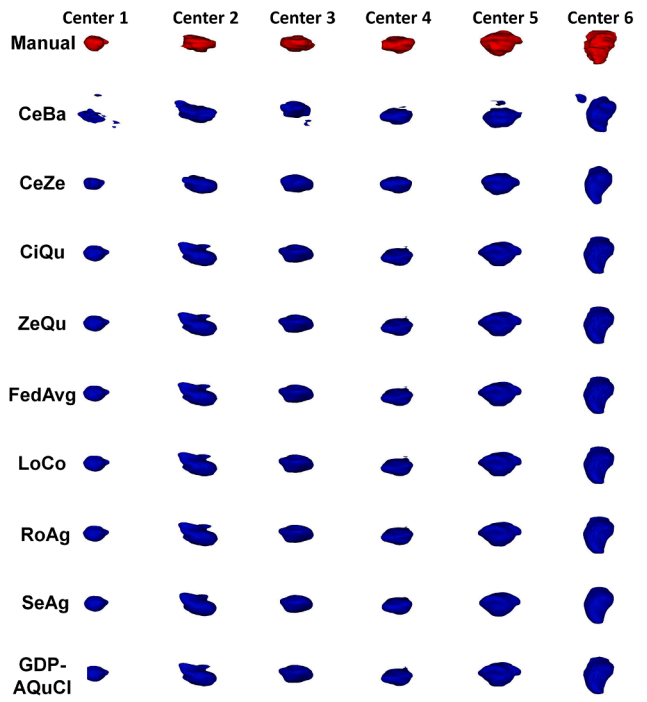

Fig. 2. 3D views of PET/CT segmentation obtained from manual (red) and different algorithms on representative patients from different centers.(1–6 from left to right): center-based (CeBa), centralized (CeZe), clipping with the quantile estimator (ClQu), zeroing with the quantile estimator (ZeQu), federated averaging (FedAvg), lossy compression(LoCo), robust aggregation (RoAg), secure aggregation (SeAg), Gaussian differentially private federated averaging with adaptive quantile clipping (GDP-AQuCl).

where PRG is a secure pseudo-random generator whose output space is $[0, R)^d$. Finally, the data owner sends its masked model parameters to the server. The data owner k uses the Shamir's $\frac{N}{2}$-out-of-N secret sharing protocol [47] to share $\{\mathbf{s}^t_{k,k'}\}$ and $\mathbf{s}_k$ with other data owners. Note that the operations in [5] are carried out in a finite field of integers[1] modulo a prime R, where [0, R) denotes the range of both model parameters and their summation.

- Homomorphic Encryption: In the cryptographic methods, which are mostly based on homomorphic encryption, the data owner sends an encrypted version of the data to the parameter server, and the signal processing is performed in the encrypted domain. Homomorphic encryption was initially introduced under the notion of privacy homomorphism in 1978 [48]. Since this seminal work, several homomorphic encryption techniques have been proposed, which are only able to process the encrypted data with one kind of operator, e.g., multiplication or addition operations, for a limited number of times [49]. The first fully homomorphic encryption (FHE) scheme has been proposed by Gentry [50], which allows an unlimited number of arithmetic operations in the encrypted domain. Recently, researchers try to use a Homomorphic encryption scheme in machine learning models, with possible application in medicine and biometrics ([51, 52]). We can use homomorphic encryption for the aggregation stage of an FL system, as it involves only the addition operation. Alternatively, one can use homomorphic encryption to train the local models in an encrypted domain using FHE. The study and analysis of privacy homomorphism are beyond the scope of this paper.
- Differential Privacy (DP) DP is the most popular context-free notion of privacy, which is inspired by the stability of likelihood ratios ([53, 54]). DP adds noise to the model parameters during training and aggregation in order to protect the privacy of the training data. It is widely used in deep learning models ([38,55-58]). Informally, a randomized computation over a database $\mathscr{D}$ is differentially private if the sensitive data of individuals contributing to $\mathscr{D}$ is protected against arbitrary adversaries with query access to $\mathscr{D}$ [59]. Although DP is primarily designed to protect the privacy of the training data, it can also improve the robustness of FL model.

Definition 1. *Let $\varepsilon \geq 0$ and $0 \leq \delta \leq 1$; a randomized algorithm $\mathscr{M}$ is said to be* (ε, δ)*-differentially private* [59] *if for any two neighbouring inputs (datasets) $\mathscr{D}_1$ and $\mathscr{D}_2$ and for every event $E \subseteq \mathbb{R}$, its output distributions are* (e^{ε},δ)*-close,* i.e.*, for every event E:*

$$\Pr[\mathscr{M}(\mathscr{D}_1) : E] \leq e^{\varepsilon} \Pr[\mathscr{M}(\mathscr{D}_2) : E] + \delta, \tag{6}$$

where $\Pr[\mathscr{M}(\mathscr{D}_1) : E]$ denotes the probability of event E in the distribution obtained by running the algorithm $\mathscr{M}$ on dataset $\mathscr{D}_1$, ε is the privacy budget, and δ denotes the probability of information leakage. The $\delta = 0$ refers to pure DP, while $\delta > 0$ refers to approximate DP. When $\delta = 0$, the (ε, δ)-DP mechanism $\mathscr{M}$ relaxed to ε-DP mechanism.

The intuition behind the definition of DP is that an individual has little incentive to participate in a statistical study, as the individual's data has limited effect on the outcome [60]. The Laplace and Gaussian noise mechanisms are the two most widely used practical mechanisms to achieve DP.

Let $f \in \mathscr{D} \rightarrow \mathbb{R}^d$ be function with L_2-sensitivity $\psi_f \triangleq \max_{\mathscr{D}_1, \mathscr{D}_2 \in \mathscr{D}, \mathscr{D}_1 \sim \mathscr{D}_2} \| f(\mathscr{D}_1) - f(\mathscr{D}_2) \|_2$, where $\mathscr{D}_1 \sim \mathscr{D}_2$ denotes that $\mathscr{D}_1$ and $\mathscr{D}_2$ are two neighbouring data sets. The Gaussian noise mechanism is defined as follows:

$$\mathscr{M}(\mathscr{D}) \triangleq f(\mathscr{D}) + \mathscr{N}\left(0, \sigma^2 \psi_f^2 \cdot \mathbf{I}_d\right), \tag{7}$$

where $\mathscr{N}(0, \sigma^2 \psi_f^2 \cdot \mathbf{I}_d)$ is a zero-mean multivariate Gaussian noise vector. Using the Gaussian mechanism, each data owner adds Gaussian noise to its local model parameter before forwarding to the server. The parameter σ is chosen based on ψ_f^2 and δ [38]. Note that the clipping (Cl) approach, described in Section 2.3.2, also bounds the L_2 sensitivity of

[1] The model parameters can be mapped to integers on the range [0, R) using a linear transform followed by a non-linearity, e.g., clipping and quantization.

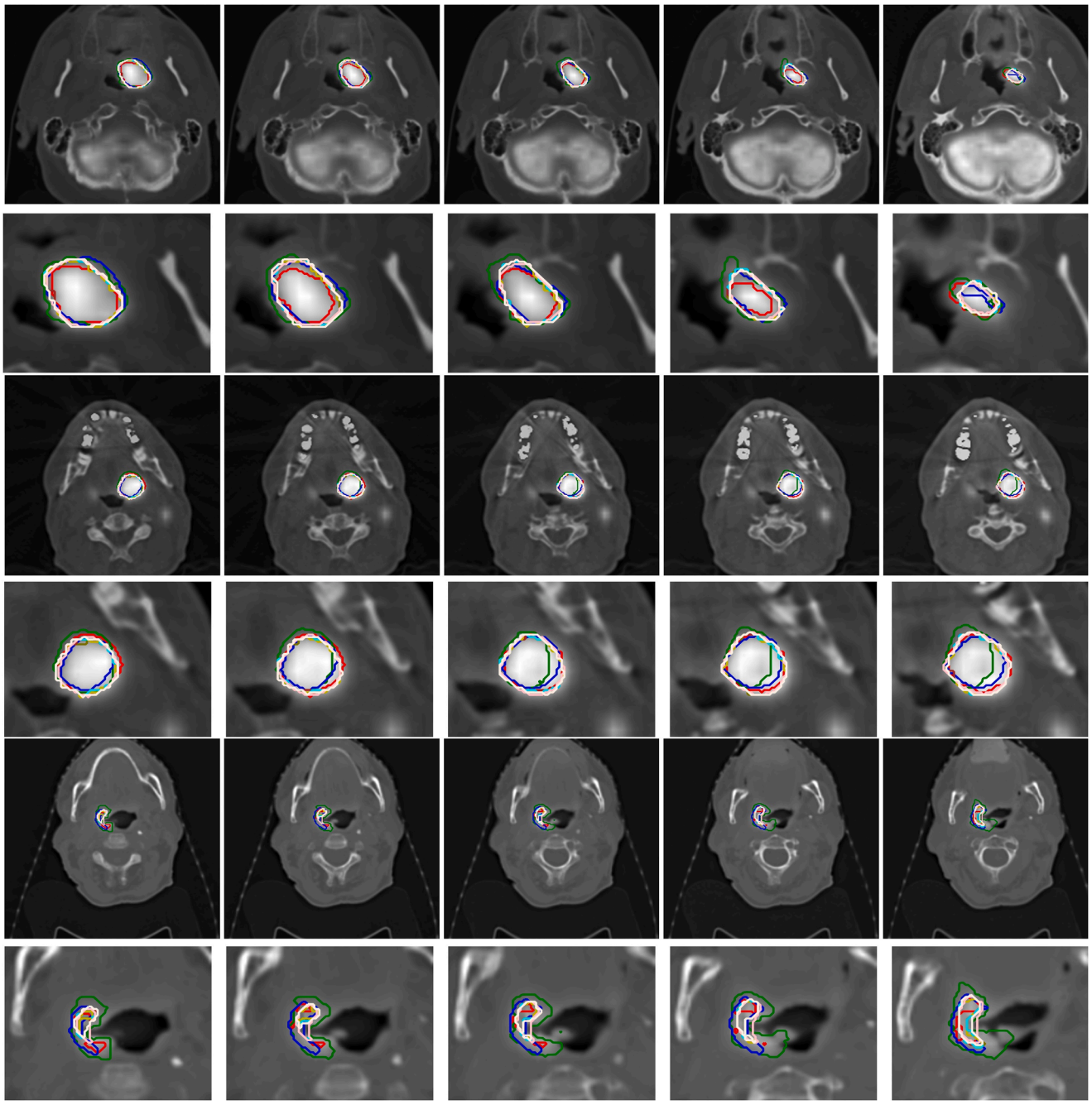

Fig. 3. 2D views of PET/CT segmentations obtained manually in three different cases: Red. CeBa: Green, CeZe: blue, ClQu: Brown, ZeQu: Olive, FedAvg: Orange, LoCo: Cyan, RoAg: Pink, SeAg: Linen, GDP-AQuCl: Yellow.

the model parameter aggregate with respect to the removal or addition of data samples of one participant (data owner). Therefore, we can add Gaussian noise to the clipped model parameters to obtain a central DP guarantee. Gaussian noise can be added (i) during local training, (ii) to the aggregated local model parameters before forwarding to the server, or (iii) to the global model parameter at the server side before sharing with the participants. A combination of DP and secure aggregation is employed for medical image FL in [41].

In [61], the authors proposed a private adaptive strategy for tuning the clipping threshold C to approximate it at a specified quantile of the update norm distribution, which can be viewed as minimizing the clipping probability. In this research, we utilize the *Gaussian differentially private federated averaging with adaptive quantile clipping* approach [61], which we refer to it as GDP-AQuCl. Moreover, we applied the proposed quantile scheme to the fixed zeroing and fixed clipping approaches described in Section 2.3.2, which results in (i) *zeroing with adaptive quantile estimator* (ZeQu), and (ii) *clipping with the quantile estimator* (ClQu) approaches. We compare all these SOTA approaches in our experiments.

2.3.4. Communication efficiency in FL

Communication efficiency in FL refers to the ability of the learning model to minimize the amount of communication required among the participating nodes in order to train the global model. In the literature, several strategies have been proposed [62–65] to optimize the communication-efficiency compared with the naïve SGD method. The communication between participants (data owners) and the parameter server is a fundamental stage for the FL frameworks. The proposed

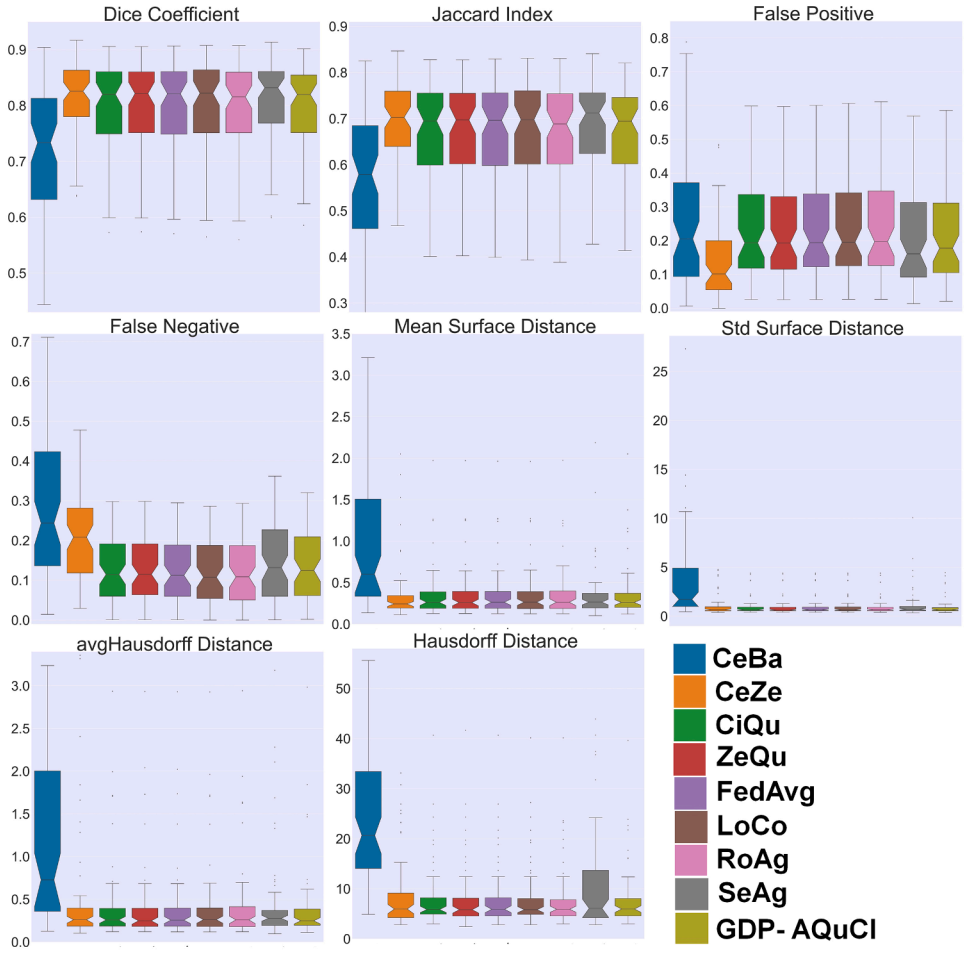

Fig. 4. Comparison of the performance of the different frameworks in terms of quantitative segmentation metrics of Dice similarity coefficient, Jaccard similarity coefficient, false-negative rate (1-Specificity), false-positive rate (1-Sensitivity), mean and standard deviation (SD) of surface distance as well as Hausdorff distance, average Hausdorff distance.

solutions in the literature to reduce the communication costs in FL are to reduce (i) the model parameter update size (including model compression and/or pruning and/or quantization and/or sparsification techniques), (ii) the number of participating data owners, and (iii) the total number of updates performed by each data owner. They are mainly based on lossy compression techniques, such as quantization and sparsification. Model compression is a technique that involves reducing the size of the model parameters in order to reduce the amount of data that needs to be transmitted during training and aggregation. Pruning is a technique that involves removing redundant or unnecessary connections from the model in order to reduce the size of the model and the amount of data that need to be transmitted. Quantization is a technique that involves representing the model parameters using a smaller number of bits to reduce the model size and the amount of data that need to be transmitted.

Lossy Compression (LoCo) Approach. A common solution to reduce the communication costs in the FL framework is to utilize lossy compression techniques on the global model sent from the server to participating parties ([66,67]). Lossy compression techniques are commonly studied through the rate-distortion theory framework. Shannon's work laid the groundwork for digital circuit design and made the current digital era possible. Since then, abundant research done on designing the lossy compression schemes [68–71]. It is worth mentioning that lossy compression meets privacy from the lens of information theory [72–79]. In this paper, we use simple probabilistic uniform quantization, which is parameterized by the number of quantization bits (q) and the compression threshold. For a vector $\boldsymbol{\theta} = [\theta_1,\cdots, \theta_d]^T$, we denote its minimum and maximum components by $\theta_{\min} = \min_j\{\theta_j\}_{j=1}^d$ and $\theta_{\max} = \max_j\{\theta_j\}_{j=1}^d$, respectively. For a probabilistic uniform binary (1-bit) quantization, one can replace every element θ_j by $\theta_{\max}$ with probability $\frac{\theta_j-\theta_{\max}}{\theta_{\max}-\theta_{\min}}$, and by $\theta_{\min}$ otherwise [80]. That is, the quantized value for each coordinate j is generated as follows:

$$Q(\theta_j) = \begin{cases} \theta_{\max} & , \text{ with probability } \dfrac{\theta_j - \theta_{\max}}{\theta_{\max} - \theta_{\min}} \\ \theta_{\min} & , \text{ otherwise} \end{cases} \tag{8}$$

Now, we can generalize the above stochastic 1-bit uniform quantization to stochastic q-bit uniform quantization. The process is based on equally dividing $[\theta_{\min},\theta_{\max}]$ into $k = 2^q$ intervals, and defining a new interval bounded by θ' and θ'' which plays a role of $\theta_{\min}$ and $\theta_{\max}$ in the above simple 1-bit uniform quantization method. More precisely, let us partition the interval $[\theta_{\min},\theta_{\max}]$ into sub-intervals $I_l \triangleq (B_k(l),B_k(l+1)]$, $l \in \{0, \cdots, k-1\}$, where $B_k(l)$ are given as:

$$B_k(l) \triangleq \theta_{\min} + l\,\frac{S}{k-1},\ \forall l \in \{0,\ldots,k-1\}, \tag{9}$$

where S satisfies $\theta_{\min} + S \geq \theta_{\max}$. Now we assign each coordinate of $\boldsymbol{\theta}$ into one of $B_k(l)$'s stochastically. To do this, for $\theta_j \in (B_k(l),B_k(l+1)]$ we quantize it as follows:

Table 3
Summary of Quantitative Image Segmentation Performance Metrics (Mean ± Sd and CI95%) for different algorithms.

	Model	Dice Score	Jaccard Coefficient	False Negative rate	False Positive rate	Mean Surface Distance	Std Surface Distance	Hausdorff Distance	avgHausdorff Distance
Mean ± Sd	**CeBa**	0.69 ± 0.17	0.55 ± 0.18	0.30 ± 0.21	0.25 ± 0.21	1.43 ± 3.23	3.62 ± 4.44	27.6 ± 23.3	2.49 ± 6.96
	CeZe	**0.80 ± 0.11**	**0.68 ± 0.13**	0.22 ± 0.15	**0.14 ± 0.12**	0.37 ± 0.37	1.04 ± 1.02	8.71 ± 7.30	0.48 ± 0.65
	ClQu	0.79 ± 0.10	0.66 ± 0.13	**0.14 ± 0.13**	0.23 ± 0.14	0.37 ± 0.32	0.96 ± 0.85	8.19 ± 6.46	0.41 ± 0.48
	ZeQu	0.79 ± 0.10	0.67 ± 0.13	**0.14 ± 0.13**	0.23 ± 0.14	0.36 ± 0.32	0.96 ± 0.86	8.14 ± 6.58	0.41 ± 0.49
	FedAvg	0.79 ± 0.10	0.66 ± 0.13	**0.14 ± 0.13**	0.23 ± 0.15	0.37 ± 0.32	0.96 ± 0.85	8.19 ± 6.48	0.41 ± 0.49
	LoCo	0.79 ± 0.11	0.67 ± 0.13	**0.14 ± 0.13**	0.24 ± 0.15	0.36 ± 0.32	0.96 ± 0.85	8.18 ± 6.46	0.41 ± 0.48
	RoAg	0.79 ± 0.11	0.66 ± 0.13	**0.14 ± 0.14**	0.24 ± 0.15	0.36 ± 0.31	0.94 ± 0.82	7.85 ± 6.16	0.41 ± 0.47
	SeAg	**0.80 ± 0.11**	0.67 ± 0.13	0.16 ± 0.15	0.20 ± 0.14	0.40 ± 0.52	1.18 ± 1.48	11.81 ± 15.51	0.53 ± 0.91
	GDP-AQuCl	0.79 ± 0.10	0.67 ± 0.12	0.15 ± 0.14	0.22 ± 0.14	**0.36 ± 0.30**	**0.93 ± 0.81**	**7.78 ± 5.98**	**0.40 ± 0.46**
CI95%	**CeBa**	0.65 to 0.74	0.51 to 0.60	0.25 to 0.35	0.20 to 0.30	0.66 to 2.20	2.57 to 4.67	22.06 to 33.13	0.83 to 4.14
	CeZe	0.77 to 0.82	0.64 to 0.71	0.19 to 0.26	0.12 to 0.17	0.28 to 0.46	0.80 to 1.28	6.97 to 10.44	0.32 to 0.63
	ClQu	0.77 to 0.82	0.64 to 0.69	0.11 to 0.17	0.20 to 0.27	0.29 to 0.44	0.76 to 1.16	6.65 to 9.72	0.30 to 0.53
	ZeQu	0.77 to 0.82	0.64 to 0.70	0.11 to 0.17	0.19 to 0.26	0.29 to 0.44	0.75 to 1.16	6.58 to 9.71	0.30 to 0.53
	FedAvg	0.77 to 0.82	0.63 to 0.69	0.11 to 0.17	0.20 to 0.27	0.29 to 0.44	0.76 to 1.16	6.65 to 9.73	0.30 to 0.53
	LoCo	0.77 to 0.82	0.64 to 0.70	0.10 to 0.17	0.20 to 0.27	0.29 to 0.44	0.76 to 1.16	6.65 to 9.72	0.30 to 0.53
	RoAg	0.76 to 0.81	0.63 to 0.69	0.10 to 0.17	0.20 to 0.27	0.29 to 0.44	0.75 to 1.14	6.39 to 9.31	0.30 to 0.52
	SeAg	0.77 to 0.82	0.64 to 0.70	0.13 to 0.20	0.17 to 0.23	0.28 to 0.52	0.83 to 1.53	8.12 to 15.5	0.31 to 0.75
	GDP-AQuCl	0.77 to 0.82	0.64 to 0.70	0.12 to 0.19	0.18 to 0.25	0.28 to 0.43	0.74 to 1.12	6.35 to 9.20	0.29 to 0.51

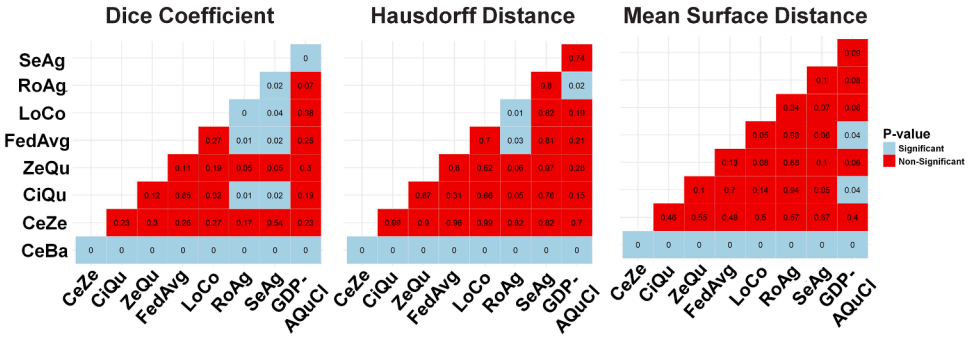

Fig. 5. Comparison of different models (*p*-values) in terms of different metrics of Dice coefficient, Hausdorff Distance, and Mean Surface Distance. Manual segmentation is used as the criterion standard.

$$Q(\theta_j) = \begin{cases} l+1 & , \text{ with probability } \dfrac{\theta_j - B_k(l)}{B_k(l+1) - B_k(l)} \\ l & , \text{ with probability } \dfrac{B_k(l+1) - \theta_j}{B_k(l+1) - B_k(l)} \end{cases}. \tag{10}$$

In our experiments, we set $q = 8$ and consider the natural choice $S = \theta_{max} - \theta_{min}$.

A comparison of the different utilized FL algorithm properties is summarized in Table 2.

2.4. Deep neural network transformers

In this study, we implemented a purely attention-based transformer without convolutions, inspired by work reported in [12,81], as a modified version of ([12,81]). This follows a very active line of research pioneered by [82], which is motivated by the significant success of the transformer structures [11] within the NLP domains and aims to bring the power of the self-attention mechanism of the transformers into the image-based and vision-based applications. The architecture consists of an encoder, a bottleneck block, a decoder, and skip-connections [12], and is based primarily on the Swin-transformer (**S**hifted **win**dows) block, which was originally proposed in [81]. The images are first split into non-overlapping blocks of dimension 4 × 4, followed by a linear projection to form the input sequences to the network. The encoder consists of patch-merging blocks for signal down-sampling, followed by Swin-transformer blocks responsible for representation learning. This forms a hierarchical representation, where, similar to the U-shaped structure of the U-Net, has a symmetric decoder layer that consists of

Table 4
Summary of Quantitative PET Metrics (Mean ± Sd and CI95%) for different algorithms.

	Method	SUV_{max}	SUV_{peak}	SUV_{mean}	SUV_{median}	TLG
Mean ±Sd	CeBa	1.28 ± 6.92	1.53 ± 10.98	5.14 ± 17.7	8.91 ± 23.19	3.70 ± 48.24
	CeZe	**0 ± 0**	**0 ± 0**	4.77 ± 10.38	6.21 ± 13.45	-3.09 ± 18.12
	ClQu	**0 ± 0**	**0 ± 0**	-2.39 ± 11.85	-3.15 ± 15.41	12.91 ± 22.93
	ZeQu	**0 ± 0**	**0 ± 0**	-2.2 ± 11.74	-2.90 ± 15.29	12.35 ± 22.65
	FedAvg	**0 ± 0**	**0 ± 0**	-2.58 ± 11.84	-3.38 ± 15.40	13.22 ± 23.03
	LoCo	**0 ± 0**	**0 ± 0**	-2.86 ± 11.87	-3.73 ± 15.44	13.7 ± 23.24
	RoAg	**0 ± 0**	**0 ± 0**	-2.86 ± 12.01	-3.87 ± 15.63	14.1 ± 23.52
	SeAg	**0 ± 0**	**0 ± 0**	1.74 ± 13.48	1.58 ± 16.88	**9.56 ± 23.94**
	GDP-AQuCl	**0 ± 0**	**0 ± 0**	**-0.87 ± 12.02**	**-1.24 ± 15.75**	9.79 ± 22.09
CI95%	CeBa	-0.36 to 2.93	0.94 to 9.35	-1.08 to 4.14	3.4 to 14.42	-7.77 to 15.16
	CeZe	0 to 0	0 to 0	2.3 to 7.24	3.01 to 9.41	-7.4 to 1.21
	ClQu	0 to 0	0 to 0	-5.2 to 0.43	-6.82 to 0.51	7.46 to 18.36
	ZeQu	0 to 0	0 to 0	-4.99 to 0.6	-6.54 to 0.73	6.97 to 17.73
	FedAvg	0 to 0	0 to 0	-5.39 to 0.23	-7.04 to 0.28	7.75 to 18.7
	LoCo	0 to 0	0 to 0	-5.68 to -0.04	-7.4 to -0.06	8.18 to 19.23
	RoAg	0 to 0	0 to 0	-5.72 to -0.01	-7.59 to -0.16	8.51 to 19.69
	SeAg	0 to 0	0 to 0	-1.46 to 4.95	-2.44 to 5.59	3.86 to 15.25
	GDP-AQuCl	0 to 0	0 to 0	-3.73 to 1.98	-4.99 to 2.50	4.54 to 15.04

Swin-transformer layers and patch-expander units. Between the encoder and the decoder, skip connections facilitate the signal flow. At the bottom of the encoder, a bottleneck consisting of two consecutive Swin-transformer blocks without up- or down-sampling provides a further connection between the encoder and the decoder.

As an alternative to the traditional sliding-window approach, the Swin-transformer block is based on the idea of shifted-windows [81]. A regular partitioning of the patches is used at one layer, while the next layer uses a shifted version of them. This provides connections between windows with different shapes using self-attention. The Swin-transformer block consists of a layer-norm (LN), the multihead self-attention (MSA), multi-layer perceptrons (MLP), and several skip-connections, such that:

$$\hat{z}^{l} = \text{W-MSA}\left(\text{LN}\left(z^{l-1}\right)\right) + z^{l-1}, \tag{11}$$

$$z^{l} = \text{MLP}\left(\text{LN}\left(\hat{z}^{l}\right)\right) + \hat{z}^{l}, \tag{12}$$

$$\hat{z}^{l+1} = \text{SW-MSA}\left(\text{LN}\left(z^{l}\right)\right) + z^{l}, \tag{13}$$

$$z^{l+1} = \text{MLP}\left(\text{LN}\left(\hat{z}^{l+1}\right)\right) + \hat{z}^{l+1}, \tag{14}$$

where, $\hat{z}^{l}$ and z^{l} represent the outputs of the (S)W-MSA, and the MLP module of the l^{th} block, respectively, and the self-attention mechanism of [11] is computed as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^{T}}{\sqrt{d}}\right)V, \tag{15}$$

where Q, K and $V \in \Re^{M^2 \times d}$ denote the *query*, *key* and *value* matrices, with M^2 representing the number of patches in a window, and d being the dimension of the query/key.

2.5. Training

We evaluated different frameworks (single-center based, centralized, and seven FL algorithms) by using 68 PET/CT images (20% of each center's local data). The training was performed on axial slices, as PET and CT images with batch size 32 were fed to models simultaneously (dual channel input). During each iteration, stratified mini-batch approaches are used, in which half of the batches with tumor segmentation and half without tumors were fed to the model to avoid bias during training. All DL models were implemented in the TensorFlow framework. FL algorithms were implemented by using TensorFlow Federated (TFF). TFF is an open-source framework developed for simulating and implementing different FL algorithms. All networks were trained in a 2D manner with an Adam optimization with a learning rate starting with 0.001, as well as a weight decay of 0.0001. Dice loss was used, and models trained using 300 epochs and 100 rounds in FL.

2.6. Quantitative evaluation

Different evaluation metrics, including standard segmentation quantitative metrics, image-derived PET metrics, and radiomics features, were considered to evaluate and compare the performances of different frameworks. Standard segmentation quantitative metrics, including the Dice similarity coefficient, Jaccard similarity coefficient, false-negative rate (1-Sensitivity), false-positive rate (1-Specificity), mean and standard deviation (SD) of surface distance (mm) as well as Hausdorff distance, average Hausdorff distance (mm) are considered. Image-derived PET metrics for clinical evaluation of the different frameworks, including variants of the standardized uptake value (SUV) SUV_{peak}, SUV_{mean}, SUV_{median}, SUV_{max}, metabolic tumor volume (MTV) and total lesion glycolysis (TLG, $MTV \times SUV_{mean}$) were also analyzed. For radiomics analysis, we extracted intensity, histogram, and shape radiomics features using SERA package [83]. All these metrics were calculated on test sets (20% of each center data).

2.7. Statistical analysis

Percent relative error (RE%) was calculated for PET image metrics with respect to manual segmentation. The Kolmogorov-Smirnov test was used for normal evaluation and then, based on the distribution paired Wilcoxon signed rank test was chosen for evaluation, and p-value < 0.05 was defined as the threshold for statistical significance. Comparison of the different models using various metrics (Dice coefficient, Hausdorff Distance, and Mean Surface Distance) was performed using paired Wilcoxon signed rank test. All p-values were corrected by Benjamini-Hochberg correction, and Intra-class correlation (ICC) ([84,85]) test was performed for radiomics feature reproducibility in different approaches with respect to manual segmentation; we classified radiomics features based on ICC value into four Groups of poor reproducibility (ICC < 0.40), fair reproducibility ($0.40 <$ ICC < 0.59), good reproducibility ($0.60 <$ ICC < 0.74), and excellent reproducibility ($0.75 <$ ICC < 1.00).

2.8. Code and data availability

All PET and CT images are available in The Cancer Imaging Archive ([3,22,26-32]). Different implementations would be available in the Authors GitHub repository.

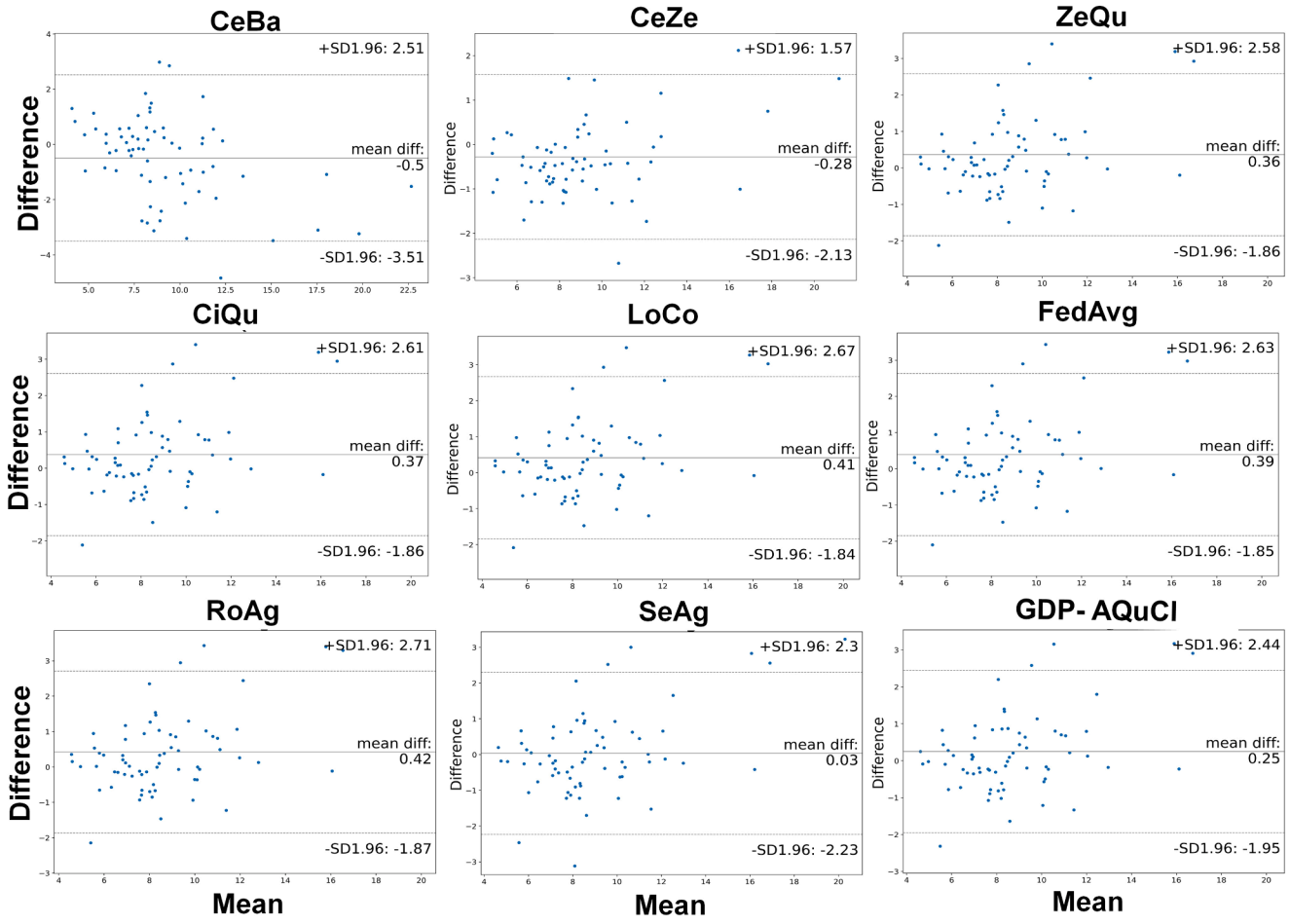

Fig. 6. Bland Altman plots of SUV_{mean} for different frameworks compared to manual segmentation.

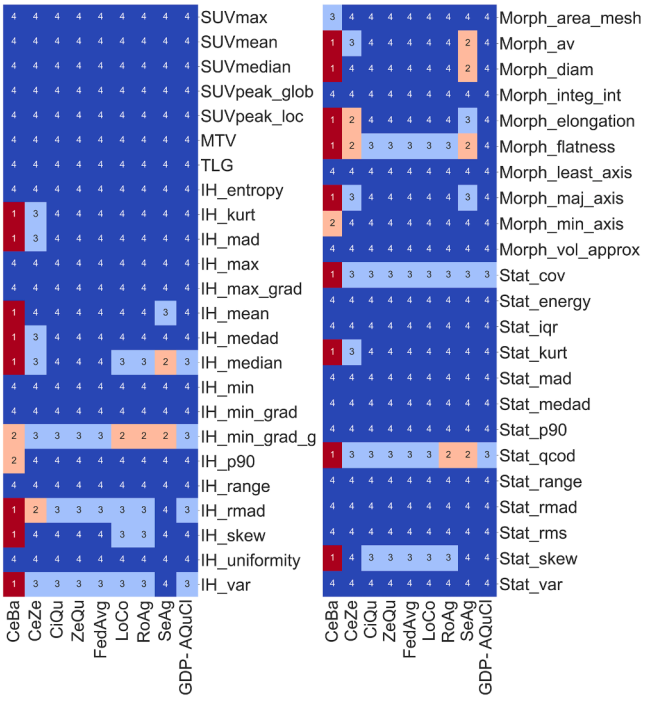

Fig. 7. ICC value of radiomic features for the different frameworks compared to manual segmentation. Less than 0.40-poor [1], Between 0.40 and 0.59-fair [2], Between 0.60 and 0.74-good [3], and Between 0.75 and 1.00-excellent [4].

3. Results

3.1. Segmentation description

Fig. 2 depicts, for visual comparison, examples of 3D-rendered volumes of segmentation of GTVs from six different clinical centers (columns) with manual segmentation (red) as well as single-center-based, centralized, and different FL approaches (blue). For a visual comparison of different approaches with respect to manual segmentation, Fig. 3 represents 2D axial views of different patients in both original and magnified versions of GTVs. As shown in these figures, segmentation provided by different FL approaches are in good agreement with centralized and manual segmentations in different textures and sizes of GTVs.

3.2. Quantitative segmentation metrics

Fig. 4, compares the performance of different models with different approaches and a summary of the results is also presented in Table 3. Centralized (CeZe) and SeAg models showed the best performance in terms of the Dice coefficient (0.80±0.11 versus 0.80±0.11), without any significant difference between the two (p-value> 0.05). In terms of the false negative rate, ClQu, ZeQu, FedAvg, LoCo and RoAg achieved the lowest values of 0.14±0.13 (CI95%: 0.11 to 0.17); however, CeZe showed the lowest false positive rate of 0.14±0.12 (CI95%: 0.12 to 0.17). In terms of Hausdorff distances GDP-AQuCl method achieved the lowest value of 7.78±5.98 mm (CI95%: 6.35 to 9.2 mm) followed by RoAg with the value 7.85±6.16 mm (CI95%: 6.39 to 9.31 mm). In terms of the Hausdorff distance (7.78±5.98 mm), Mean Surface Distance (0.36

Table 5
Summary of Quantitative Image Segmentation Performance Metrics (Mean ± Sd and CI95%) for the Different Centres.

	Centre	Dice Score	Jaccard Coefficient	False Negative rate	False Positive rate	Mean Surface Distance	Std Surface Distance	Hausdorff Distance	Mean Hausdorff Distance
Mean ±Sd	**Center 1**	0.56 ± 0.23	0.42 ± 0.21	0.48 ± 0.27	0.26 ± 0.26	1.92 ± 2.91	4.85 ± 6.45	30.78 ± 31.78	3.78 ± 6.47
	Center 2	0.67 ± 0.19	0.53 ± 0.19	0.26 ± 0.22	0.32 ± 0.24	2.12 ± 4.07	5.42 ± 7.99	32.13 ± 34.6	3.81 ± 9.36
	Center 3	0.63 ± 0.19	0.49 ± 0.18	0.3 ± 0.19	0.34 ± 0.23	2.11 ± 3.5	5.27 ± 6.41	37.6 ± 27.5	3.03 ± 5.39
	Center 4	0.72 ± 0.18	0.58 ± 0.18	0.26 ± 0.21	0.24 ± 0.2	1.23 ± 3.18	3.2 ± 4.05	29.63 ± 20.39	2.04 ± 6.85
	Center 5	0.62 ± 0.21	0.47 ± 0.19	0.43 ± 0.25	0.22 ± 0.23	1.29 ± 2.14	3.06 ± 4.24	21.27 ± 19.06	2.55 ± 6.56
	Center 6	0.7 ± 0.17	0.56 ± 0.18	0.3 ± 0.19	0.23 ± 0.19	1.2 ± 1.55	3.69 ± 4.34	31.84 ± 28.1	1.81 ± 2.76
	CeBa	0.69 ± 0.17	0.55 ± 0.18	0.3 ± 0.21	0.25 ± 0.21	1.43 ± 3.23	3.62 ± 4.44	27.6 ± 23.3	2.49 ± 6.96
CI95%	**Center 1**	0.5 to 0.61	0.37 to 0.47	0.41 to 0.54	0.2 to 0.32	1.23 to 2.61	3.32 to 6.38	23.23 to 38.33	2.24 to 5.31
	Center 2	0.62 to 0.71	0.48 to 0.57	0.21 to 0.31	0.26 to 0.38	1.16 to 3.09	3.52 to 7.32	23.91 to 40.36	1.58 to 6.03
	Center 3	0.59 to 0.68	0.44 to 0.53	0.26 to 0.35	0.29 to 0.4	1.28 to 2.94	3.75 to 6.79	31.06 to 44.14	1.75 to 4.31
	Center 4	0.67 to 0.76	0.54 to 0.62	0.21 to 0.31	0.19 to 0.29	0.47 to 1.98	2.23 to 4.16	24.78 to 34.47	0.41 to 3.67
	Center 5	0.57 to 0.67	0.43 to 0.52	0.37 to 0.49	0.17 to 0.28	0.78 to 1.8	2.06 to 4.07	16.74 to 25.8	0.99 to 4.11
	Center 6	0.65 to 0.74	0.51 to 0.6	0.25 to 0.35	0.19 to 0.28	0.83 to 1.57	2.66 to 4.73	25.16 to 38.52	1.16 to 2.47
	CeBa	0.65 to 0.74	0.51 to 0.6	0.25 to 0.35	0.2 to 0.3	0.66 to 2.2	2.57 to 4.67	22.06 to 33.13	0.83 to 4.14

±0.30 mm), Std Surface Distance (0.93±0.81 mm), and average Hausdorff Distance (0.40±0.46 mm), GDP-AQuCl outperformed all federated, single-center, and centralized approaches. Statistical analysis showed significant differences (p-value$<$ 0.05) when comparing single center-based models with centralized and federated algorithms for different quantitative metrics. However, for almost all segmentation metrics, statistical tests showed no significant difference (p-value $>$ 0.05) between centralized and different FL approaches and also among different FL algorithms. Fig. 5 presents a comparison of different models (p-values) in terms of three metrics.

3.3. PET quantitative metrics ICC and reproducibility

Results of PET quantitative metrics in terms of RE%, are presented in Table 4 for different approaches. SUV_{max} and SUV_{peak} values of all federated and centralized approaches achieved RE% of zero. However, for CeBa, RE% of 1.28±6.92 and 1.53±10.98 were achieved for SUV_{max} and SUV_{peak}, respectively. Lowest RE% of SUV_{mean} (−0.87±12.02) and SUV_{median} (−1.24±15.75) was achieved by GDP-AQuCl, whereas the lowest RE% of TLG (9.79±22.09) was achieved by SeAg. All quantitative PET metrics showed excellent repeatability in terms of the ICC analysis (ICC$>$ 0.75). Fig. 6 depicts the Bland Altman figure of SUV_{mean} for different approaches, which was computed with respect to manual segmentation, and shows good agreement between different approaches and manual segmentation. Fig. 7 represents the ICC value of different radiomic features of different algorithms with respect to manual segmentation, and as shown in this figure most features showed excellent repeatability (ICC$>$ 0.75). The CeBa approach had fewer repeatable features (ICC$>$ 0.75).

3.4. Center-based analysis

In addition to centralized and FL-based frameworks, we also analyzed training and testing with each center's data-set separately, and the quantitative metrics for these single-center approaches are presented in Table 5. In CeBa analysis, results of training and testing of data using the same center (for training and testing) were presented. Training on one center and testing on other sets (datasets from centers not presented in training) showed low generalizability for different centers (mean Dice score of 0.56 to 0.72). In Table 6, we present the quantitative PET metrics for different centers, and as seen in the table, all metrics showed high variability across different centers. In Fig. 8, we present 2D axial views of different patients in both original and magnified versions of GTVs by training on different centers' datasets. This figure shows low accuracy of segmentation when training on a single center's dataset and testing on another center's.

4. Discussion

PET/CT image segmentation is a crucial step toward quantitative analysis in monitoring treatment response and radiation therapy. However, it suffers from a number of challenges due to inherent limitations in image quality and the high variability in HN regions. Inter-observer variability with an average Dice score of 0.57 in CT, 0.61 in PET/CT, and 0.69 in PET/CT was reported in previous studies between different human observers ([7,86,87]). Various DL algorithms have been developed to address these challenges by automating the segmentation process. Centralized training on data pooled from multiple centers is ideal for building generalizable models. However, this approach faces privacy, security, legal, ethical, and ownership challenges. These challenges could be addressed by a shared global model using FL. In the current study, we evaluated the performance of different decentralized FL frameworks for multi-institutional PET/CT image co-segmentation. The HECKTOR challenge was organized to address HN tumor segmentation using PET/CT images. Since we used different datasets from those used in the HECKTOR challenge, our results are not directly comparable. However, compared to models proposed in HECKTOR, our proposed centralized (CeZe) and FL (SeAg) models performed better than the winner of the challenge (in terms of Dice score 0.80±0.11 vs. 0.78) by taking advantage of the transformer architecture.

We implemented seven different FL algorithms in the current study and compared their performance with centralized and single-center-

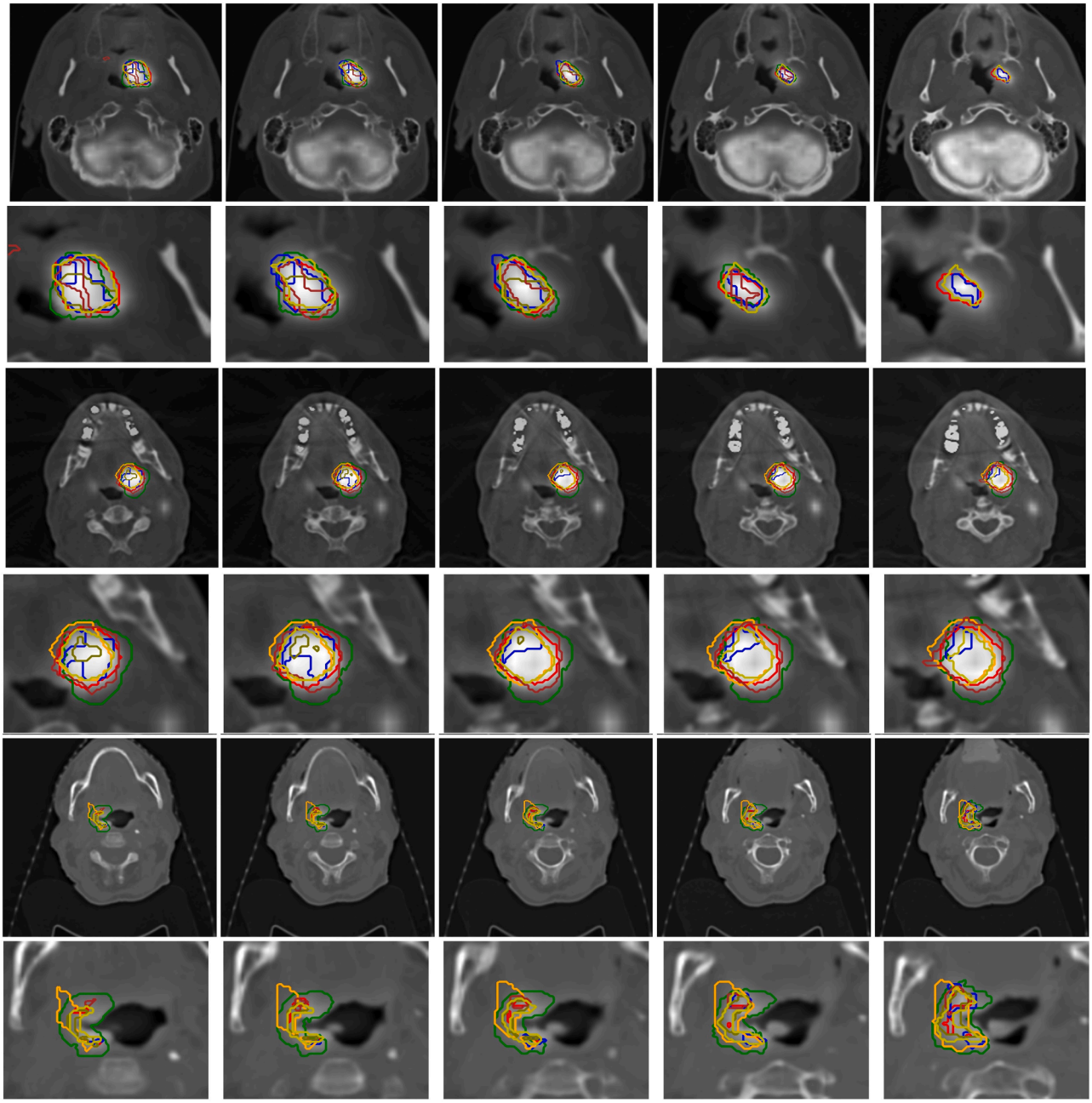

Fig. 8. 2D views of PET/CT segmentation obtained in three different cases: from the manual: Red. Center 1: Green, Center 2: Blue, Center 3: Brown, Center 4: Olive, Center 5: Orange, Center 6: Cyan.

based approaches for HN GTV segmentation from PET/CT images using vision transformers. All FL approaches achieved centralized learning model performance with no statistically significant difference. Among FL algorithms, SeAg and GDP-AQuCl outperformed other FL algorithms considering different quantitative metrics. However, there were no statistically significant differences between these FL algorithms. Conversely, single-center-based models showed low accuracy and generalizability. From all segmentation frameworks, GDP-AQuCl produced the highest number of reproducible radiomic features. Among the plausible reasons would be the lowest value of surface distance and Hausdorff distance compared to other frameworks and concurrently the same value of the Dice score. We conclude that collaboration between different centers is highly crucial for generalizable DL model development. Notwithstanding the variability in PET scanner models, image acquisition and reconstruction protocols, and different sizes of datasets across different centers, all FL approaches achieved centralized learning model performance with no statistically significant difference.

FL algorithms have some inherent challenges in medical imaging, including data partitioning, data distribution, privacy, and security, as well as communication and computation capabilities of the infrastructure. Choosing the right data partitioning is an important step toward addressing limited sample or feature sizes, or both, resulting in horizontal FL (HFL), vertical FL (VFL), or federated transfer learning (FTL), respectively ([13,14]). In the current study, we implemented horizontal FL algorithms, where there is no overlap between data from different centers. However, we used both PET and CT images for DL models. The second issue in FL is data distribution, which is a statistical data heterogeneity challenge due to the decentralized nature of datasets as each center generates their local data. As the data is decentralized, the distribution of data across each center could be significantly different,

Table 6
Summary of Quantitative PET Metrics (Mean ± Sd and and CI95%) for different training sets by different centres.

	Center	SUV_{max}	SUV_{peak}	SUV_{mean}	SUV_{median}	TLG
Mean ± Sd	**Center 1**	0.73 ± 8.97	0.89 ± 13.33	15.7 ± 27.19	22.05 ± 34.97	-10.2 ± 61.33
	Center 2	-0.15 ± 14.1	0.21 ± 16.51	2.8 ± 21.83	5.2 ± 26.6	37.31 ± 117.2
	Center 3	1.31 ± 7.22	1.18 ± 11.43	-3.16 ± 17.82	-0.52 ± 23.02	30.24 ± 126.2
	Center 4	1.53 ± 7.15	1.34 ± 10.58	4.27 ± 16.37	7.8 ± 23.87	4.42 ± 45.2
	Center 5	-1.71 ± 18.57	-1.84 ± 20.12	8.69 ± 29.83	14.76 ± 36.61	-18.43 ± 38.56
	Center 6	-0.08 ± 14.24	-0.38 ± 16.24	4.09 ± 19.55	6.85 ± 23.45	3.37 ± 48.59
	CeBa	1.28 ± 6.92	1.53 ± 10.98	5.14 ± 17.7	8.91 ± 23.19	3.7 ± 48.24
CI95%	**Center1**	-1.4 to 2.86	9.24 to 22.16	-2.28 to 4.06	13.73 to 30.36	-24.78 to 4.38
	Center2	-3.5 to 3.2	-2.39 to 7.99	-3.71 to 4.13	-1.12 to 11.53	9.45 to 65.17
	Center3	-0.4 to 3.03	-7.4 to 1.07	-1.53 to 3.9	-5.99 to 4.95	0.24 to 60.24
	Center4	-0.17 to 3.23	0.38 to 8.16	-1.18 to 3.85	2.13 to 13.47	-6.32 to 15.16
	Center5	-6.12 to 2.71	1.6 to 15.78	-6.62 to 2.94	6.06 to 23.46	-27.6 to -9.26
	Center6	-3.46 to 3.31	-0.56 to 8.73	-4.24 to 3.48	1.28 to 12.42	-8.18 to 14.92
	CeBa	-0.36 to 2.93	0.94 to 9.35	-1.08 to 4.14	3.4 to 14.42	-7.77 to 15.16

which is known as non-independent and identically distributed (non--IID) data. Different centers equipped with different scanner models and using different image acquisition and processing protocols, employing different segmentation techniques, may result in non-IID data in medical imaging ([13,14]). To address heterogeneity in our data, we employed automated pre-processing, including cropping, metal artifact reduction, and resizing to isotropic voxels. In addition, as the sample size of each center is different (from 23 to 99), we used stratified mini-batch approaches during each iteration (where half of the batches consisted of tumor segmentation while the other half did not include tumors), to avoid biased training. In our study, all FL approaches achieved centralized method performance in PET/CT image segmentation and outperformed single-center-based approaches. Another issue in FL is privacy and security, as the number of centers could potentially be increased to hundreds and even thousands, in which case all centers cannot be considered trustable parties. Different kinds of attacks, including membership inference and model inversion attacks, could be performed by curious parties to discover whether a specific data sample exists within the training set of other centers, or to regenerate training sets from the trained model during model training, respectively ([13, 14]). These attacks result in the leaking of sensitive information about patients during decentralized training, which can be a serious concern impeding the adoption of FL techniques in large-scale medical applications. Different methods, such as data perturbation or encryption, can be implemented for data privacy and security purposes. Controlled random noise can be added to samples during training to guarantee DP ([13, 14]). Additionally, encryption can be used during the aggregation process to preserve privacy. Membership and model inversion attacks can be addressed by the DP mechanism ([13,14]). Other attacks, including data and model poisoning (i.e. adversarial attacks), can be performed by malicious parties. We implemented DP as well as secure FL approaches and showed that they both achieved centralized FL performance in PET/CT image segmentation while preserving patient privacy and security against potential attacks.

In our study, an experienced nuclear medicine physician evaluated and checked all PET/CT segmentations and edited/modified them to offset plausible errors, which is unrealistic in a real-world FL setup. However, it was necessary for our study as the dataset was gathered from an online dataset that contained a few errors that had to be mitigated before building models. In real FL, images and segmentations should be checked and modified in case of errors; otherwise, this could be an issue for the training process. Another challenge in FL is the statistical variation resulting from image pre-processing at different centers ([13,14]). As PET/CT images are in DICOM format, in real-world scenarios, image pre-processing could be shared with the client to provide pre-processed images with the same setting across the different centers. We implemented fully automated pre-processing steps in the current study toward reproducing data preparation. One of the limitations of the current study is that all the analysis has been performed in one server with multiple GPUs treated as different centers; thus, good communication between centers and the parameter server is assumed. To address the communication limitations in practice, quantized model transmission is considered, and it is observed that the centralized benchmark performance can be achieved while reducing the communication load significantly. Further studies should consider practical communication bottlenecks for real clinical applications. In the current study, we used a limited number of data and clients for model development to demonstrate the effectiveness of FL for tumor segmentation to achieve the performance achieved by the centralized level model. However, further studies need to be conducted using more data and clients to prove the effectiveness of FL algorithms for FL segmentation models.

5. Conclusion

FL-based algorithms have proven to be highly effective for HN tumor segmentation in PET/CT images, achieving performance on par with centralized deep learning models. These algorithms enable the training of generalizable PET/CT image segmentation models by providing access to large, diverse datasets from multiple centers without compromising patient privacy or security. This decentralized approach to model training allows for the creation of more robust and accurate models, particularly in situations where communication among centers is limited. The use of FL-based algorithms represents a novel and innovative approach to HN tumor segmentation in PET/CT images.

Declaration of Competing Interest

None

Acknowledgements

Ethics approval and consent to participate were unnecessary since the study was performed on open access online dataset. This work was supported by the Swiss National Science Foundation under grant SNSF 320030–176052, the Private Foundation of Geneva University Hospitals under Grant RC-06–01, the German Research Foundation (DFG, Project IDs 322900939, 454024652, 432698239 & 445703531) and the European Research Council under Grant ERC-101001791. 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.

References

[1] E.M. Rohren, T.G. Turkington, R.E Coleman, Clinical applications of PET in oncology, Radiology 231 (2004) 305–332.
[2] B. Foster, U. Bagci, A. Mansoor, Z. Xu, D.J. Mollura, A review on segmentation of positron emission tomography images, Comput. Biol. Med. 50 (2014) 76–96.
[3] I. Shiri, H. Arabi, A. Sanaat, E. Jenabi, M. Becker, H. Zaidi, Fully automated gross tumor volume delineation from PET in head and neck cancer using deep learning algorithms, Clin. Nucl. Med. 46 (11) (2021) 872–883.
[4] editorsV Andrearczyk, V Oreiller, M Vallires, J Castelli, H Elhalawani, M Jreige, et al. (Eds.), Automatic Segmentation of Head and Neck Tumors and Nodal Metastases in PET-CT Scans, Medical imaging with deep learning, 2020.

[5] L. Zhao, Z. Lu, J. Jiang, Y. Zhou, Y. Wu, Q. Feng, Automatic nasopharyngeal carcinoma segmentation using fully convolutional networks with auxiliary paths on dual-modality PET-CT images, J. Digit. Imaging 32 (3) (2019) 462–470.

[6] Z. Guo, N. Guo, K. Gong, Q. Li, Gross tumor volume segmentation for head and neck cancer radiotherapy using deep dense multi-modality network, Physics in Medicine \& Biology. 64 (20) (2019), 205015.

[7] V. Oreiller, V. Andrearczyk, M. Jreige, S. Boughdad, H. Elhalawani, J. Castelli, et al., Head and neck tumor segmentation in PET/CT: the HECKTOR challenge, Med. Image Anal. 77 (2022), 102336.

[8] V. Andrearczyk, V. Oreiller, M. Jreige, M. Vallires, J. Castelli, H. Elhalawani, et al., Overview of the HECKTOR Challenge At MICCAI 2020: Automatic Head and Neck Tumor Segmentation in PET/CT. 3D Head and Neck Tumor Segmentation, editors, PET/CT Challenge, 2020.

[9] editorsJ Xie, Y Peng (Eds.), The Head and Neck Tumor Segmentation Based On 3D U-Net. 3D Head and Neck Tumor Seg in PET/CT Challenge, 2021.

[10] Chen J., Lu Y., Yu Q., Luo X., Adeli E., Wang Y., et al. Transunet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:210204306. 2021.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, et al., Attention is all you need, Adv. Neural Inf. Process. Syst. (2017) 30.

[12] Cao H., Wang Y., Chen J., Jiang D., Zhang X., Tian Q., et al. Swin-unet: unet-like pure transformer for medical image segmentation. arXiv preprint arXiv: 210505537. 2021.

[13] K.M.J. Rahman, F. Ahmed, N. Akhter, M. Hasan, R. Amin, K.E. Aziz, et al., Challenges, Applications and Design Aspects of Federated Learning: A survey, IEEE Access, 2021.

[14] C.-R. Shyu, K.T. Putra, H.-C. Chen, Y.-Y. Tsai, K.S.M.T. Hossain, W. Jiang, et al., A systematic review of federated learning in the healthcare area: from the perspective of data properties and applications, Appl. Sci. 11 (23) (2021) 11191.

[15] I. Dayan, H.R. Roth, A. Zhong, A. Harouni, A. Gentili, A.Z. Abidin, et al., Federated learning for predicting clinical outcomes in patients with COVID-19, Nat. Med. 27 (10) (2021) 1735–1743.

[16] 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, Int. MICCAI Brainlesion Workshop; (2018).

[17] Bercea C.I., Wiestler B., Rueckert D., Albarqouni S. Feddis: Disentangled federated learning for unsupervised brain pathology segmentation. arXiv preprint arXiv: 210303705. 2021.

[18] K.V. Sarma, S. Harmon, T. Sanford, H.R. Roth, Z. Xu, J. Tetreault, et al., Federated learning improves site performance in multicenter deep learning without data sharing, J. Am. Med. Inf. Asso (2021).

[19] Li W, Milletar F, Xu D, Rieke N, Hancox J, Zhu W, et al., editors. Privacy-preserving federated brain tumour segmentation. International workshop on machine learning in medical imaging; 2019.

[20] D. Yang, Z. Xu, W. Li, A. Myronenko, H.R. Roth, S. Harmon, et al., Federated Semi-Supervised Learning for COVID Region Segmentation in Chest CT Using Multi-National Data from China, Medical image analysis, Italy, Japan, 2021.

[21] I. Shiri, A. Vafaei Sadr, A. Akhavan, Y. Salimi, A. Sanaat, M. Amini, et al., Decentralized collaborative multi-institutional PET attenuation and scatter correction using federated deep learning, Eur. J. Nucl. Med. Mol. Imag. (2022) 1–17.

[22] I. Shiri, A.V. Sadr, M. Amini, Y. Salimi, A. Sanaat, A. Akhavanallaf, et al., Decentralized distributed multi-institutional PET image segmentation using a federated deep learning framework, Clin. Nucl. Med. (2022) 10–1097.

[23] Isik-Polat E., Polat G., Kocyigit A., Temizel A. Evaluation and analysis of different aggregation and hyperparameter selection methods for federated brain tumor segmentation. arXiv preprint arXiv:220208261. 2022.

[24] Pati S., Baid U., Zenk M., Edwards B., Sheller M., Reina G.A., et al. The federated tumor segmentation (fets) challenge. arXiv preprint 210505874. 2021.

[25] S. Pati, U. Baid, B. Edwards, M. Sheller, S.-H. Wang, G.A. Reina, et al., Federated learning enables big data for rare cancer boundary detection, Nat. Commun. 13 (1) (2022) 1–17.

[26] H.J.W.L. Aerts, E.R. Velazquez, R.T.H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, et al., Corrigendum: decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach, Nature Comm. (2014).

[27] MICCAI/MD. Anderson Cancer Center Head and Neck Quantitative Imaging Working Group. Matched computed tomography segmentation and demographic data for oropharyngeal cancer radiomics challenges. Sci Data. 2017;4:170077.

[28] A.J. Grossberg, A.S.R. Mohamed, H. Elhalawani, W.C. Bennett, K.E. Smith, T. S. Nolan, et al., Imaging and clinical data archive for head and neck squamous cell carcinoma patients treated with radiotherapy, Sci Data 5 (1) (2018) 1–10.

[29] A. Grossberg, H. Elhalawani, A. Mohamed, S. Mulder, B. Williams, A.L. White, et al., MD Anderson Cancer Center Head and Neck Quantitative Imaging Working Group, HNSCC, 2020, https://doi.org/10.7937/k9/tcia.2020:a8sh-7363. The Cancer Imaging Archive.

[30] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, et al., The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository, J. Digit. Imaging 26 (6) (2013) 1045–1057.

[31] M. Vallieres, E. Kay-Rivest, L.J. Perrin, X. Liem, C. Furstoss, H.J.W.L. Aerts, et al., Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer, Sci. Rep. 7 (1) (2017) 1–14.

[32] I. Shiri, M. Amini, F. Yousefirizi, A. Vafaei Sadr, G. Hajianfar, Y. Salimi, et al., Information fusion for fully automated segmentation of head and neck tumors from PET/CT images, Med. Phys. (2023). In press.

[33] A. Mehranian, M.R. Ay, A. Rahmim, H. Zaidi, 3D prior image constrained projection completion for X-ray CT metal artifact reduction, IEEE Trans. Nucl. Sci. 60 (5) (2013) 3318–3332.

[34] I. Shiri, H. Arabi, Y. Salimi, A. Sanaat, A. Akhavanallaf, G. Hajianfar, et al., COLI-Net: deep learning-assisted fully automated COVID-19 lung and infection pneumonia lesion detection and segmentation from chest computed tomography images, Int. J. Imaging Syst. Technol. 32 (1) (2022) 12–25.

[35] E. Ozfatura, S. Ulukus, D. Gündüz, Straggler-aware distributed learning: communication–computation latency Trade-Off, Entropy 22 (5) (2020).

[36] editorsB McMahan, E Moore, D Ramage, S Hampson, BA y Arcas (Eds.), Communication-efficient Learning of Deep Networks from Decentralized Data, Artificial intelligence and statistics, 2017.

[37] Pillutla K., Kakade S.M., Harchaoui Z. Robust aggregation for federated learning. arXiv preprint arXiv:191213445. 2019.

[38] Deep learning with differential privacy, in: M Abadi, A Chu, I Goodfellow, HB McMahan, I Mironov, K Talwar, et al. (Eds.), Proc of the ACM SIGSAC Conf on Computer and Comm Security, 2016.

[39] Membership inference attacks against machine learning models, in: R Shokri, M Stronati, C Song, V Shmatikov (Eds.), 2017 IEEE Symposium on Security and Privacy (SP), 2017.

[40] Exploiting unintended feature leakage in collaborative learning, in: L Melis, C Song, E De Cristofaro, V Shmatikov (Eds.), 2019 IEEE Symposium on Security and Privacy (SP), 2019.

[41] Malekzadeh M., Hasircioglu B., Mital N., Katarya K., Ozfatura M.E., Gndz D. Dopamine: differentially private federated learning on medical data. CoRR. 2021; abs/2101.11693.

[42] Byzantine-robust distributed learning: towards optimal statistical rates, in: D Yin, Y Chen, R Kannan, P Bartlett (Eds.), International Conference on Machine Learning, 2018.

[43] Practical secure aggregation for privacy-preserving machine learning, in: K Bonawitz, V Ivanov, B Kreuter, A Marcedone, HB McMahan, S Patel, et al. (Eds.), Proc of the ACM SIGSAC Conference on Computer and Communications Security, 2017.

[44] Completeness theorems for non-cryptographic fault-tolerant distributed computing, in: S Goldwasser, M Ben-Or, A Wigderson (Eds.), Proc of the 20th STOC, 1988.

[45] Multiparty computation from somewhat homomorphic encryption, in: I Damgrd, V Pastro, N Smart, S Zakarias (Eds.), Annual Cryptology Conference, 2012.

[46] New directions in cryptography, in: M Hellman (Ed.), IEEE transactions on Information Theory, 1976, p. 22.

[47] A. Shamir, How to share a secret, Commun. ACM (1979).

[48] On data banks and privacy homomorphisms, in: RL Rivest, L Adleman, ML Dertouzos (Eds.), Foundations of secure computation, 1978.

[49] C. Aguilar-Melchor, S. Fau, C. Fontaine, G. Gogniat, R. Sirdey, Recent advances in homomorphic encryption: a possible future for signal processing in the encrypted domain, IEEE Signal Process. Mag. 30 (2) (2013) 108–117.

[50] Gentry C, editor Fully homomorphic encryption using ideal lattices. 41st annual ACM Symp on Theory of computing; 2009.

[51] $$BatchCrypt$}$: efficient homomorphic encryption for $$Cross-Silo$}$ federated learning, in: C Zhang, S Li, J Xia, W Wang, F Yan, Y Liu (Eds.), USENIX Annual Technical Conference (USENIX ATC 20), 2020.

[52] A. Wood, K. Najarian, D. Kahrobaei, Homomorphic encryption for machine learning in medicine and bioinformatics, ACM Comput. Surv. (CSUR) 53 (4) (2020) 1–35.

[53] Calibrating noise to sensitivity in private data analysis, in: C Dwork, F McSherry, K Nissim, A Smith (Eds.), Theory of cryptography conference, 2006.

[54] C. Dwork, A. Roth, The algorithmic foundations of differential privacy, Found. Trends Theor. Comput. Sci. 9 (3–4) (2014) 211–407.

[55] Geyer R.C., Klein T., Nabi M. Differentially private federated learning: a client level perspective. arXiv preprint 171207557. 2017.

[56] Sok: security and privacy in machine learning, in: N Papernot, P McDaniel, A Sinha, MP Wellman (Eds.), 2018 IEEE European Symposium on Security and Privacy (EuroS\&P), 2018.

[57] K. Wei, J. Li, M. Ding, C. Ma, H.H. Yang, F. Farokhi, et al., Federated learning with differential privacy: algorithms and performance analysis, IEEE Trans. Info Forens. Secur. 15 (2020) 3454–3469.

[58] Differentially private federated learning: an information-theoretic perspective, in: S Asoodeh, W-N Chen, FP Calmon, A zgr (Eds.), 2021 IEEE International Symposium on Information Theory (ISIT), 2021.

[59] Beyond differential privacy: composition theorems and relational logic for f-divergences between probabilistic programs, in: G Barthe, F Olmedo (Eds.), Int Coll on Automata, Languages, and Prog, 2013.

[60] J. Duchi, Lecture notes for statistics 311/elec, Engineering 377 (2016).

[61] G. Andrew, O. Thakkar, B. McMahan, S. Ramaswamy, Differentially private learning with adaptive clipping, Adv. Neural Inf. Process. Syst. 34 (2021).

[62] Konecny J., McMahan H.B., Yu F.X., Richtrik P., Suresh A.T., Bacon D. Federated learning: strategies for improving communication efficiency. arXiv preprint arXiv: 161005492. 2016.

[63] Singh A., Vepakomma P., Gupta O., Raskar R. Detailed comparison of communication efficiency of split learning and federated learning. arXiv preprint arXiv:190909145. 2019.

[64] CMFL: mitigating communication overhead for federated learning, in: WANG Luping, WANG Wei, LI Bo (Eds.), 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS), 2019.

[65] Amiri M.M., Gunduz D., Kulkarni S.R., Poor H.V. Federated learning with quantized global model updates. arXiv preprint arXiv:200610672. 2020.

[66] Caldas S., Koneny J., McMahan H.B., Talwalkar A. Expanding the reach of federated learning by reducing client resource requirements. arXiv preprint arXiv: 181207210. 2018.
[67] Federated learning with compression: unified analysis and sharp guarantees, in: F Haddadpour, MM Kamani, A Mokhtari, M Mahdavi (Eds.), Int Conference on Artificial Intelligence and Statistics, 2021.
[68] D.A. Lelewer, D.S. Hirschberg, Data compression, ACM Comput. Surv. (CSUR) 19 (3) (1987) 261–296.
[69] O.K. Al-Shaykh, R.M. Mersereau, Lossy compression of noisy images, IEEE Trans. Image Proces. 7 (12) (1998) 1641–1652.
[70] D. Salomon, Data compression: the Complete Reference, Springer Science \& Business Media, 2004.
[71] K. Sayood, Introduction to Data Compression, Morgan Kaufmann, 2017.
[72] editor Information theory and privacy in data banks. Proceedings of the June 4-8, in: IS Reed (Ed.), 1973, national computer conference and exposition, 1973.
[73] L. Sankar, S.R. Rajagopalan, H.V. Poor, Utility-privacy tradeoffs in databases: an information-theoretic approach, IEEE Trans. Inf. Forens. Secur. 8 (6) (2013) 838–852.
[74] L. Sankar, S.R. Rajagopalan, S. Mohajer, H.V. Poor, Smart meter privacy: a theoretical framework, IEEE Trans. Smart Grid 4 (2) (2012) 837–846.
[75] Privacy-preserving outsourced media search using secure sparse ternary codes, in: B Razeghi, S Voloshynovskiy (Eds.), IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018.
[76] Privacy-preserving image sharing via sparsifying layers on convolutional groups, in: S Ferdowsi, B Razeghi, T Holotyak, FP Calmon, S Voloshynovskiy (Eds.), IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020.
[77] Y. Yakimenka, H.-Y. Lin, E. Rosnes, J. Kliewer, Optimal rate-distortion-leakage tradeoff for single-server information retrieval, IEEE J. Sel. Areas in Commun. (2022).
[78] Variational Leakage: the Role of Information Complexity in Privacy Leakage, in: AA Atashin, B Razeghi, D Gndz, S Voloshynovskiy (Eds.), Proceedings of the 3rd ACM Workshop on Wireless Security and Machine Learning, 2021.
[79] B. Razeghi, F.P. Calmon, D. Gunduz, S. Voloshynovskiy, Bottlenecks CLUB: unifying information-theoretic Trade-offs among complexity, leakage, and Utility, IEEE Trans. Inf. Forens. Secur. 18 (2023) 2060–2075.
[80] Distributed mean estimation with limited communication, in: AT Suresh, XY Felix, S Kumar, HB McMahan (Eds.), International Conference on Machine Learning, 2017.
[81] Swin transformer: hierarchical vision transformer using shifted windows, in: Z Liu, Y Lin, Y Cao, H Hu, Y Wei, Z Zhang, et al. (Eds.), Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.
[82] Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., et al. An image is worth 16×16 words: transformers for image recognition at scale. arXiv preprint arXiv:201011929. 2020.
[83] S. Ashrafinia, Quantitative Nuclear Medicine Imaging Using Advanced Image Reconstruction and Radiomics, The Johns Hopkins University, 2019.
[84] P.E. Shrout, J.L. Fleiss, Intraclass correlations: uses in assessing rater reliability, Psychol. Bull. 86 (2) (1979) 420.
[85] J.P. Weir, Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM, J. Strengtth Condit. Res. 19 (1) (2005) 231–240.
[86] S. Gudi, S. Ghosh-Laskar, J.P. Agarwal, S. Chaudhari, V. Rangarajan, S.N. Paul, et al., Interobserver variability in the delineation of gross tumour volume and specified organs-at-risk during IMRT for head and neck cancers and the impact of FDG-PET/CT on such variability at the primary site, J. Med. Imag. Radiat. Sci. 48 (2) (2017) 184–192.
[87] V. Andrearczyk, V. Oreiller, S. Boughdad, C.C.L. Rest, H. Elhalawani, M. Jreige, et al., Overview of the HECKTOR Challenge At MICCAI 2021: Automatic Head and Neck Tumor Segmentation and Outcome Prediction in PET/CT images. 3D Head and Neck Tumor Segmentation in PET/CT Challenge, Springer, 2021, pp. 1–37.