Medical Image-to-Image Translation with Spatial Self-Attention for Radiotherapy in Federated Learning

Tariq Bdair*, Hiba Saadeh†, Ban Qaqish‡, Aya Sulaq§, Majdi Rawashdeh¶‖
*Data Science Department, Princess Sumaya University for Technology, Amman, Jordan
Email: t.bdair@psut.edu.jo
†Data Science Department, Princess Sumaya University for Technology, Amman, Jordan
Email: hib20200126@std.psut.edu.jo
‡Data Science Department, Princess Sumaya University for Technology, Amman, Jordan
Email: ban20200175@std.psut.edu.jo
§Data Science Department, Princess Sumaya University for Technology, Amman, Jordan
Email: aya20200964@std.psut.edu.jo
¶Software Engineering Department, Al Yamamah University, Riyadh, Saudi Arabia
‖Business Information Technology Department, Princess Sumaya University for Technology, Amman, Jordan
Email: m.rawashdeh@psut.edu.jo

***Abstract*—Globally, cancer remains a leading cause of death, affecting millions of people each year. Accurate medical imaging is crucial for the effective planning of radiotherapy. However, repeated exposure to radiation from Computed Tomography (CT) scans during treatment planning can put patients at more risk. Fortunately, the recent improvement in automated image-to-image translation using deep learning methods has reached a superior performance. However, this might be challenged by data limitation of requiring a large amount of annotated data assembled in one location, privacy, and motion artifacts in medical imaging. Yet, finding such conditions usually is not feasible. To address this, we propose RadiaSync. RadiaSync is a federated learning framework proposed to train decentralized models in a privacy-preserved fashion. In our method, we use CycleGAN architecture for image translation within an FL environment, ensuring patient privacy and collaborative learning across different clients. Further, we propose a spatial attention mechanism that enhances the translated image quality with more than 55% improvement over the baseline.**

***Index Terms*—Federated Learning, Radiotherapy, MRI-to-CT Translation, CycleGAN, Deep Learning, Spatial Self-Attention, Artificial Intelligence in Healthcare, Generative Adversarial Networks.**

I. Introduction

Data-driven machine learning (ML) has evolved as a highly useful approach for developing precise and resilient statistical models using the vast volumes of medical data generated by contemporary healthcare systems. Nowadays, medical images are crucial in cancer diagnosis and patient treatment, leading to both the enhancements of medical care. The utilization of medical data such as medical images allows for more accurate diagnoses and individualized treatment procedures [1]. Recent studies emphasize the role of deep learning in using patient data to predict disorder progression, therapy response, and the probability of damaging developments, thus fostering aggressive interventions and better resource allocation in healthcare techniques [2]. Moreover, the combination of medical data and computer vision can improve the efficiency and usefulness of patient treatments across various medical fields [3], [4], such as cancer radiotherapy [5].

Image-guided radiation therapy has evolved as a revolutionary process in radiotherapy, showing improved accuracy in the targeting of tumors while minimizing harm to surrounding healthy tissues while balanced with the safety of minimizing unnecessary radiation exposure [6], [7]. The therapeutic journey of a cancer patient involves the detailed localization of the tumor for the radiologist to prescribe the beam configuration required for the patient in radiotherapy, ensuring maximal beneficence and minimal exposure to healthy tissues. This procedure requires the patient to undergo a computed tomography (CT) scan that exposes the patient to excessive radiation [7], [8].

Medical imaging utilizes various techniques to capture spatial information about organs and tissues in vivo, such as computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET). These modalities depend on diverse physical principles, resulting in images with inconsistent contrasts and dimensionalities. While this diversity enlarges diagnostic capabilities, it also introduces challenges when merging data across different imaging modalities. Usually, multiple modalities provide complementary details, requiring their combined use for accurate diagnosis. For example, PET/CT combination imaging allows for PET attenuation correction through CT data, and CT is typically used in radiation oncology, augmenting MRI-based diagnostic planning [9].

979-8-3503-5475-1/24/$31.00 ©2024 IEEE

To optimize diagnostic procedures, image quality must be improved before the examination, especially when using computerized methods, as the accuracy of the results is highly dependent on image quality [9]. In some cases, additional image data can be generated from existing ones without further examination. This demands computational methods capable of translating between modalities, improving workflow efficiency, and reducing the burden on patients. However, translating medical images between modalities presents challenges, particularly the risk of introducing unrealistic or unreliable information.

II. Related Works

Current advances in computational methods, including deep learning methods such as Generative Adversarial Networks (GANs) [10], have enhanced the translation and generation of medical images, offering improved accuracy and efficiency in both diagnostic and post-processing tasks.

When GANs [10] was first proposed, it transformed the field of image translation as its suggested architecture enables the generation of realistic synthetic images, instead of the commonly used simple mapping. A vanilla GAN consists of two networks; a generator that creates realistic synthetic images, and a discriminator that classifies generated images as real or fake. This highly competitive environment returns highly convincing outputs [11]. The superior efficiency of GANs, in comparison to other traditional methods, in the field of medical image translation was illustrated by Denck et. al [12]. In this work, GAN architecture was utilized to enhance the quality of MRI scans, by synthesizing MRIs with different characteristics, such as contrast, scanner type, scan location etc. It offered a solution to the challenge of standardizing MRI scans across different settings.

CycleGAN [13], on the other hand, is a specialized GAN for image-to-image translation, that employs two sets of generators and discriminators networks, each corresponding to a different domain or imaging modality, e.g. MRI or CT. The generators translate images between the two domains; for example, one generator translates images from Domain-A, e.g. MRI, to Domain-B, e.g. CT, and the second generator performs the opposite. The two discriminators' tasks are to assess these generated images to verify their authenticity within their particular domains. This architecture allows CycleGAN to translate images without the need for paired training data [14].

CycleGAN served as a basis for the following research [15], which proposed the efficiency of integrating attention mechanisms in translating MRI scans to CT scans. A pivotal innovation in this method was the introduction of an attention-gated classifier, multi-scale feature modulation, and a layer for efficient data compression and reconstruction, all integrated into the existing CycleGAN architecture called Cycle-Consistent GAN. By ingraining an attention mechanism within the discriminator network, the model was able to concentrate on relevant regions of the images, enhancing the accuracy of the translations. This enhancement further validated the selection of CycleGAN as the framework for the proposed image translation pipeline.

However, without access to sufficient data, the above methods will be deterred from achieving their full potential and, eventually, from making the transition from research to clinical trial. One solution that can be contributing to this problem, is federated learning (FL) [16], [17], which is the focus of this paper.

The Federated Learning (FL) framework employed in this study comprises a central server and multiple local clients, where each client represents a participating medical institution within the network. Both the server and all clients are equipped with the CycleGAN architecture. Initially, the central server distributes its base CycleGAN model to all clients, enabling the local models to initialize their weights accordingly. Each client performs local training on its dataset for a set number of epochs. Upon completion of the local training, the locally trained weights are sent back to the central server, where they are aggregated using algorithms such as FedAvg [16]. The FedAvg algorithm combines local model weights through stochastic gradient descent (SGD), ensuring synchronization of learning rates and optimization epochs across all clients. The aggregated average of the weights is then assigned as the updated weights of the central server. This process of model distribution, local training, and weight aggregation repeats over multiple rounds until a set number of rounds is completed or the model converges. In this framework, each client and its local data represent the unique model and dataset of each participating medical institution.

The FL approach facilitates collaborative learning across different institutions with a specific shared goal while preserving patient privacy, through decentralized data storage, and addressing overfitting and restrictions followed by the necessity of unifying the medical image format. Note that the FL paradigm is highly scalable, allowing any medical center to join the network at any given time, and supports continuous learning, meaning the model is adaptive and provides real-time model updates when new data is introduced. Since no actual patient data is being exchanged between servers, the paradigm provides all stated advantages with reduced cost and high bandwidth efficiency.

Previous research [18] proposed an FL pipeline that incorporated CycleGAN for translating brain images from one MRI modality to another. However, it was conducted using the vanilla CycleGAN architecture for translating between distinct MRI modalities. Due to our more complex image translation approach, translating between different medical image modalities, we introduced a spatial self-attention mechanism within the CycleGAN architecture. This mechanism allows the model to get a better sight of long-range relations, and at the same time, to pay more attention to the main features [19] yielding more accurate translations as well as images with better visual interpretability.

In this paper, we propose RadiaSync; a Federated Learning framework, where the fundamental components are medical image translation, utilizing accurate deep learning architectures, spatial self-attention, and a decentralized learning paradigm to ensure patient privacy without limiting the ability of cooperative learning for medical image in radiotherapy. Given the lack of extensive research, further investigation in the area of FL in the medical field is encouraged, leading to the proposal of this research.

Our **contributions** include the following:

- We propose a federated learning framework for the image-to-image translation of MRI scans to CT scans in brain radiotherapy.
- We propose the spatial self-attention mechanism within our CycleGAN network for a more useful sight of long-range relations between different regions in the brain.

III. Methodology

Our methodology, see Fig.1, consists of multiple components including CycleGAN, attention mechanism, and decentralized federated learning paradigm. These components work together to achieve higher-quality CT images suitable for brain radiotherapy. The multidisciplinary approach employed in this paper aims to address common issues in medical imaging, such as noise and motion artifacts. It also ensures decentralized data storage, increasing data privacy while improving image quality and reliability. Next, we explain our methodology in detail.

A. CycleGAN Architecture

As previously explained, the CycleGAN architecture incorporates two distinct generators, the first converts MRIs into CT scans while the second performs the reverse operation. As well as two distinct discriminators that are responsible for validating the authenticity of the generated images, one discriminator for each imaging modality. The generator attempts to synthesize realistic images that the discriminators would not be able to detect, while the discriminators try to enhance their accuracy by increasing their detection strength. Generators are the core of our research since our main goal is to generate the most accurate images possible.

1) Generator: The basic unit in our CycleGAN generators is the UNet model, see Fig.2. The input image is processed, as a tensor of normalized size, through five layers of convolutional down sampling along with Leaky ReLU and 2D Instance Normalization, to reduce spatial dimensions and increase channel depth. In downsampling layers, the generator aims to compress complex features from the input image, which justifies the choice of employing LeakyReLU. Since LeakyReLU allows a small, non-zero unit, for negative gradients it is used for downsampling to preserve important information and prevent the ‘dead neuron’ issue that could be caused by zeroing the effect of important neurons, this is essential when dealing with complex-featured dataset such as medical images.

At the layer where most important features are extracted, the spatial self-attention layer is added along with a residual block to prevent possible gradient explosion or vanishing. Further architectural explanation of the spatial self-attention will be presented afterwards.

Consequently, the data undergoes up sampling via five layers of transposed convolutions, ReLU, and skip connections to restore the original spatial dimensions of the images. The final convolutional layer in the generators ensures that the output image matches the original input image in dimension and format. When upsampling images ReLU has is utilized to ensure the pixel values in the output images are positive, yielding more realistic images. ReLU also introduces non-linearity into the network, capturing more complex patterns to ensure the model can reconstruct intricate details of the original image. This is illustrated in the last five layers in Fig.2. The generator loss calculation involves the Mean Squared Error (MSE), Eq.1, for assessing the error of generator predictions against valid targets and the Mean Absolute Error (MAE), Eq.2, for pixel-wise comparison between the synthetically generated images and their real counterparts. Including both losses yields better image quality. MAE calculates how close the generated image is to the target image in pixel intensity. MAE tends to maintain fine details and makes models less sensitive to outliers. MSE, on the other hand, ensures large differences are penalized more severely, leading to sharper corrections in significant areas of the image.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_{\text{true}} - y_{\text{predicted}})^2 \quad (1)$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_{\text{true}} - y_{\text{predicted}}| \quad (2)$$

2) Discriminator: Within the CycleGAN framework, two discriminators are utilized: one to assess the authenticity of generated CT scans, and the other for MRI scans. The discriminators consist of a sequence of four convolutional layers that progressively downsample the image, apply Leaky ReLU activations and instance normalization. The design implemented allows the discriminators to extract the abstract, most complex, features from the images. The final convolutional layer is set to output a raw scalar map that, after average pooling, outputs a single authenticity score per image. The discriminator acts as a binary classification network, as illustrated in Fig.3.

Training the discriminators consists of presenting a real image, followed by calculating MSE to determine the loss associated with real images based on the output, compared against a valid modality. The same procedure is applied to assess the authenticity of generated images, calculating the loss from the output against an imitation modality. It is worth noting that the process of discriminating the fake images is excluded from the generator’s optimization computations, this prevents the discriminators’ adjustments from influencing the gradients of the generator. This process is essential when training both components to ensure the independent assessment of the quality of generated data without influencing the generator’s internal state during this computation.

The cumulative loss for each discriminator, shown in Eq.3, is the sum of $\mathcal{L}_1$ losses of the real and fake images within the same modality, with overall model loss being the average

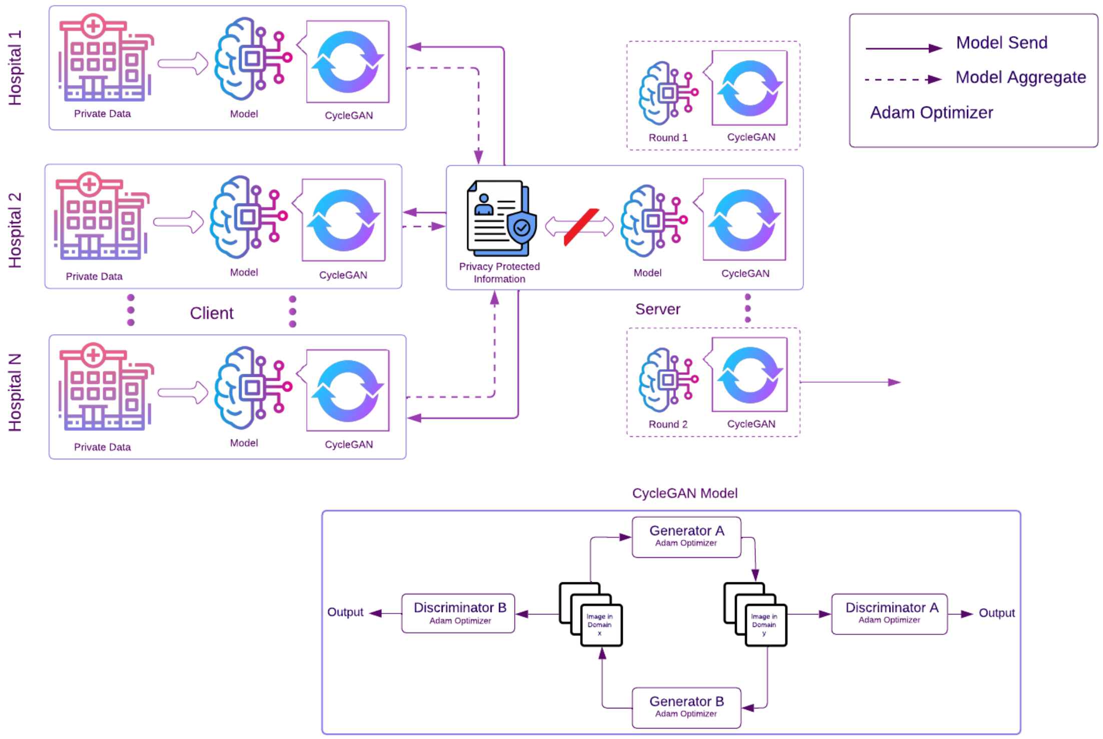

Fig. 1. RadiaSync Pipeline Design: our method consists of multiple components including CycleGAN, attention mechanism, and decentralized federated learning paradigm. These components work together to achieve higher-quality CT images suitable for brain radiotherapy

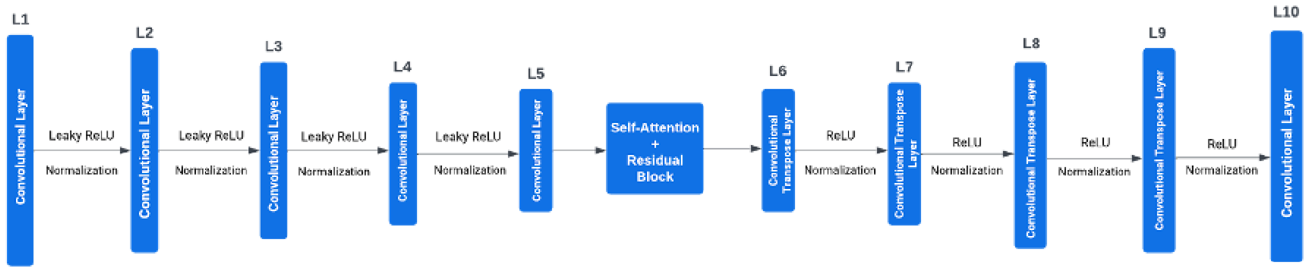

Fig. 2. CycleGAN Generator Network: The generator attempts to synthesize realistic images that the discriminators would not be able to detec.

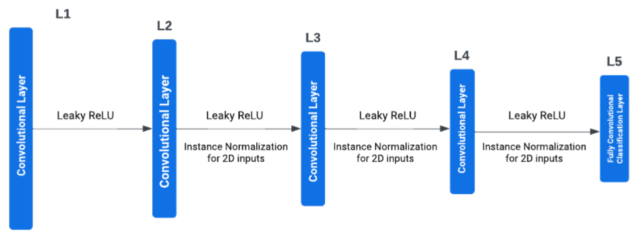
L1
Convolutional Layer
Leaky ReLU
L2
Convolutional Layer
Leaky ReLU
Instance Normalization for 2D inputs
L3
Convolutional Layer
Leaky ReLU
Instance Normalization for 2D inputs
L4
Convolutional Layer
Leaky ReLU
Instance Normalization for 2D inputs
L5
Fully Convolutional Classification Layer

Fig. 3. CycleGAN Discriminator Network: the two discriminators assess the authenticity of generated images for both modalities.

of these sums, facilitating balanced training dynamics across both image modalities using the Adam optimizer.

$$\begin{aligned}\mathcal{L}_{\text{discriminator total}} =& \alpha \cdot (\mathcal{L}_1\text{Loss}_{\text{real}B} + \mathcal{L}_1\text{Loss}_{\text{fake}B}) \\ &+ \beta \cdot (\mathcal{L}_1\text{Loss}_{\text{real}A} + \mathcal{L}_1\text{Loss}_{\text{fake}A})\end{aligned} \quad (3)$$

Where α and β are hyper-parameters.

3) Spatial Self-Attention: In contrast to the previous works, we include the spatial self-attention in CNN-based networks to improve feature representation by allowing each pixel in the feature map to consider all other pixels. The mechanism, illustrated in Fig.4, involves transforming the input features into query, key, and value representations using 1x1 convolutions. The query and key representations are multiplied to calculate attention scores. A SoftMax function is then used to normalize the results. The value representations are weighted by these attention scores, aggregating important information from all positions in the image. The summed weights of values combined with the input via a skip connection, and this shapes up the input features with global context information. The implementation of this mechanism enhances tasks like capturing long-range dependencies and focusing on relevant features.

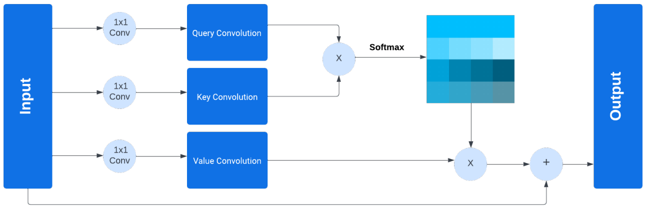

Fig. 4. Spatial Self-Attention Algorithm: the implementation of this mechanism enhances tasks like capturing long-range dependencies and focusing on relevant features.

B. CycleGAN Within a Federated Learning Framework

The pipeline design revolves around a central server that houses both private data and the CycleGAN model, as displayed in Figure 1. Note that the implementation of this simulation was done locally on one machine. Assume the number of clients is n. In practice, dictionary of total, $2 \cdot n$, generators and a dictionary of total , $2 \cdot n$, discriminators are initialized, where each component is named after the image modality they are responsible for, a for MRI and b for CT. Another dictionary is initialized to house the generators and discriminators of the central server. A for-loop is written to iterate over every couple of generators in the dictionary to train them, then through a nested for-loop that iterates over the corresponding discriminators in the discriminators dictionary. After every client has been trained for the predetermined number of epochs and the loss functions for their components is calculated, the state dictionaries of every client are stored in an array that is sent to the server. The server object accepts this array, calculates the average of weights, and recognizes them as its own. It then sends the aggregated weights back to the clients and continues this round until the set round number is reached or until model convergence.

IV. Evaluation Measures

Following previous works utilizing a FL setting [18], MAE, Eq.2, was used to calculate the sum of absolute difference between the predicted and actual values and MSE, Eq.1, was used to calculate the average square difference between the predicted and actual values. The lower the values of both terms, the better the reconstruction of the image compared to the original image.

Peak Signal-to-Noise Ratio (PSNR), referred to in Equation 4, is mainly used to evaluate the quality of the image reconstruction by evaluating the context, or edge, of neuroimages. Since allowing radiologists to visually interpret the generated images is an essential part of this research, this measure is allegedly essential to ensure the image quality is high enough to be humanly interpretable. The higher the PSNR value the better the image quality.

$$PSNR = 10\log_{10}\left(\frac{1}{n}\sum_{i=1}^{n}(y_{\text{true}} - y_{\text{predicted}})^2\right) \tag{4}$$

In addition to that, we utilized Structural Similarity Index (SSIM) to measure the similarity between two images. In contrast to MSE or MAE, which focus on pixel-wise differences, SSIM evaluates the structural and perceptual quality of images by considering changes in luminance, contrast, and structure.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{5}$$

V. Experiments and Results

A. Dataset

SynthRAD2023 dataset [20] is a carefully curated collection of 1080 paired medical images; 540 paired MRI-CT images and 540 paired CBCT-CT images, from patients receiving radiotherapy to the brain and pelvic regions. The dataset was collected for testing synthetic CT generation algorithms in modern radiotherapy. It was sourced between 2018 and 2022 and include patients aged between 3 and 93. The total volume of the data is approximately 25.4 GB, consisting of extensive data across varied imaging modalities and patient conditions. The task of this research is limited to translating MRI to CT scans for the brain area, hence only relative volumes were extracted, leaving us with approximately 6.69 GB, a total of 180 paired MRI-CT images.

The dataset was split into an Independent and Identically Distributed format (IID). To start with, the dataset (180 paired images) was split as 80% training data (144 paired images) and 20% test images (36 paired images) to evaluate the model's final performance. Decentralized data distribution is considered the core of FL, where each client's model is tasked to learn and adapt to its local data characteristics. In this project, four clients were initialized with equal weights of 0.25, indicating that each client possessed a quarter of the total training dataset to locally learn, this leaves every client with 36 paired images (144 images divided by 4 clients). The model is designed to train on 2D images, thus every image was sliced into 30 slices, leaving each client with 1080 tensor slices for training. However, when the single model inference was implemented on the test data, only one slice per image was considered. The testing slice was slice number 100, since it is located right in the middle of the image, and in the middle of the range of slices taken for training; slices 85 to 115. To facilitate the understanding of the splitting process it is illustrated in Fig.5.

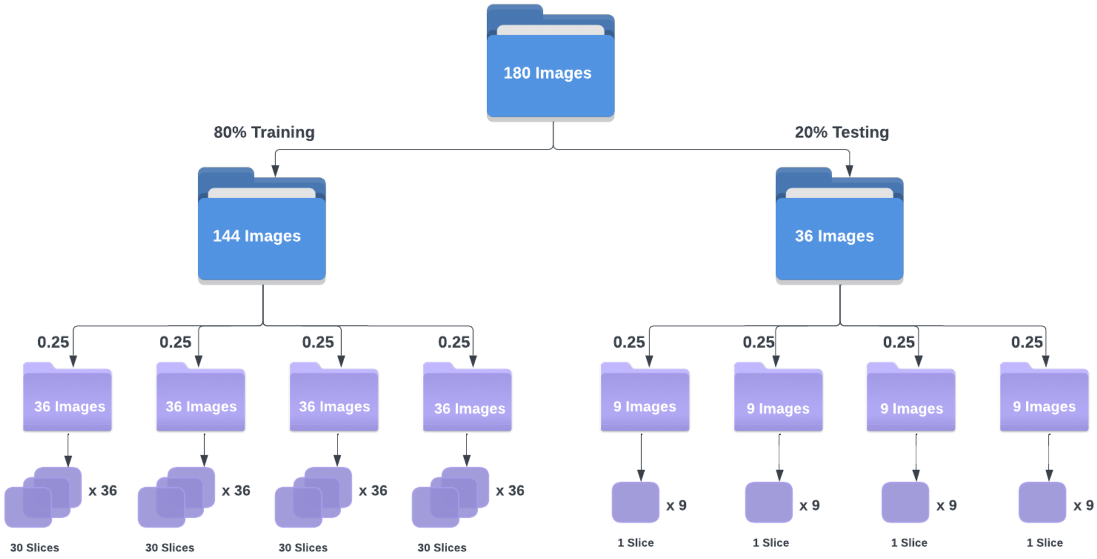

Fig. 5. Data Splitting Mechanism

B. Implementation Details

Dataset images were normalized in their original form before conversion to the '.npy' format. Subsequently, they were expanded into another tensor dimension and normalized using MinMax normalization to transform them into a PyTorch tensor of size 224. Images pre and post-processing are displayed in Fig.6.

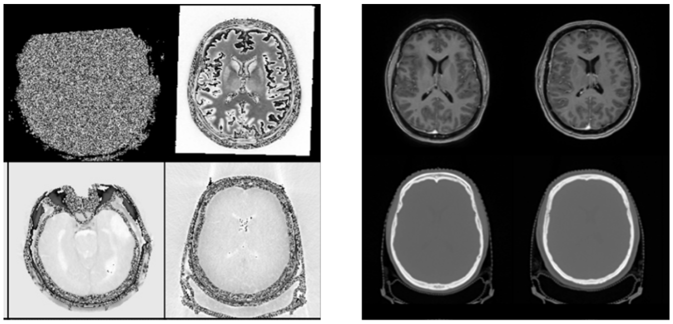

Fig. 6. Images on the left are unprocessed while images on the right show the input images after pre-processing

For all experiments Adam Optimizer is used to optimize CycleGAN components during training, with $\beta1$, $\beta2$, decay rate and learning rate are 0.5, 0.999, 2, and 0.0001, respectively. Locally, the model was trained for 3 epochs per client, and a total of 10 rounds, of weight exchange globally, in the FL setting. The participation rate of clients in the FL framework is 100%, indicating that all clients were chosen to train in every round, and the weights of all clients were aggregated every round. The time taken to train the pipeline from the beginning to the end of 10 rounds ranged approximately from 12 hours to 14 hours. Experiments that deploy spatial self-attention have an output size, for query and key convolutions, of 8 with kernel size 1. All experiments with self-attention also underwent a residual block that utilized a kernel size, stride, padding, and bias of 3, 1, 1, and False, respectively.

C. Experiments Results

1) Central Server vs. Client Accuracy in Federated Learning environment: The main goal of this experiment is to evaluate the training of multiple models that are geographically scattered while maintaining client privacy and ensuring the model captures diverse datasets. This would be a fair comparison since the data split on all clients resembles the ratio of data a hospital would have acquired to the data all hospitals would have when an FL environment is established, collaborating the knowledge and results. The final SSIM of the FL environment central server, after 10 rounds, was 0.6860, while clients' SSIM values ranged from 0.6385 to 0.6962, as shown in Fig.7. The results validate the benefits international organizations would acquire from utilizing such an environment for training their systems. The reason behind the accuracy value increase in clients is due to the state dictionaries that were sent from the server to the clients every round, which include other clients' knowledge and dataset training. Therefore, the FL framework was endorsed in further experiments.

2) CycleGAN vs. UNet: An experiment was conducted to compare the baseline UNet architecture with the proposed CycleGAN architecture, in an FL environment. The results in Table I show that CycleGAN has an SSIM value of 0.6860, while the UNet model recorded an SSIM value of 0.1328, with an improvement of more than 55%. The results confirm the choice of CycleGAN has established a more stable and

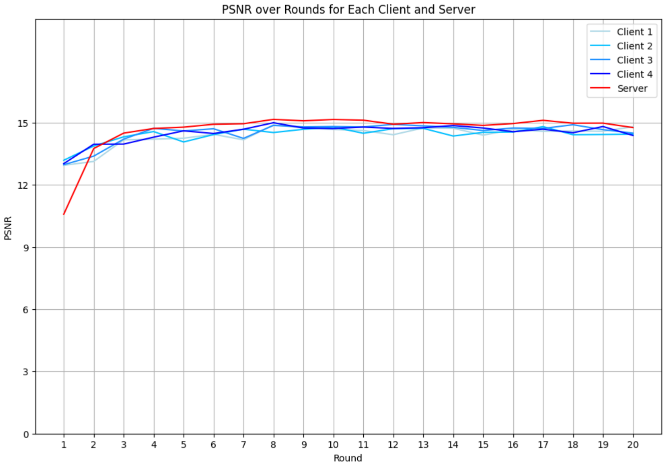

Fig. 7. PSNR values per round for every client and centralized server

accurate training. Further, Figure 8 qualitatively compares the difference between the final results of both architectures on a testing data sample. Even though UNet is considered as the baseline of image translation tasks, its ability to map images from one modality to another is rather weak in an FL environment due to asynchronous communication that could cause the model to become unstable. Hence, the CycleGAN model was pursued for further investigation to explore the possibility of enhancing its image translation ability.

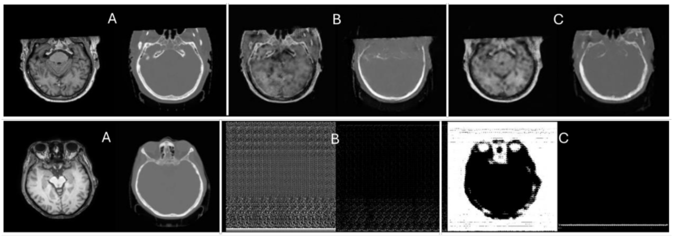

Fig. 8. Top row showcases CycleGAN model results, while bottom row displays UNet results, where the paired images A are the ground truth, the set B are the generated images from the set A, and the final set C are the synthetic images generated from the set B

3) Spatial Self-Attention vs. Without Spatial Self-Attention: To inspect the effect of adopting the spatial self-attention mechanism a test was conducted to observe the difference between a typical CycleGAN structure, and a modified CycleGAN that incorporates a spatial self-attention layer and a residual block at the final downsampling layer in the generator. Both architectures were evaluated within the same FL environment and reported the results in Table I. Images generated with self-attention recorded an SSIM of 0.6942, while those without self-attention achieved a value of 0.6860. Self-attention was visually easier to interpret and more closely resembled the ground truth compared to CycleGAN without self-attention, as displayed in Figure 9. This abides by the hypothesis of this research since the attributes of a CT scan rely on multiple factors that can be scattered along an MRI scan, and not necessarily in the corresponding position of pixels, allowing the model to consider the full image when generating one image modality to another. The qualitative results, displayed in Fig.9, confirm our finding in Table I.

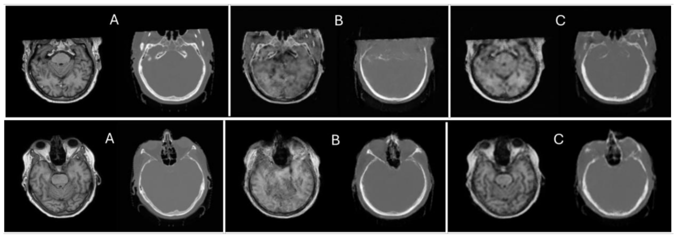

Fig. 9. Top row showcases CycleGAN model results without self-attention, while bottom row displays CycleGAN model results with self-attention, where the paired images A are the ground truth, the set B are the generated images from the set A, and the final set C are the synthetic images generated from the set B

TABLE I
COMPARISON OF EVALUATION METRICS ACROSS DIFFERENT BASELINES

	U-Net	Without Self-Attention	Self-Attention
SSIM	0.1328	0.6860	0.6971
PSNR	4.1069	14.7589	15.2728
MAE	0.3319	0.0373	0.0315

In summary, the proposed method demonstrated superiority in terms of accuracy compared to other possible baselines quantitively and in terms of visual interpretability. This result emphasized the efficiency of the proposed model to translate MRI scans to CT scans, through adopting a spatial self-attention incorporated CycleGAN architecture. While maintaining patient privacy and allowing collaborative learning across medical institutions, by establishing an FL framework, and its potential to assist radiologists in planning radiotherapy while ensuring patient safety.

VI. DISCUSSION

As demonstrated, the pipeline design chosen has excelled in performance when compared to other possible architectures. The finalized architecture utilizes a CycleGAN model that incorporates a spatial self-attention mechanism in an FL environment.

The first test encompassed the advantage of employing an FL learning paradigm for international industries that serve a common purpose. Noticeably, the aggregation of client state dictionaries in the central server allowed the central server to learn from more diverse datasets, capturing a wider range of patterns than a single local client could. Clients are prone to effectively learn important aspects of their local data while neglecting others; aggregation of knowledge in the central server gives it the power to effectively consider all features simultaneously. It is to be noted that the convergence of clients' updated files by the FedAvg algorithm can serve as regularization and secure privacy, i.e. it avoids overfitting problems on a single client's data. This can increase the generalization power of the model. Therefore, the server's performance is better than the performance of any single client.

The internal architecture of the CycleGAN model allowed for the exploration of CT to MRI image translation. Given that the CycleGAN already employs a generator responsible for generating CT scans to MRI, that generator was being trained as well throughout the whole process. When both modalities were compared to one another, it was clear that MRI scans were more complex, detailed, and filled with information, as opposed to CT scans. This difference played a huge role in the difference in accuracies of both generators in the CycleGAN, as the generator responsible for MRI to CT scans always possessed a higher accuracy.

When looking at the results of self-attention from that perspective, the generated MRI scans are more detailed and preserve more information from the CT scans than the traditional image translation task. This result suggests that when further investigating the task of translating CT scans to MRI scans, self-attention is a mechanism that can ensure the effectiveness of this process. The attention mechanism is important in binding together the critical parts of the images, hence allowing the model to focus and translate the meaningful features. Adding an attention mechanism to our work has also yielded images that are better generated by highlighting and preserving the essential details, hence improving the quality and accuracy of the translated images.

VII. Conclusion and Future Work

In conclusion, the implementation of a CycleGAN in a FL environment has proven to be one of the most efficient and accurate architectures to perform MRI-to-CT translation tasks with an SSIM value of 0.6971, PSNR of 15.2728, and MAE of 0.0315. Qualitative results indicate the synthetic CT scan output is relatively close to the ground truth, with a high SSIM value and low error term, and a visually reliable result indicated by the high value of PSNR. Results at hand yielded promising results that suggest the future of image-to-image translation tasks in the medical field has become more reliable with the development of new computer vision models. Future work and enhancements could address the data distributed on all four clients. In this work, we assume that all clients are IID. However, this is not necessarily the case when this architecture is implemented on a larger scale in real life. One could investigate non-IID settings and domain shifts with FL. Further work could include developing more personalized federated learning techniques that would be able to handle different data distributions.

References

[1] P. Lambin, R. T. Leijenaar, T. M. Deist, J. Peerlings, E. E. De Jong, J. Van Timmeren, S. Sanduleanu, R. T. Larue, A. J. Even, A. Jochems *et al.*, "Radiomics: the bridge between medical imaging and personalized medicine," *Nature reviews Clinical oncology*, vol. 14, no. 12, pp. 749–762, 2017.

[2] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual review of biomedical engineering*, vol. 19, no. 1, pp. 221–248, 2017.

[3] J. Gao, Y. Yang, P. Lin, and D. S. Park, "Computer vision in healthcare applications," *Journal of healthcare engineering*, vol. 2018, 2018.

[4] A. Khang, V. Abdullayev, E. Litvinova, S. Chumachenko, A. V. Alyar, and P. Anh, "Application of computer vision (cv) in the healthcare ecosystem," in *Computer Vision and AI-Integrated IoT Technologies in the Medical Ecosystem*. CRC Press, 2024, pp. 1–16.

[5] J. Olveres, G. González, F. Torres, J. C. Moreno-Tagle, E. Carbajal-Degante, A. Valencia-Rodríguez, N. Méndez-Sánchez, and B. Escalante-Ramírez, "What is new in computer vision and artificial intelligence in medical image analysis applications," *Quantitative imaging in medicine and surgery*, vol. 11, no. 8, p. 3830, 2021.

[6] D. A. Jaffray, "Image-guided radiotherapy: from current concept to future perspectives," *Nature reviews Clinical oncology*, vol. 9, no. 12, pp. 688–699, 2012.

[7] G. X. Ding and C. W. Coffey, "Radiation dose from kilovoltage cone beam computed tomography in an image-guided radiotherapy procedure," *International Journal of Radiation Oncology* Biology* Physics*, vol. 73, no. 2, pp. 610–617, 2009.

[8] H. I. Garcia Schüler, M. Pavic, M. Mayinger, N. Weitkamp, M. Chamberlain, C. Reiner, C. Linsenmeier, P. Balermpas, J. Krayenbühl, M. Guckenberger *et al.*, "Operating procedures, risk management and challenges during implementation of adaptive and non-adaptive mr-guided radiotherapy: 1-year single-center experience," *Radiation Oncology*, vol. 16, pp. 1–10, 2021.

[9] K. Armanious, C. Jiang, M. Fischer, T. Küstner, T. Hepp, K. Nikolaou, S. Gatidis, and B. Yang, "Medgan: Medical image translation using gans," *Computerized medical imaging and graphics*, vol. 79, p. 101684, 2020.

[10] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[11] S. AI, "Image to image translation," 2024, accessed: 2024-06-27.

[12] J. Denck, J. Guehring, A. Maier, and E. Rothgang, "Mr-contrast-aware image-to-image translations with generative adversarial networks," *International Journal of Computer Assisted Radiology and Surgery*, vol. 16, pp. 2069–2078, 2021.

[13] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[14] A. Alotaibi, "Deep generative adversarial networks for image-to-image translation: A review," *Symmetry*, vol. 12, no. 10, p. 1705, 2020.

[15] V. Kearney, B. P. Ziemer, A. Perry, T. Wang, J. W. Chan, L. Ma, O. Morin, S. S. Yom, and T. D. Solberg, "Attention-aware discrimination for mr-to-ct image translation using cycle-consistent generative adversarial networks," *Radiology: Artificial Intelligence*, vol. 2, no. 2, p. e190027, 2020.

[16] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[17] N. Rieke, J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein *et al.*, "The future of digital health with federated learning," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–7, 2020.

[18] J. Wang, G. Xie, Y. Huang, J. Lyu, F. Zheng, Y. Zheng, and Y. Jin, "Fedmed-gan: Federated domain translation on unsupervised cross-modality brain image synthesis," *Neurocomputing*, vol. 546, p. 126282, 2023.

[19] S. Li, X. Zhang, J. Xiong, C. Ning, and M. Zhang, "Learning spatial self-attention information for visual tracking," *IET Image Processing*, vol. 16, no. 1, pp. 49–60, 2022.

[20] A. Thummerer, E. van der Bijl, A. Galapon Jr, J. J. Verhoeff, J. A. Langendijk, S. Both, C. N. A. van den Berg, and M. Maspero, "Synthrad2023 grand challenge dataset: Generating synthetic ct for radiotherapy," *Medical physics*, vol. 50, no. 7, pp. 4664–4674, 2023.