

Decentralised Approaches to Training AI Models in Healthcare

Zur Erlangung des akademischen Grades eines

DOKTORS DER INGENIEURWISSENSCHAFTEN (Dr.-Ing.)

von der KIT-Fakultät für

Elektrotechnik und Informationstechnik

des Karlsruher Instituts für Technologie (KIT)

genehmigte

DISSERTATION

von

Ciro Benito Raggio, M.Sc.

geb. in Soveria Mannelli, Italien

Tag der mündlichen Prüfung:

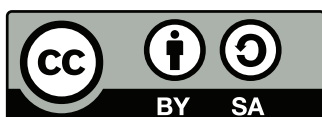
22. Oktober 2026

Referentin:

Prof. Dr.-Ing. Maria Francesca Spadea

Korreferent:

Priv.-Doz. Dr. rer. hum. biol. habil. Oliver Blanck



This document - excluding the cover, pictures, tables and graphs - is licensed under the Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0): <https://creativecommons.org/licenses/by-sa/4.0/>

Abstract

In the last decade, Artificial Intelligence (AI) has transformed the healthcare field, particularly medical image analysis, by enabling the automated learning of feature representations from large-scale datasets across a wide range of medical tasks. However, its broader clinical adoption, especially in the imaging context, remains constrained by the difficulty of collecting large, heterogeneous training datasets. Patient data are subject to strict regulatory frameworks, such as the General Data Protection Regulation (GDPR), which impose substantial barriers to multi-centre data sharing and often render large multi-centre centralised training impractical. Furthermore, AI models trained on data from a single or a limited number of institutions are prone to important performance degradation when deployed in unseen clinical environments, owing to variability in acquisition protocols, scanner hardware, and patient populations.

Federated Learning (FL) is able to address these concerns by enabling collaborative model training without sharing raw data, thereby preserving privacy and improving exposure to diverse data distributions. Despite its demonstrated effectiveness for image classification and segmentation tasks, the application of FL to medical Image-to-Image (I2I) translation was explored only in the field of cross-modality Magnetic Resonance Imaging (MRI) translation prior to this work. This thesis aims to address this gap by focusing on Synthetic Computed Tomography (sCT) generation as a relevant use case and further supporting the FL in healthcare.

In detail, the research work presented in this thesis spanned four interconnected areas: feasibility, efficiency, security, and translational tooling.

The feasibility of federated sCT generation was established through multi-centre studies. The first study investigated brain MRI-to-sCT translation across five institutions, giving rise to the FedSynthCT framework. The second extended this framework to head and neck Cone Beam Computed Tomography (CBCT)-to-sCT translation, using four publicly available institutional datasets. In both studies, federated models trained on heterogeneous distributed data achieved stable convergence and consistent generalisation to independent external datasets, without requiring data centralisation or site-specific adaptations. These results confirmed that sCT generation is a viable application domain for FL, and that inter-site data heterogeneity does not preclude effective decentralised collaboration.

Practical barriers to FL applicability and deployment were subsequently investigated. Within the FedSynthCT framework, an adaptive parameter-freezing strategy was proposed to substantially reduce the computational demand, and therefore the time and energy required for federated training, in line with the increasingly important principles of Green AI. Fur-

thermore, a security analysis of the FedSynthCT framework revealed that gradient inversion attacks did not represent a viable threat in the evaluated setting, while residual vulnerabilities to membership inference and poisoning attacks were identified, motivating the future adoption of system-level safeguards beyond the federated algorithm itself.

The implications of communication-efficiency techniques, including model compression, sparsification, and weight clustering, were further characterised in the context of federated brain tumour segmentation setting, demonstrating that only moderate and well-balanced levels of these techniques can substantially reduce communication overhead while preserving segmentation accuracy.

Open-source software tools were developed and released to bridge the gap between research and its application in healthcare. ImageAugmenter, a medical image augmentation extension for the 3D Slicer Imaging Computing Platform, was published to support data preparation for both centralised and federated pipelines. Subsequently, the 3D Slicer extension named ModalityConverter was released to integrate the pre-trained federated I2I models developed throughout this research alongside existing centralised models, enabling their use and further development within a widely adopted medical research environment. The extension also includes an automated quality assessment module, based on a Deep Learning (DL) framework that predicts quantitative sCT Mean Absolute Error (MAE) without requiring ground-truth Computed Tomography (CT) data, thereby supporting direct output validation in a single environment, without requiring specialised infrastructure or programming expertise.

The contributions of this thesis confirmed that decentralised learning can be a practically viable paradigm for collaborative medical image analysis and established the methodological, analytical, and software foundation for future multi-centre clinical validation and broader deployment of federated I2I translation in healthcare.

Zusammenfassung

Im vergangenen Jahrzehnt hat die Künstliche Intelligenz (KI) das Gesundheitswesen, insbesondere die medizinische Bildanalyse, nachhaltig geprägt. Durch die automatisierte Erlernung von Merkmalsrepräsentationen aus umfangreichen Datensätzen konnte eine Vielzahl medizinischer Aufgaben wesentlich vorangebracht werden. Die breitere klinische Etablierung, insbesondere im Bereich der medizinischen Bildgebung, wird jedoch nach wie vor durch die begrenzte Verfügbarkeit großer und heterogener Trainingsdatensätze erschwert. Patientendaten unterliegen strengen regulatorischen Vorgaben wie der General Data Protection Regulation (GDPR), die den standortübergreifenden Austausch medizinischer Daten erheblich einschränken und multizentrische, zentralisierte Trainingsansätze häufig praktisch nicht umsetzbar machen. Darüber hinaus weisen KI-Modelle, die ausschließlich auf Daten einzelner oder weniger Einrichtungen trainiert wurden, beim Einsatz in bislang unbekannten klinischen Umgebungen oftmals deutliche Leistungsdefizite auf. Ursachen hierfür sind Unterschiede in den Akquisitionsprotokollen, der Scannerhardware sowie den jeweiligen Patientenkollektiven.

Federated Learning (FL) bietet einen vielversprechenden Ansatz zur Bewältigung dieser Herausforderungen, indem ein gemeinsames Modelltraining ohne Austausch der zugrunde liegenden Rohdaten ermöglicht wird. Dadurch kann einerseits der Datenschutz gewahrt und andererseits die Berücksichtigung heterogener Datenverteilungen verbessert werden. Obwohl die Wirksamkeit von FL für Klassifikations- und Segmentierungsaufgaben bereits umfassend gezeigt wurde, beschränkten sich Anwendungen im Bereich der medizinischen Image-to-Image (I2I)-Translation vor dieser Arbeit auf die modalitätsübergreifende Magnet-Resonanz-Tomographie (MRT)-Translation. Ziel der vorliegenden Dissertation war es dementsprechend, diese Forschungslücke zu schließen. Hierfür, wurde die Generierung von Synthetische Computertomographie (sCT) als relevantes Anwendungsbeispiel untersucht, wobei gleichzeitig die Weiterentwicklung föderierter Lernverfahren im Gesundheitswesen unterstützt wurde.

Die in dieser Dissertation vorgestellten Forschungsarbeiten lassen sich vier eng miteinander verknüpften Themenbereichen zuordnen: Machbarkeit, Effizienz, Sicherheit sowie translationsorientierten Softwarewerkzeugen.

Die Machbarkeit der föderierten sCT-Generierung wurde im Rahmen multizentrischer Studien nachgewiesen. Die erste Studie untersuchte die Übertragung von MRT-Bildern des Gehirns in sCT Bildern über fünf Einrichtungen hinweg und führte zur Entwicklung des FedSynthCT-Frameworks. In einer zweiten Studie wurde dieses Framework auf die

Konvertierung von Cone Beam Computed Tomography (CBCT)-zu-sCT Bildern im Kopf- und Halsbereich ausgeweitet, wobei vier öffentlich verfügbare institutionelle Datensätzen herangezogen wurden. In beiden Untersuchungen erzielten die auf heterogenen, verteilt vorliegenden Daten trainierten föderierten Modelle eine stabile Konvergenz sowie eine konsistente Generalisierungsfähigkeit auf unabhängigen externen Datensätzen, ohne dass eine Zentralisierung der Daten oder standortspezifische Anpassungen erforderlich waren. Die Ergebnisse belegen, dass die sCT-Generierung ein geeignetes Anwendungsfeld für FL darstellt und dass die Heterogenität der Daten zwischen verschiedenen Standorten einer erfolgreichen dezentralen Zusammenarbeit nicht widerspricht.

Im weiteren Verlauf wurden praktische Herausforderungen hinsichtlich der Anwendbarkeit und Implementierung von FL untersucht. Innerhalb des FedSynthCT-Frameworks wurde eine adaptive Strategie zum Einfrieren von Modellparametern entwickelt, die den Rechenaufwand und damit den Zeit- und Energiebedarf des föderierten Trainings erheblich reduziert und somit den zunehmend an Bedeutung gewinnenden Prinzipien der Green AI entspricht. Darüber hinaus zeigte eine Sicherheitsanalyse des FedSynthCT-Frameworks, dass Gradient-Inversion-Angriffe unter den untersuchten Bedingungen keine realistische Bedrohung darstellten. Verbleibende Schwachstellen wurden durch Membership-Inference- und Poisoning-Angriffe identifiziert, was die zukünftige Implementierung zusätzlicher Schutzmechanismen auf Systemebene über den föderierten Algorithmus hinaus nahelegt.

Ferner wurden die Auswirkungen verschiedener Verfahren zur Steigerung der Kommunikationseffizienz – darunter Modellkompression, Sparsifizierung und Weight Clustering – im Kontext der föderierten Segmentierung von Hirntumoren untersucht. Es zeigte sich, dass insbesondere moderate und ausgewogen eingesetzte Ausprägungen dieser Techniken den Kommunikationsaufwand erheblich reduzieren können, ohne die Segmentierungsgenauigkeit wesentlich zu beeinträchtigen.

Zur Überführung der Forschungsergebnisse in die praktische Anwendung wurden Open-Source-Softwarewerkzeuge entwickelt und veröffentlicht. Mit ImageAugmenter wurde eine Erweiterung für die 3D Slicer Imaging Computing Platform bereitgestellt, die die medizinische Bildaugmentation sowohl für zentralisierte als auch für föderierte Arbeitsabläufe unterstützt. Anschließend wurde die 3D Slicer-Erweiterung ModalityConverter veröffentlicht, welche die im Rahmen dieser Forschungsarbeiten entwickelten vortrainierten föderierten I2I-Modelle in bestehende zentralisierte Modelle integrieren und so deren Einsatz und Weiterentwicklung innerhalb einer etablierten medizinischen Forschungsumgebung ermöglicht. Die Erweiterung umfasst zudem ein automatisiertes Modul zur Qualitätsbewertung, das auf einem Deep Learning (DL)-basierten Framework beruht und die quantitative sCT-Mean Absolute Error (MAE) ohne Ground-Truth-CT-Daten vorhersagen kann. Dadurch wird eine unmittelbare Validierung der erzeugten Ergebnisse innerhalb einer einzigen Umgebung ermöglicht, ohne dass spezialisierte Infrastruktur oder Programmierkenntnisse erforderlich sind.

Die Ergebnisse dieser Dissertation zeigen, dass dezentrales Lernen ein praktikables Paradigma für die kollaborative medizinische Bildanalyse darstellt. Zugleich schaffen sie die methodischen, analytischen und softwaretechnischen Voraussetzungen für zukünftige multi-

zentrische klinische Validierungsstudien sowie für eine breitere Etablierung der förderierten I2I-Translation im Gesundheitswesen.

Acknowledgments

This work would not have been possible without the support of the many people who accompanied me throughout the three years of my doctoral research, and to whom I owe my deepest gratitude.

My special thanks go to Prof. Dr.-Ing. Maria Francesca Spadea for giving me the opportunity to carry out my doctoral research at the IBT on a project that I have genuinely enjoyed from the very beginning. Thank you for your guidance, support, flexibility and trust throughout these years.

I would also like to thank everyone who has supported this research from the beginning, as well as those who have contributed along the way: Priv.-Doz. Dr. rer. hum. biol. habil. Oliver Blanck, Prof. Dr.-Ing. Paolo Zaffino, Prof. Dr. rer. nat. Francesco Cicone, Prof. Dr.-Ing. Lucia Migliorelli, Mathias Zabaleta, Nils Skupien, Lina Bucher, Dr.-Ing. Asaf Raza, Prof. Dr.-Ing. Antonella Guzzo and Prof. Dr.-Ing. Giancarlo Fortino. All of you provided valuable support in different ways, and much of this research would not have been possible without it.

I am truly grateful to have been part of the IBT. Throughout these years, I have always found support, mutual respect and a stimulating environment. Everyone with whom I had the opportunity to interact, collaborate or travel has taught me something and left a lasting impression on me. I sincerely thank you all and wish you the very best for the future.

I would also like to thank all the students I had the pleasure of supervising over the years. Your enthusiasm, curiosity and commitment made this journey even more meaningful and rewarding.

Thanks to my family, who have always supported me, cared about my well-being, and encouraged me throughout this journey, despite the distance.

Finally, thank you, Alessia, for believing in me, supporting and encouraging me to take this path, and for sharing with me all the ups and downs of this far from easy journey.

Contents

Abstract	i
Zusammenfassung	iii
Acknowledgments	vii
Abbreviations	xiii
1 Introduction	1
1.1 Motivation	1
1.2 Objective of the Thesis	3
1.3 Structure of the Thesis	4
<hr/>	
I Fundamentals	7
<hr/>	
2 Clinical Context	9
2.1 The Role of Multimodal Imaging for Planning and Guidance	9
3 Technical Fundamentals	25
3.1 Deep Learning in Medical Image Analysis	25
3.2 Principles and Applications of Federated Learning in Medical Imaging . . .	36
<hr/>	
II Enabling Robust and Privacy-Preserving Synthetic CT Generation with Federated Learning	41
<hr/>	
4 FedSynthCT-Brain: A Federated Brain MRI-to-CT Synthesis Framework .	43
4.1 Introduction	43
4.2 Manuscript	45
5 Privacy-Preserving and Generalizable Federated Head and Neck CBCT-to-CT Translation	65
5.1 Introduction	65
5.2 Manuscript	66
<hr/>	

III Efficiency, Sustainability and Reliability of Federated Learning in Healthcare	83
6 Energy-Efficient Federated Learning via Adaptive Encoder Freezing for MRI-to-CT Translation	85
6.1 Introduction	85
6.2 Manuscript	86
7 Security Analysis of a Federated Learning Framework for Medical Image-to-Image Translation	107
7.1 Introduction	107
7.2 Manuscript	108
8 Robust Neurocomputing Model in Efficient Federated Brain Tumour Segmentation with Sparsification and Weights Clustering	123
8.1 Introduction	123
8.2 Manuscript	125
IV Tooling, Reproducibility and Translation	139
9 ImageAugmenter: A user-friendly 3D Slicer tool for medical image augmentation	141
9.1 Introduction	141
9.2 Manuscript	142
10 A Quality Control Tool to Predict Synthetic Computed Tomography Hounsfield Unit Accuracy	153
10.1 Introduction	153
10.2 Manuscript	154
11 ModalityConverter: Medical Image-to-Image Translation in 3D Slicer	167
11.1 Motivation and significance	167
11.2 Software architecture and module ecosystem	168
11.3 Illustrative example	178
11.4 Conclusions	181
V Key Findings and Future Directions	183
12 Discussion and Conclusion	185
12.1 Enabling Federated Learning for sCT Generation	185
12.2 Sustainability, Efficiency, and Security of Federated Learning	187
12.3 Tooling, Reproducibility, and Research Translation	189

12.4 Limitations and Future Directions	190
12.5 Conclusion	191
References	193
List of Publications and Supervised Theses	203

Abbreviations

AI	Artificial Intelligence i, 1, 3, 35 f., 85 f., 123, 167, 186
BraTS	Brain Tumor Segmentation 123
CBCT	Cone Beam Computed Tomography i, iv, 1 f., 9, 14 f., 23, 39, 65 f., 153, 185 f., 190
cGAN	conditional Generative Adversarial Network 30, 66, 186
CNN	Convolutional Neural Network 1, 25, 27 f., 30 f., 38
CT	Computed Tomography ii, 1 f., 9, 12–15, 17, 19, 27, 29, 36, 39, 65, 189 f.
DDPM	Denoising Diffusion Probabilistic Models 31, 35
DL	Deep Learning ii, iv, 1 f., 4, 25, 27, 31, 34 ff., 38 f., 43, 65, 85 f., 107, 123, 141 f., 153, 168, 171, 189, 191
FCN	Fully Convolutional Network 28
FID	Fréchet Inception Distance
FL	Federated Learning i, iii f., 2 ff., 36–39, 86, 107, 123, 142, 154, 167, 185–191
GAN	Generative Adversarial Network 30 f., 35, 38 f.
GDPR	General Data Protection Regulation i, iii, 1, 35, 123
HIPAA	Health Insurance Portability and Accountability Act
HU	Hounsfield Unit 13 f., 65 f.
I2I	Image-to-Image i–v, 1 ff., 29–32, 38 ff., 65, 153, 167, 171, 181, 185, 187 ff., 191
IGRT	image-guided radiotherapy 65 f.
JSON	JavaScript Object Notation 169
MAE	Mean Absolute Error ii, iv, 32 f., 153, 171, 189 f.
ML	Machine Learning 25, 36
MRI	Magnetic Resonance Imaging i, 1 f., 9, 17 f., 20, 29, 31, 36, 38 f., 44, 123, 153, 185 f., 190
ONNX	Open Neural Network Exchange 174
PET	Positron Emission Tomography 1, 9, 15, 17
PSNR	Peak Signal-to-Noise Ratio 32 f.
RELU	Rectified Linear Unit 26, 28
RMSE	Root Mean Squared Error 32 f.

sCT	Synthetic Computed Tomography i f., 2 ff., 22, 33, 44, 66, 153, 167, 185 ff., 189 ff.
SSIM	Structural Similarity Index Measure 32 f.
VGG	Visual Geometry Group 26
ViT	Vision Transformer 31

Introduction

1.1 Motivation

Medical imaging plays a key role in modern healthcare, supporting diagnosis, prognosis, treatment planning and monitoring across a wide range of clinical domains. Imaging modalities such as Magnetic Resonance Imaging (MRI), Computed Tomography (CT), Cone Beam Computed Tomography (CBCT), and Positron Emission Tomography (PET) provide complementary three-dimensional (3D) anatomical and functional information and are routinely employed in fields including oncology, neurology, and radiotherapy. The increasing availability of digital medical images, together with advances in computational power, has fuelled a rapid adoption of Artificial Intelligence (AI) and, in particular, Deep Learning (DL) techniques in healthcare [1–4].

Over the last decade, DL-based methods have demonstrated state-of-the-art performance in a variety of medical imaging tasks, including classification, segmentation, detection, and image synthesis. Convolutional Neural Network (CNN) and, more recently, generative models have achieved remarkable accuracy in disease classification, organ and lesion segmentation, and Image-to-Image (I2I) translation, often surpassing traditional rule-based approaches [1–4]. These advances have positioned DL as a key enabling technology for data-driven medical imaging pipelines. Despite their success, most DL approaches in medical imaging rely on a centralised training paradigm, in which data from one or more institutions are aggregated into a single dataset and used to train a global model. While effective from a methodological standpoint, this paradigm presents substantial limitations in real-world healthcare settings. Medical images are considered sensitive personal data, as they directly relate to an individual’s health status. Consequently, their collection, storage, and processing are subject to strict regulatory frameworks, such as the General Data Protection Regulation (GDPR) in Europe [5].

Although data centralisation is not inherently prohibited, complying with these regulations necessitates comprehensive legal agreements, extensive governance procedures, and robust technical safeguards. These measures often impede or even preclude multi-centre collaborations [6–9].

In addition to regulatory and bureaucratic barriers, centralised DL models are particularly vulnerable to data heterogeneity and distribution shifts. Medical images acquired at different institutions frequently differ in scanner vendors, acquisition protocols, reconstruction algorithms, and patient demographics. As a result, models trained on centralised datasets—especially when dominated by data from a limited number of centres—often show degraded performance when applied to unseen clinical settings [10–16]. This lack of robustness and generalisability represents a major obstacle to the deployment of DL systems in routine clinical practice.

Federated Learning (FL) has emerged as a promising alternative to centralised training, addressing both privacy and generalisation concerns. In FL, multiple institutions collaboratively train a shared model by keeping the data local and exchanging only model updates with a central server. This paradigm enables learning from distributed, heterogeneous datasets without transferring raw medical images, thereby reducing privacy risks and simplifying compliance with data protection regulations. Beyond privacy preservation, FL offers the opportunity to expose models to a broader range of data distributions, potentially improving robustness and performance on unseen domains [15, 16].

Within medical imaging, FL has been extensively investigated for classification and segmentation tasks, demonstrating competitive performance compared to centralised baselines while maintaining data privacy [4, 14–19]. However, its application to more complex generative tasks remains limited.

Image synthesis, and in particular medical I2I translation, poses additional challenges due to its sensitivity to subtle anatomical details, intensity distributions, and domain-specific artefacts. These challenges are amplified in federated settings, where data heterogeneity across sites is the standard rather than the exception. Medical I2I translation represents a pivotal component within a range of clinical workflows. One of the most relevant applications is the generation of Synthetic Computed Tomography (sCT) images from MRI or CBCT.

Synthetic CT enhances the accuracy for MRI-only radiotherapy workflows, reducing image registration uncertainties and patient exposure to ionising radiation, and facilitating attenuation correction in hybrid imaging systems such as PET/MRI. Numerous DL-based approaches for sCT generation have been proposed in the literature, achieving high accuracy under controlled experimental conditions [1, 3, 20]. Nevertheless, these methods are typically trained using a centralised paradigm and are often evaluated only on data drawn from the same distribution as the training set, limiting their applicability in real multi-centre scenarios.

The combination of strict data protection requirements, pronounced inter-centre heterogeneity, and the clinical relevance of sCT synthesis highlights a clear gap in the current literature. While FL offers a solution to address these challenges, its potential for medical image synthesis—and for sCT generation in particular—has not yet been systematically explored.

1.2 Objective of the Thesis

General Objective:

The aim of this thesis is to investigate and advance decentralised learning strategies in healthcare, focusing on FL for medical I2I translation, with particular emphasis on sCT generation. The thesis addresses challenges related to data heterogeneity, privacy preservation, computational efficiency, and robustness, aiming to develop robust solutions suitable for real-world clinical deployment.

Specific Objectives:

To achieve this general objective, the following specific objectives have been defined:

1. To assess the feasibility of FL for MRI-to-sCT translation and validating federated frameworks on multi-centre datasets.
2. To evaluate the impact of inter-site data heterogeneity and domain shifts in federated settings, analysing the generalisation capability of the federated models using an independent dataset from an institution not involved during the federated training.
3. To improve the efficiency and sustainability of federated training by developing optimisation strategies to reduce convergence time, communication overhead, and energy consumption, in line with Green AI principles.
4. To extend the federated framework to additional modalities and anatomical regions, such as CBCT-to-sCT translation.
5. To investigate the robustness and security of the FL paradigm by analysing the effects of common attack scenarios and their implications for medical imaging synthesis applications.
6. To develop and provide open-source tools supporting federated learning workflows, including offline image augmentation, model distribution and quality assessment, facilitating adoption by clients within and outside the federation.

1.3 Structure of the Thesis

Part I – Fundamentals presents the required fundamentals which are introduced at a medical and a technical level.

- **Chapter 2** introduces the clinical background relevant to this thesis including principles of medical imaging, Computed Tomography, Cone-Beam Computer Tomography, Positron Emission Tomography and Magnetic Resonance Imaging.
- **Chapter 3** provides the technical and methodological background, covering DL and FL in medical imaging, and image-to-image translation applications.

Part II – Enabling Robust and Privacy-Preserving Synthetic CT Generation with Federated Learning presents the core research, development, application and validation of FL for sCT generation across multiple institutions.

- **Chapter 4** introduces the first study on MRI-to-sCT translation using federated learning across four different institutions, describing the framework, validation, and results.
- **Chapter 5** extends the approach to CBCT-to-sCT synthesis and evaluates generalisation, and performance.

Part III – Efficiency, Sustainability and Reliability of Federated Learning in Healthcare addresses computational and operational aspects of federated training in the healthcare environment.

- **Chapter 6** presents a federated Green AI approach to enhance sustainability, reduce training time and energy consumption for federated sCT synthesis.
- **Chapter 7** presents an analysis of potential attack vectors and defences within the FedSynthCT framework.
- **Chapter 8** evaluates the implication of clients sparsification, compression and clustering for brain tumour segmentation under different conditions of data heterogeneity, highlighting implications for performance and reliability.

Part IV – Tooling, Reproducibility, and Translation describes software frameworks developed to support federated and centralised medical AI, enhancing reproducibility and accessibility.

- **Chapter 9** presents an open-source 3D Slicer module that enables intuitive medical image augmentation through a graphical interface.
- **Chapter 10** presents a deep learning-based quality control framework for sCT that predicts slice-wise mean absolute error without ground-truth CT, supporting clinical deployment of centralised and federated models.
- **Chapter 11** presents an open-source 3D Slicer extension that integrates pre-trained deep learning and federated learning models for MRI-to-CT, CBCT-to-CT, and PET-to-CT translation, enabling accessible and reproducible inference within the 3D Slicer environment.

Part V – Key Findings and Future Directions summarises the contributions and presents perspectives for future work.

- **Chapter 12** provides an overview of the key findings, research achievements and limitations, as well as presenting the broader implications, open challenges and future directions of this thesis.

PART I

FUNDAMENTALS

Clinical Context

2.1 The Role of Multimodal Imaging for Planning and Guidance

Medical imaging comprises a variety of modalities, each employing distinct physical principles to visualise anatomical structures and functions. Common modalities include X-ray radiography, Computed Tomography (CT), Cone Beam Computed Tomography (CBCT), Positron Emission Tomography (PET), and Magnetic Resonance Imaging (MRI) [21].

These imaging techniques can be broadly categorised into anatomical modalities (e.g., X-ray radiography, CT, CBCT) and functional or molecular modalities (e.g., PET). MRI primarily provides high-resolution anatomical information, although specialised sequences can also probe functional or physiological properties of tissues. Anatomical imaging primarily delivers structural information, such as morphology and spatial relationships between tissues, whereas functional imaging enables the assessment of metabolic activity, perfusion, receptor expression, or other physiological processes. The integration of these complementary sources of information has become increasingly important in modern clinical practice, with the strengths of each modality compensating for the limitations of the others. This integration enhances the detection and delineation of pathological lesions, as well as risk stratification, thereby supporting more precise therapeutic decision-making.

Among the clinical fields that particularly benefit from multimodal imaging, radiation oncology represents a paradigmatic example. In this domain, multimodal imaging integration plays a pivotal role throughout the treatment pathway, from target definition and dose calculation to image guidance and adaptive strategies. The ability to combine geometric accuracy with biological characterisation underpins the development of increasingly personalised and adaptive treatment approaches, which are discussed in the following sections.

This chapter introduces the fundamental physical principles underlying the most widely used medical imaging modalities and outlines their clinical roles in diagnosis, treatment planning, and therapy guidance.

However, each technique produces images characterised by different contrast mechanisms, spatial resolution, and quantitative properties (presented in Table 2.1). Consequently, information obtained from one modality is not directly interchangeable with another. In many clinical workflows, including radiotherapy planning and image-guided therapy, it becomes necessary to combine, register, or computationally translate images across modalities. The development of reliable methods to bridge this modality gap has become an important research topic in medical image analysis.

Table 2.1: Comparative overview of commonly used medical imaging modalities.

Modality	Key characteristics
X-ray Radiography	Type: Anatomical Contrast: X-ray attenuation (density/composition) Strengths: Fast, cost-effective, excellent for bone imaging and fractures Limitations: Poor soft-tissue contrast, 2D projection only, ionising radiation exposure
CT	Type: Mainly anatomical Contrast: Volumetric X-ray attenuation; Hounsfield Unit (HU) scale Strengths: High-resolution 3D anatomical imaging, good for thoracic/abdominal organs, trauma, vasculature Limitations: Limited soft-tissue contrast, ionising radiation exposure, lower functional information
CBCT	Type: Anatomical Contrast: Cone-beam X-ray attenuation (not HU-calibrated) Strengths: Compact, lower dose for small FOV, high-resolution for bone visualisation, in-room image-guided radiotherapy imaging Limitations: Lower quantitative accuracy than CT, ionising radiation exposure, limited soft-tissue contrast, prone to scatter and beam-hardening artefacts
PET	Type: Functional / molecular Contrast: Radiotracer distribution; SUV, kinetic modelling Strengths: Sensitive metabolic/functional imaging, whole-body imaging, tumour detection/staging, therapy monitoring Limitations: Low spatial resolution, internal ionising radiation from the radiotracer, expensive, requires radiotracers with short half-life, infrastructure-intensive
MRI	Type: Mainly anatomical Contrast: Proton density, T_1/T_2 relaxation, diffusion, perfusion, susceptibility Strengths: Excellent soft-tissue contrast, multiplanar, non-ionising, functional imaging (fMRI, DWI), versatile Limitations: Long acquisition times, motion sensitivity, contraindications for some implants, higher cost, limited bone imaging

Finally, the chapter is organised as follows. Section 2.1.1 presents X-ray-based imaging modalities and their clinical applications, followed by CT and CBCT (Section 2.1.2). PET and hybrid techniques are introduced in Section 2.1.3, while MRI and its anatomical and

functional contrasts are discussed in Section 2.1.4. The chapter concludes with a practical application of multimodal imaging in image-guided radiotherapy (Section 2.1.5).

2.1.1 X-ray radiography

X-ray imaging represents a cost-effective and widely adopted technique that is widely employed in both medical diagnostics and industrial non-destructive testing. Its capability to penetrate matter, including the human body, has enabled major progress in non-invasive visualization of internal structures. The strategic and technological relevance of this modality has consequently driven the rapid advancement of high-performance X-ray detectors, along with a broad range of related imaging applications [22].

In detail, X-rays are electromagnetic waves belonging to the ionizing radiation spectrum, with wavelengths approximately spanning from 0.01 to 10 nm. As X-ray photons propagate through matter, they may be transmitted, absorbed, or deflected from their original trajectory. Absorption and scattering phenomena are governed by the attenuation properties of the material and can be described by the Beer-Lambert law:

$$I = I_0 e^{-\mu d}, \quad (2.1)$$

where I denotes the transmitted photon intensity, I_0 is the incident photon intensity, μ represents the linear attenuation coefficient of the medium, and d corresponds to the photon path length through the material.

The overall attenuation results from the combined contribution of three main interaction mechanisms: (i) the photoelectric effect, (ii) Compton scattering, and (iii) Rayleigh scattering. These three interaction mechanisms are presented in Figure 2.1. Their relative contribution depends both on the atomic composition of the material (in terms of atomic number Z) and on the energy of the incident photons. At lower photon energies, attenuation is primarily driven by the photoelectric effect, whereas Compton scattering becomes increasingly relevant for higher photon energies and for materials characterised by low atomic number. Rayleigh scattering, also referred to as coherent scattering, corresponds to an elastic interaction in which the photon is deflected without energy loss and without ionisation. Although its contribution to total attenuation in the diagnostic energy range (10-150 keV) is generally modest compared with the photoelectric and Compton processes, it may become non-negligible at lower photon energies in some imaging modalities (such as mammography) and contributes to scattered radiation, thereby slightly affecting image contrast [22, 23].

In essence, the purpose of an X-ray imaging system is to convert the residual X-ray photons emerging from an object into a visual representation suitable for analysing its internal structure. To this end, the detector is positioned downstream of the object in order to capture the transmitted radiation and register the resulting intensity distribution. Consequently, the signal is transformed into a projection pattern, which is then processed to generate a two-dimensional (2D) radiographic image [22]. Variations in the transmitted photon intensity,

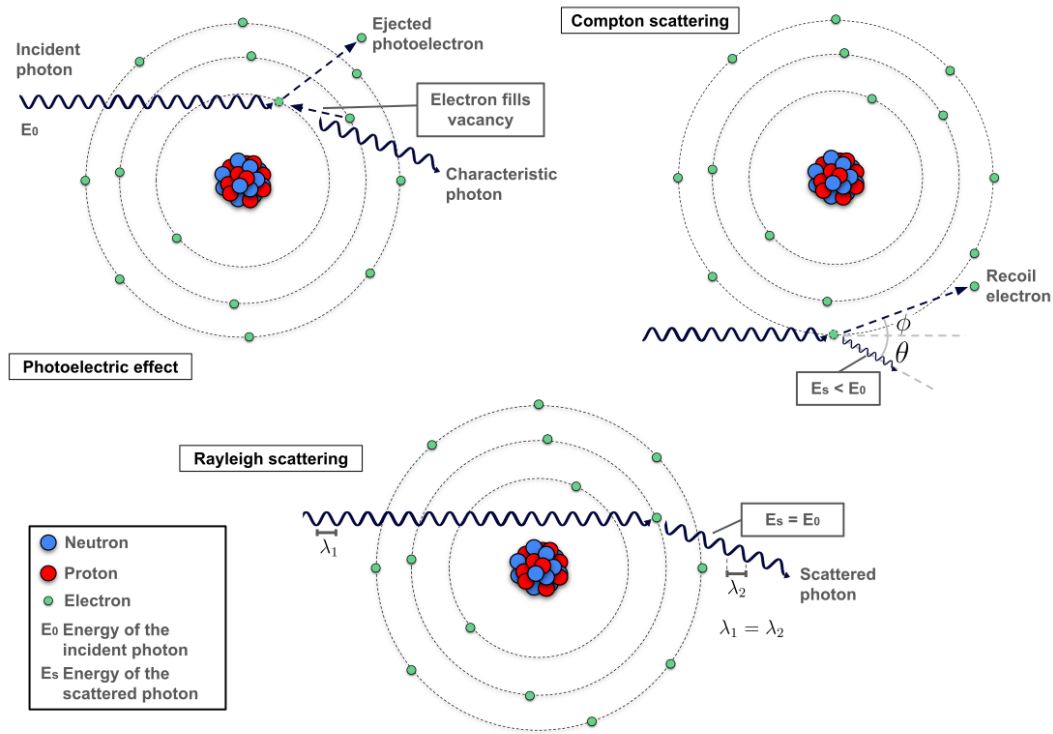


Figure 2.1: Representation of the principal interaction mechanisms of the X-ray photons. The photoelectric effect (left) involves complete photon absorption and the ejection of an inner-shell electron. Compton scattering (right) describes an inelastic interaction with a weakly bound outer shell electron, resulting in photon deflection with reduced energy and electron recoil. Rayleigh scattering (centre) involves an elastic interaction with the entire atom, resulting in photon deflection without energy transfer or ionisation. This process occurs when the wavelength of the photon is significantly longer than the radius of the atom, causing the atom to scatter the radiation coherently. E_0 denotes the energy of the incident photon, E_s the energy of the scattered photon, while λ_1 and λ_2 indicate the wavelengths of the incident and scattered photons, respectively.

determined by the spatial distribution of attenuation coefficients, give rise to contrast in the final image, thereby enabling the visualization of distinct internal features [22].

Owing to their remarkable penetrating capability, X-rays have become a key modality in medical imaging. Continuous technological progress significantly enhanced diagnostic radiography, enabling detailed visualization of skeletal structures and supporting the detection of fractures, joint dislocations, bone pathologies, and foreign bodies. Such information plays a crucial role in clinical decision-making and surgical planning [22].

2.1.2 Computed Tomography and Cone-Beam CT

While conventional radiography provides a 2D projection of a three-dimensional (3D) object, CT enables volumetric imaging through the acquisition of multiple projections at different angular positions around the subject. In CT, a series of X-ray projections is collected over a

rotational trajectory around the cranio-caudal axis of the patient, and cross-sectional images are reconstructed by solving the inverse problem associated with X-ray attenuation [22, 24–26]. Stacking the reconstructed slices allows the generation of a full 3D representation of the scanned anatomy, which corresponds to the 3D volumetric map of X-ray attenuation, thereby overcoming the superimposition limitations inherent to planar radiography.

Early acquisition systems were based on a translate-rotate geometry with pencil-beam collimation and single detectors. Subsequent technological generations progressively introduced fan-beam geometries, multi-detector arrays, continuous rotation enabled by slip-ring technology, and helical acquisition modes. These innovations significantly reduced acquisition times while improving spatial resolution and volumetric coverage [25, 26].

Modern multi-detector CT (MDCT) systems employ wide detector arrays and fan- or cone-shaped beams, enabling the simultaneous acquisition of multiple slices per rotation. The introduction of helical (spiral) scanning further allowed continuous data acquisition during patient table translation, leading to faster examinations and improved 3D reconstructions. In addition, dual-energy and multi-energy CT systems, based on the use of two distinct X-ray spectra (either through dual sources, rapid kVp switching, or energy-discriminating detectors), allow for material differentiation and virtual monoenergetic imaging [27]. Such techniques enhance tissue characterisation, improve contrast resolution, and enable applications such as virtual non-contrast imaging and advanced angiographic studies.

In conventional CT, voxel intensities are expressed in Hounsfield Unit (HU), defined as a linear transformation of the reconstructed linear attenuation coefficient relative to water:

$$HU = 1000 \frac{\mu - \mu_{\text{water}}}{\mu_{\text{water}}}. \quad (2.2)$$

By definition, distilled water at standard conditions is assigned a value of 0 HU, while air corresponds to approximately -1000 HU. Other biological tissues exhibit characteristic ranges of attenuation values depending on their physical density and composition. For example, adipose tissue typically presents values around -100 to -80 HU, whereas most soft tissues lie between approximately 30 and 70 HU. Bone structures show substantially higher values: trabecular bone commonly ranges from about 50 to 200 HU, while dense cortical bone can extend from roughly 250 up to 2000 HU depending on mineral density and acquisition parameters.

The scale is open-ended, yet in practice, it is generally considered to reach a limit of approximately 3000 HU (e.g. metals or highly concentrated contrast agents) [25]. This definition ensures standardised and reproducible quantitative measurements across scanners, provided appropriate calibration, scatter correction, and beam-hardening compensation are applied.

Clinically, CT has become essential for comprehensive anatomical assessment. It provides high-resolution visualisation of osseous structures, thoracic and abdominal organs, vascular anatomy, and traumatic injuries. In the domain of thoracic imaging, for instance, CT facilitates comprehensive evaluation of pulmonary nodules, interstitial patterns, and consolidations, thereby supporting diagnosis, staging, and treatment planning [25, 26]. Nevertheless,

image contrast in CT arises from differences in the X-ray attenuation properties of tissues, primarily determined by their physical density and effective atomic number. While this mechanism enables excellent delineation of structures with substantial density differences, such as bone and air-filled spaces, the intrinsic soft-tissue contrast of CT is generally limited.

A related but technically distinct modality is the CBCT, in which a cone-shaped X-ray beam and a flat-panel detector rotate around the subject, acquiring a large number of 2D projections over a single rotation. Compared with conventional fan-beam CT, CBCT systems are typically characterised by a more compact design, lower radiation dose in specific applications, and high spatial resolution for small fields of view.

However, despite the geometric similarities with multi-detector CT, CBCT differs substantially in terms of quantitative accuracy. In CBCT, grey values are generally not directly calibrated to HU. Moreover, limited scatter correction, beam hardening effects, and variability in acquisition protocols introduce non-linearities in the reconstructed attenuation values [28, 29]. These differences in image appearance and soft-tissue definition can be appreciated in Figure 2.2, which provides a direct visual comparison of axial, sagittal, and coronal views acquired with CT and CBCT in the same patient.

Voxel intensities in CBCT may deviate from true attenuation coefficients and cannot be assumed to be quantitatively comparable to CT-derived HU without dedicated calibration procedures [30]. Thus, while CBCT provides adaptive in-room imaging and optimal spatial resolution for high-contrast structures such as bone, it is generally less reliable than conventional CT for quantitative tissue characterization and soft-tissue contrast assessment. Therefore, CBCT has gained widespread adoption in image-guided radiotherapy (for patient positioning and monitoring before, during or after the dose delivery), interventional procedures, dental and maxillofacial imaging, where accurate 3D anatomical visualisation is required in a relatively limited anatomical region [1, 31, 32].

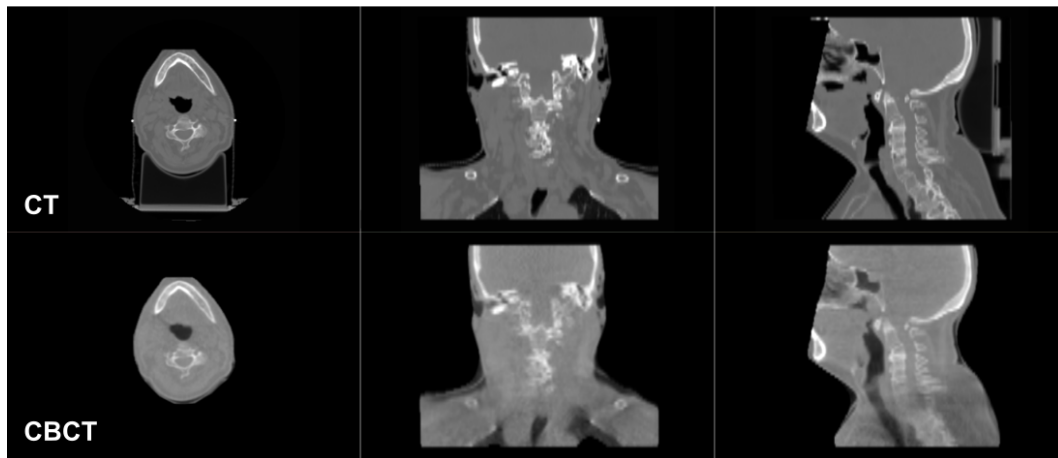


Figure 2.2: A comparison of axial, sagittal and coronal views from CT and CBCT scans. The patient was selected from the public SynthRAD2025 challenge dataset [33].

Overall, CT and CBCT represent natural technological evolutions of projection radiography. By combining rotational acquisition geometries with advanced reconstruction algorithms, these modalities enable true volumetric imaging, significantly expanding diagnostic capabilities and supporting both clinical decision-making and translational research applications.

2.1.3 Positron Emission Tomography

PET is a quantitative nuclear medicine imaging modality designed to visualise and measure physiological and biochemical processes *in vivo*. Unlike purely anatomical techniques, PET provides functional information by mapping the spatial distribution of radiolabelled molecules (also named radiotracers) within the body. The functional image contrast in PET is primarily determined by the biological behaviour of the radiotracer rather than by the radionuclide itself. While the radionuclide influences physical properties such as half-life and positron energy, the tracer molecule defines the physiological process being imaged and therefore the resulting contrast between tissues. The technique is widely applied in oncology, neurology, and cardiology, where early detection of metabolic alterations is clinically relevant [21, 34–36].

PET is based on the injection of a biologically active compound labelled with a positron-emitting radionuclide. Commonly used radionuclides include carbon-11 (^{11}C), nitrogen-13 (^{13}N), oxygen-15 (^{15}O), and fluorine-18 (^{18}F). These isotopes are characterised by relatively short physical half-lives, which allows repeated examinations while limiting radiation exposure [21, 35, 36]. Following the β^+ decay process, the emitted positron travels a short distance in tissue before annihilating with an electron. This annihilation event produces two photons emitted in nearly opposite directions. Coincidence detection of these photon pairs by a ring of scintillation detectors enables localisation of the annihilation event along a line of response, forming the basis of tomographic reconstruction [35].

The most widely employed PET radiotracer is 2-deoxy-2- ^{18}F fluoro-D-glucose (FDG), an analogue of glucose. FDG accumulates in tissues with high glycolytic activity due to increased expression of glucose transporters and hexokinase activity, features commonly observed in malignant cells. For this reason, FDG-PET has become integral to tumour detection, staging, treatment response assessment, and recurrence monitoring in many solid malignancies [37]. An illustrative example of a whole-body PET is provided in Figure 2.3.

A relevant strength of PET lies in its ability to perform true whole-body imaging within a single examination. This feature is particularly important in oncological practice, where systemic staging and the detection of distant metastases are crucial for treatment planning and prognostic stratification [37], reducing sampling bias and improving disease characterisation.

PET systems measure the *in vivo* radioactivity concentration within tissues (typically expressed in kBq/mL), which reflects the local concentration of the injected radiotracer. However, absolute activity values are influenced by several practical factors, most notably

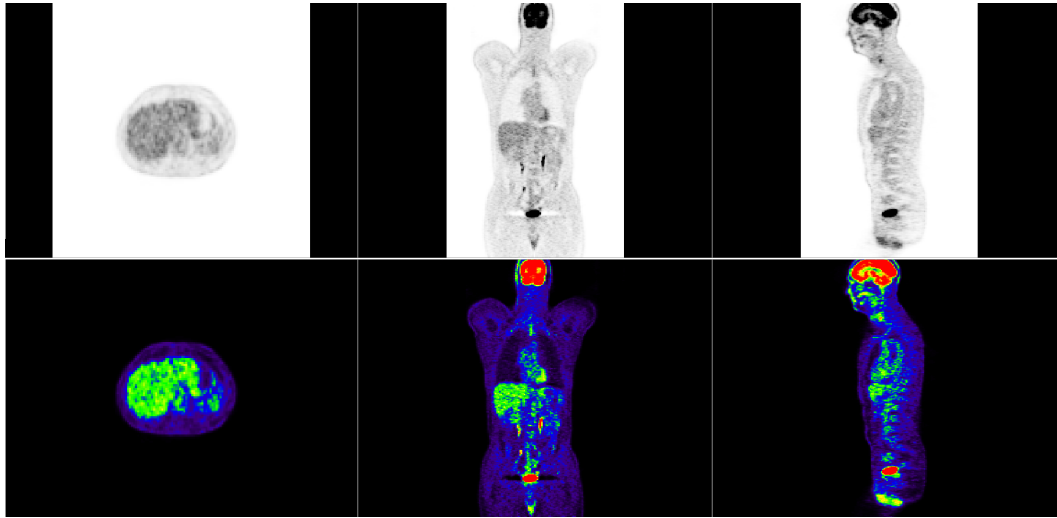


Figure 2.3: Whole-body ^{18}F -FDG PET acquisition displayed using two different intensity mapping schemes. Although the underlying quantitative activity distribution remains unchanged, alternative colour scales are applied in order to enhance visual interpretation of tracer uptake patterns. The example is derived from the MICCAI 2018 ^{18}F -FDG PET Radiomics Risk Stratifiers in Head and Neck Cancer dataset [38].

the injected dose and patient body size. In order to mitigate these sources of variability and enable meaningful inter- and intra-subject comparisons, semi-quantitative metrics such as the Standardised Uptake Value (SUV) are widely used. The SUV normalises tissue activity concentration to injected dose and a body-size parameter (e.g., body weight), thereby providing a relative index of tracer uptake. In addition to such semi-quantitative measures, more advanced kinetic modelling approaches can be applied to dynamic PET data to estimate physiological parameters including metabolic rates, receptor density and perfusion [39, 40].

Although PET has well-established clinical value, there are several limitations that must be carefully considered. A primary concern relates to exposure to ionising radiation. The radiation burden is derived from the injected radiotracer and, in hybrid systems such as PET/CT, from the additional CT acquisition. While the effective doses are generally within the accepted limits for diagnostic procedures, cumulative exposure may be a relevant consideration, particularly in younger patients or in cases where repeated follow-up examinations are required [41].

From a technical perspective, PET has a lower intrinsic spatial resolution compared with high-resolution CT or MRI. This limitation is partly attributable to fundamental physical factors, including the finite positron range in tissue and the non-collinearity of annihilation photons [40]. Consequently, small lesions may be affected by partial volume effects, potentially influencing quantitative accuracy.

Finally, PET imaging is associated with logistical and economic constraints. The limited half-lives of the radionuclides most frequently used necessitate either proximity to a cyclotron facility or access to a specialised radiopharmacy. This infrastructure requirement increases operational costs and may limit availability in certain regions [42]. Together, these factors

highlight the importance of appropriate patient selection and careful integration of PET findings within the broader clinical context.

2.1.4 Magnetic Resonance Imaging

Since its clinical introduction, MRI has become a pillar of modern diagnostic imaging. It is a non-invasive diagnostic technique that enables high-resolution visualisation of internal anatomical structures and, in selected applications, physiological and functional processes. In contrast to imaging modalities based on ionising radiation, such as CT or PET, MRI exploits the physical principles of nuclear magnetic resonance and therefore patients are not exposed to ionising radiation.

Specifically, MRI is based on the behaviour of atomic nuclei with non-zero spin, most notably the hydrogen nucleus (proton), when placed in a strong external magnetic field. In clinical systems, the static magnetic field strength –measured in Tesla (T)– typically ranges from 1.5 to 3 T, although higher-field systems (e.g. 7 T, and 12 T) are increasingly being used in research settings [21].

Once positioned within the bore of the scanner, the patient is subject to an applied static magnetic field, the effect of which is to induce partial alignment of hydrogen nuclei along the field direction. These nuclei exhibit a precession at a frequency known as the Larmor frequency, which is proportional to the magnetic field strength according to the relationship $f = \gamma B_0$, where γ is the gyromagnetic ratio. For hydrogen nuclei ($\gamma = 42.58$ MHz/T), the Larmor frequency is approximately 64 MHz at 1.5 T and 128 MHz at 3 T, reaching nearly 300 MHz in high-field systems around 7 T [43]. The application of a radiofrequency (RF) pulse at this particular frequency results in the perturbation of the equilibrium alignment of the spins. Following the interruption of the RF excitation, the nuclei revert to equilibrium via relaxation processes that are distinguished by two primary time constants: (i) longitudinal relaxation (T_1) and (ii) transverse relaxation (T_2). The MR signal is sampled after a delay known as the echo time (TE), defined as the interval between the RF excitation and the peak of the measured signal echo. Together with the repetition time (TR), TE plays a key role in determining the relative weighting of T_1 and T_2 relaxation effects in the reconstructed images. The spatial encoding of the emitted signals is achieved by superimposing magnetic field gradients, thereby enabling the reconstruction of cross-sectional images via Fourier-based algorithms [44].

The differences in proton density and relaxation properties among tissues result in intrinsic image contrast. More specifically, the detected MRI signal reflects the decay of transverse magnetisation following RF excitation. This decay can be described by an exponential relationship:

$$M_{xy}(t) = M_{xy}(0)e^{-t/T_2} \quad (2.3)$$

where $M_{xy}(t)$ denotes the transverse magnetisation at time t and T_2 represents the transverse relaxation time constant. In practice, additional magnetic field inhomogeneities introduce further dephasing effects, resulting in an effective relaxation constant T_2^* , such that:

$$M_{xy}(t) = M_{xy}(0)e^{-t/T_2^*}. \quad (2.4)$$

Because MRI signal acquisition occurs at the echo time (TE), tissues with very short transverse relaxation times may lose most of their signal before measurement. This phenomenon is particularly evident in cortical bone, tendons, and ligaments, which exhibit extremely short T_2 or T_2^* values due to their dense structure, low proton density, and restricted molecular motion. As a consequence, these tissues appear as signal voids in conventional MRI sequences, since the transverse magnetisation decays almost completely before the signal can be effectively sampled. Dedicated techniques such as ultrashort echo time (UTE) imaging are therefore required to capture signal from such short- T_2 tissues [44].

These relaxation characteristics also explain why tissues with very low proton density and extremely short T_2 values, such as cortical bone, appear dark in most conventional MRI sequences. Therefore, in the context of T_1 -weighted (T_1w) sequences, fluid is characteristically hypointense, fat is hyperintense, and intravascular flow may appear bright depending on sequence parameters. In contrast, T_2 -weighted (T_2w) sequences characterise fluid as hyperintense, while fat and rapidly flowing blood typically manifest reduced signal intensity or flow-related signal loss [45]. These signal differences can be directly appreciated in the example selected from the publicly available IXI dataset¹ shown in Figure 2.4.

Thus, MRI offers several important advantages that explain its central role in contemporary diagnostic imaging. The absence of ionising radiation is the most significant benefit, making MRI particularly well suited to younger patients and clinical scenarios requiring repeated follow-up examinations, such as oncological surveillance or monitoring of chronic neurological diseases.

Furthermore, given its superior soft tissue contrast and multiplanar capability, MRI is particularly valuable in a wide range of medical fields. In the field of neurological imaging, for instance, it facilitates detailed differentiation between grey and white matter, as well as the sensitive detection of demyelinating disease, tumours, infection, and vascular pathology. Similarly, in the field of cardiovascular medicine, MRI enables comprehensive evaluation of cardiac morphology, ventricular function, myocardial perfusion, tissue characterisation, and great vessel anatomy [21, 46].

An important characteristic of MRI is the large number of acquisition parameters that can be adjusted to modulate image contrast and sensitivity to different tissue properties. By varying parameters such as TR, TE, flip angle, and gradient timing, different pulse sequences can be designed to emphasise specific physical mechanisms, including T_1 and T_2 relaxation, proton density, diffusion, perfusion, or susceptibility effects. This flexibility makes MRI a highly versatile imaging modality capable of investigating multiple aspects of tissue structure

¹Available at: <https://brain-development.org/ixi-dataset/>. Last accessed: October 27, 2026

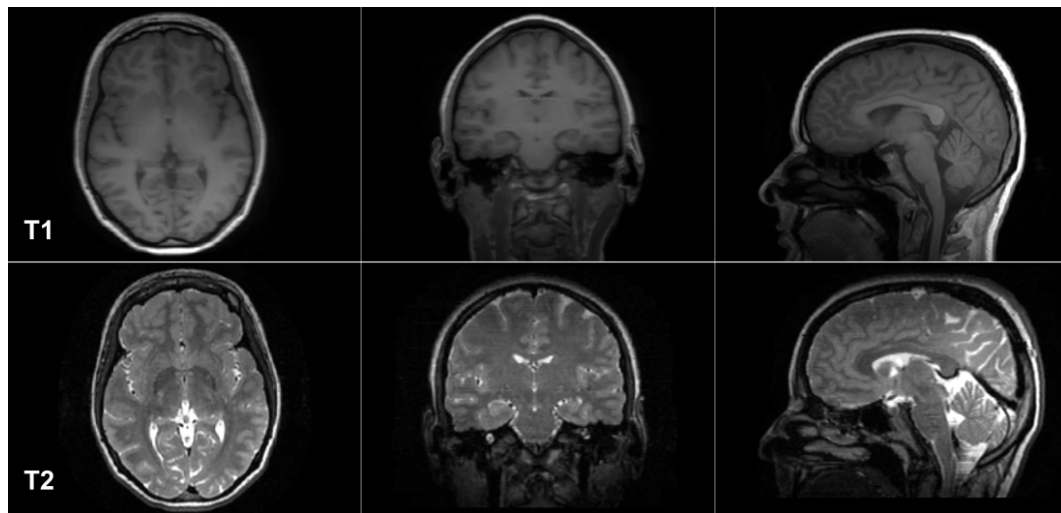


Figure 2.4: Comparison of axial, sagittal and coronal views from brain T_1w and T_2w MRI acquisitions of the same subject. In T_1w images, fluid appears hypointense and fat hyperintense, whereas in T_2w images fluid is hyperintense and fat demonstrates relatively lower signal intensity.

and physiology within the same examination. Several specialised imaging techniques have been developed by exploiting these contrast mechanisms.

Diffusion-weighted imaging (DWI) is highly sensitive to acute cerebral ischaemia and is integral to modern stroke protocols [47]. Furthermore, functional MRI (fMRI), typically based on blood oxygen level-dependent (BOLD) contrast, allows mapping of task-related or resting-state brain activity and has significantly advanced neuroscience research [48]. Magnetic resonance angiography (MRA) permits non-invasive visualisation of vascular structures without the need for catheterisation. Dynamic contrast-enhanced MRI (DCE-MRI) and other quantitative approaches provide insights into tumour perfusion and micro-environmental characteristics, contributing to oncological assessment and therapy monitoring [45].

Whole-body MRI has gained increasing acceptance for staging and surveillance in oncological practice, particularly in the evaluation of skeletal metastases [21]. In breast imaging, MRI is recommended for high-risk populations, including carriers of pathogenic *BRCA1* and *BRCA2* variants, due to its superior sensitivity relative to mammography in dense breast tissue [49].

Despite its considerable strengths, MRI is associated with specific practical and technical limitations. Examination times are generally longer than those of CT, which may pose challenges in acutely ill, claustrophobic, or uncooperative patients. Despite the reduction in acquisition times brought about by the development of accelerated imaging techniques, the presence of motion artefacts resulting from respiration, cardiac activity, or voluntary movement persists as a potential cause of image degradation. Furthermore, while respiratory and cardiac gating techniques mitigate these effects, they may further prolong the examination.

Safety considerations also impose restrictions. Patients with specific implanted medical devices, particularly those that are not certified as MR-conditional, as well as individuals with

ferromagnetic foreign bodies, may be deemed ineligible for scanning due to risks associated with the strong static magnetic field and gradient switching. Careful screening is therefore mandatory prior to every examination [21].

Economic factors must also be acknowledged. MRI systems require substantial capital investment and ongoing maintenance, including cryogenic infrastructure for superconducting magnets. Consequently, availability may be limited in certain healthcare settings.

In addition, although gadolinium-based contrast agents significantly enhance lesion detection and characterisation, there are potential risks associated with their use. Mild adverse reactions, including nausea, vomiting, pain at the injection site, headache, and dizziness, have been reported following their injection. The deposition of gadolinium in various tissues, including bone, kidney, brain, skin, and lymph nodes, has been evidenced by various studies. The specific form of these depositions is still unknown, and currently, there is no evidence of tissue damage directly associated with gadolinium accumulation. Nevertheless, studies have reported certain disorders related to gadolinium retention, including hypersensitivity reactions, nephrogenic systemic fibrosis, and gadolinium deposition disease [50].

Taken together, these limitations do not diminish the substantial clinical value of MRI, but they highlight the importance of appropriate patient selection and protocol optimisation in routine practice.

2.1.5 Application of multimodal imaging: image-guided and adaptive radiotherapy

Multimodal imaging combines anatomical, metabolic and functional information obtained from different imaging techniques in order to support diagnosis and therapy. An example of this integration is presented in Figure 2.5, showing PET, CT, and MRI acquisitions of a tumour at the base of the tongue, including both individual and fused representations. In radiation oncology, this integration is essential: imaging defines the target, guides dose calculation, verifies treatment delivery, and increasingly supports adaptation over time.

Conventional external-beam radiotherapy planning is CT-based. A planning CT acquired with the patient in treatment position provides a three-dimensional map of electron density, which is required for accurate dose calculation. Consequently, CT remains the geometric and dosimetric reference standard. However, CT alone may be insufficient for precise delineation of the gross tumour volume (GTV), clinical target volume (CTV), and organs at risk (OARs), particularly in anatomical regions characterised by limited soft-tissue contrast. Therefore, an MRI scan is routinely registered to the planning CT to improve tumour and OAR definition, especially in brain, head-and-neck, prostate, and gynaecological malignancies. Owing to its superior soft-tissue contrast and multiplanar capability, MRI has been shown to reduce inter-observer variability in contouring and to improve target definition [1, 20, 51].

MR-only radiotherapy workflows. The increasing reliance on MRI for target definition has led to the development of MR-only radiotherapy workflows, in which MRI constitutes

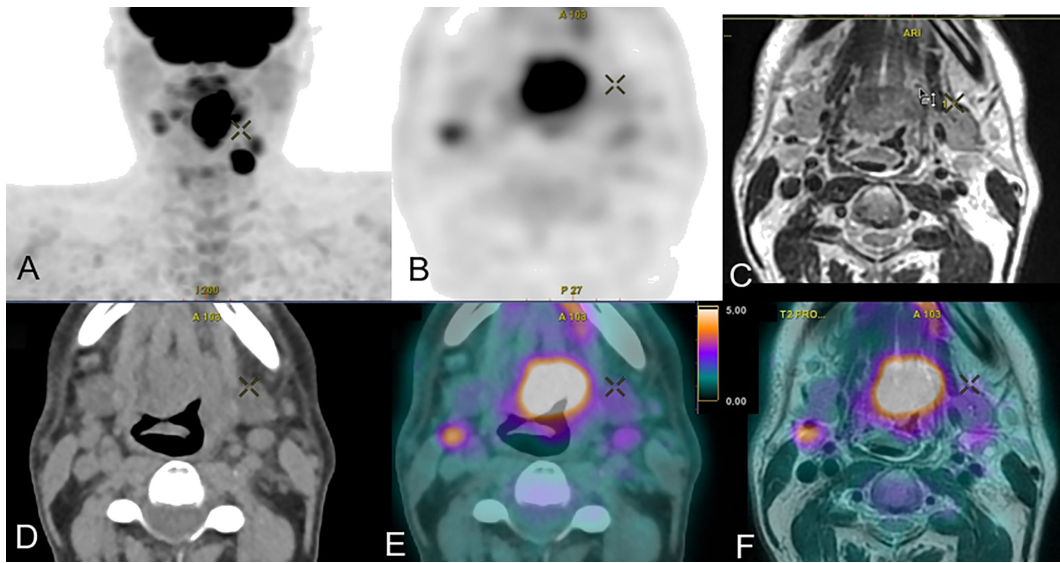


Figure 2.5: Multimodal imaging of a tumour located at the base of the tongue combining PET, MRI, and CT. The figure includes (A) a frontal PET maximum-intensity projection, (B) an axial FDG-PET image, (C) an axial T2-weighted MRI scan, (D) an axial CT image, together with fused representations obtained from (E) PET/CT and (F) PET/MRI. Taken from Decazes et al. [52], licensed under the terms of the Creative Commons Attribution License (CC BY 4.0).

the primary imaging modality for both delineation and treatment planning. In contrast to the conventional CT–MR workflow, where CT provides electron density information and MRI is used as a secondary modality for contouring, the MR-only paradigm eliminates the planning CT entirely.

The MR-only approach is typically supported by three principal arguments. First, the removal of CT–MR image registration avoids the introduction of systematic geometrical uncertainties. Registration errors arising at the planning stage propagate to every treatment fraction and therefore represent systematic spatial deviations. Reported uncertainties in CT–MR registration are in the order of a few millimetres, depending on anatomical site and registration strategy. In anatomical regions where the target and the structures used for patient positioning cannot be simultaneously registered with high accuracy, the resulting ambiguity may affect target coverage or organ-at-risk sparing. Indeed, for intracranial tumours, where rigid anatomy facilitates accurate registration, the incremental geometric benefit may be limited. In contrast, in regions such as the head and neck, where anatomical configuration can differ substantially between imaging sessions, or where positioning structures and targets cannot be simultaneously aligned with high precision, MR-only planning may reduce ambiguity and improve spatial consistency. In such scenarios, MR-only planning has the potential to reduce systematic spatial uncertainty and may justify margin reduction, particularly for small targets or highly conformal dose distributions [51].

Secondly, MR-only workflows are often considered from a logistical perspective. The elimination of the CT acquisition process serves to streamline the patient pathway, thus removing the necessity for one imaging session. However, this may be particularly relevant

in adaptive workflows driven by MRI, where repeated MR acquisitions are required and thus the time to acquire imaging is prolonged. Furthermore, current implementations require additional quality assurance (QA) steps, and the resource savings resulting from the elimination of CT have not yet been consistently demonstrated in clinical practice [51].

Thirdly, the exclusion of CT reduces exposure to ionising radiation. Although this reduction is modest compared with the therapeutic dose delivered during radiotherapy, it represents a conceptual advantage of the MR-only approach.

The primary technical challenge associated with MR-only planning stems from the inherent limitations of MRI in providing direct electron density information, which is indispensable for accurate dose calculation. Consequently, Synthetic Computed Tomography (sCT) images must be generated from the MR scan to enable dose calculation.

Several methods have been developed for sCT generation, including atlas-based approaches, voxel-based classification techniques, deformable image registration (DIR) pipelines, and more recently deep learning models. Reported dosimetric differences between CT-based and sCT-based planning are typically within 1% for many anatomical sites, indicating that clinically acceptable dose accuracy can be achieved [1, 51, 53]. Nevertheless, the sCT generation process is inherently complex and may be unsuccessful in the presence of atypical anatomy, surgical alterations, or metallic implants. Given that many sCT generation pipelines, in particular the deep learning-based, operate as partially opaque or “black-box” processing chains, rigorous patient-specific quality assurance is essential before their integration into clinical practice. In order to address this need, dedicated workflows have been introduced to quantify model uncertainty in sCT generation, alongside the development of independent validation models specifically designed to evaluate the reliability and accuracy of the generated sCT images [54, 55].

In addition, MR-only workflows are closely linked to the broader development of adaptive radiotherapy. In MR-driven adaptive strategies, anatomical changes observed during treatment necessitate re-optimisation. In this context, three options are available: (i) perform a deformable registration of the original planning CT, (ii) acquisition of a new CT, or (iii) generation of a new sCT directly from the adaptive MR scan. The latter option is inherently aligned with an MR-only philosophy and may facilitate streamlined adaptation.

Consequently, MR-only radiotherapy planning represents a logical extension of the increasing centrality of MRI in radiation oncology and constitutes a key enabling technology for fully MRI-driven treatment strategies. Its principal potential benefit lies in the reduction of systematic geometric uncertainties introduced by CT–MR registration, particularly in anatomical regions where accurate registration is difficult. While dosimetric feasibility using sCTs has been convincingly demonstrated, widespread adoption requires rigorous quality assurance frameworks and careful evaluation of workflow efficiency [1, 3, 51, 53].

CBCT Adaptive Radiotherapy. While multimodal imaging is essential at the planning stage, its role extends throughout the entire treatment course via the image-guided adaptive radiotherapy (IGART).

Radiotherapy is typically delivered in multiple fractions, and anatomical conditions can vary from day to day due to factors such as organ motion, tumour regression, weight loss, or changes in organ filling. Interfraction anatomical changes may progressively compromise the validity of the original treatment plan. Therefore, IGART aims to address this limitation by modifying the treatment plan in response to observed changes. Consequently, the necessity to employ imaging prior to (and, in certain systems, concurrently with) irradiation emerges as a pivotal element in ensuring precise target positioning [32, 56].

CBCT integrated into modern linear accelerators (LINACs), has become a standard tool for volumetric verification. By acquiring three-dimensional images immediately prior to treatment, CBCT allows alignment of the patient with respect to bony anatomy or, in favourable cases, soft-tissue structures. This daily imaging enables margin reduction and improves geometric precision [56].

Typically, CBCT-based adaptive workflows employ deformable image registration to propagate contours, estimate anatomical variations, and accumulate delivered dose over time.

Although CBCT image quality and quantitative reliability are inferior to diagnostic CT, recent developments in scatter correction, calibration, and sCT generation from CBCT scan have improved the feasibility of dose recalculation directly on daily images [1, 57]. These advances enable both offline adaptation (replanning after several fractions) and, in selected settings, online adaptation immediately prior to irradiation.

Reported dosimetric discrepancies between planning CT and CBCT-derived sCT dose calculations are generally within 1-2% for several anatomical sites, with gamma pass rates exceeding clinically accepted thresholds under standard criteria, suggesting that clinically acceptable accuracy can be achieved in selected scenarios [1, 57]. Nevertheless, performance is influenced by anatomical region, presence of metallic implants, truncation artefacts, and the robustness of the underlying correction or learning-based model.

From a practical perspective, the viability of CBCT-based sCT workflows depends not only on dosimetric accuracy but also on computational efficiency and integration within the clinical timeline. Offline adaptation, performed after several fractions, is generally compatible with current processing times. Online adaptation, in contrast, requires rapid sCT generation, contour propagation, plan re-optimisation, and quality assurance within a limited time window, typically on the order of minutes. Recent advances in GPU-accelerated deep learning pipelines have reduced inference times substantially, making near real-time sCT generation technically feasible; however, robust patient-specific quality assurance and automated consistency checks remain essential before widespread routine adoption.

Thus, in comparison with MR-only adaptive strategies, CBCT-based adaptation benefits from the availability of CBCT on conventional LINACs and the absence of a requirement for dedicated MRI-integrated systems. This broad accessibility enhances its translational potential across diverse clinical environments. Nevertheless, the intrinsic limitations of CBCT

image quality necessitate sophisticated correction or synthesis pipelines, which introduce additional layers of technical complexity and validation requirements.

However, CBCT-based adaptive radiotherapy is a pragmatic and widely implementable approach to daily anatomical adaptation. The generation of sCT from CBCT scans is fundamental to enabling accurate dose recalculation and plan adaptation. While the current state of evidence supports its dosimetric feasibility in several anatomical sites, long-term clinical validation, standardised quality assurance frameworks, and comprehensive evaluation of workflow efficiency remain critical for its consolidation as a routine adaptive strategy [1, 57].

Technical Fundamentals

3.1 Deep Learning in Medical Image Analysis

Deep Learning (DL) has fundamentally transformed medical image analysis by enabling automated, end-to-end learning of hierarchical feature representations from large-scale imaging datasets. Unlike traditional Machine Learning (ML), which relies on hand-crafted features, DL models can automatically extract relevant patterns from raw images. This capability improves the accuracy and efficiency of tasks such as medical image classification, segmentation, detection, and synthesis, facilitating the precise identification of anatomical structures and supporting automated detection and classification of diverse conditions [58].

At the core of this revolution lies the Convolutional Neural Network (CNN), a class of feed-forward neural networks specifically designed to exploit the spatial structure of images [58, 59]. A CNN (Figure 3.1) operates through a stack of differentiable building blocks, each applying a different convolutional filter to the incoming feature maps. The principal components are:

Convolutional layers. Each convolutional layer applies a set of learnable filters (kernels) to the input feature map via a discrete convolution operation. Given an input feature map $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$, where H , W , and C denote height, width, and number of channels respectively, a kernel $\mathbf{K} \in \mathbb{R}^{k_h \times k_w \times C}$ computes a dot product over local receptive fields, adding a bias term b :

$$(\mathbf{X} * \mathbf{K})_{i,j} = \sum_{m=0}^{k_h-1} \sum_{n=0}^{k_w-1} \sum_{c=0}^{C-1} \mathbf{X}_{i+m, j+n, c} \mathbf{K}_{m,n,c} + b, \quad (3.1)$$

This operation detects low-level patterns (edges, corners, textures) in shallow layers and progressively more abstract structures (shapes, organs, lesions) in deeper layers. Consequently, each filter produces a feature map and the collection of feature maps across all filters constitutes the output of the layer [58, 60, 61].

Non-linear activation functions. Following each convolution, a non-linear activation function is applied element-wise, breaking the linearity of the network and enabling the approximation of complex, non-convex decision boundaries. For instance, the Rectified Linear Unit (RELU), defined as:

$$\sigma(x) = \max(0, x) \quad (3.2)$$

is one of the most widely adopted activation functions due to its computational simplicity and its ability to mitigate the vanishing gradient problem during back-propagation [58].

Pooling layers. Spatial down-sampling is achieved through pooling operations (typically using max pooling), which partition the feature map into local neighbourhoods and retain the maximum activation within each region. Therefore, this operation reduces the spatial resolution of the feature maps [58, 60].

Fully connected layers and output heads. Following several alternating convolutional and pooling stages, the high-dimensional feature maps are flattened into a one-dimensional (1D) vector and processed by one or more fully connected layers. A final layer produces class probabilities for classification tasks, often using the Softmax function to produce class probabilities. Alternatively, regression heads or dense output layers can be used for other prediction objectives [58, 61].

Batch normalisation and regularisation. Batch normalisation layers have become a standard feature of recent network architectures, typically positioned between the convolution and activation steps. These layers normalise the distribution of activations across a mini-batch, accelerating convergence and enabling the adoption of higher learning rates. Dropout regularisation, which randomly deactivates neurons during training, reduces overfitting on limited medical datasets [58, 59].

The complexity of the feature hierarchy that a network can learn is determined by the depth and width of the network [62]. Several canonical architectures have become benchmarks in the field. AlexNet demonstrated the effectiveness of training deep networks with data augmentation and dropout on large image collections. Visual Geometry Group (VGG) networks replaced large kernels with stacked 3×3 convolutions, showing that depth was a critical determinant of representational power. Furthermore, ResNet introduced residual connections, which bypass one or more layers, enabling networks with hundreds of layers to be trained by allowing gradients to flow directly through the shortcut and thereby addressing the vanishing gradient problem [58, 61].

3.1.1 Image Classification and Object Detection

Image classification, defined as the process of assigning a label to an entire image, was the pioneering task in which CNN exhibited remarkable performance in the field of medical

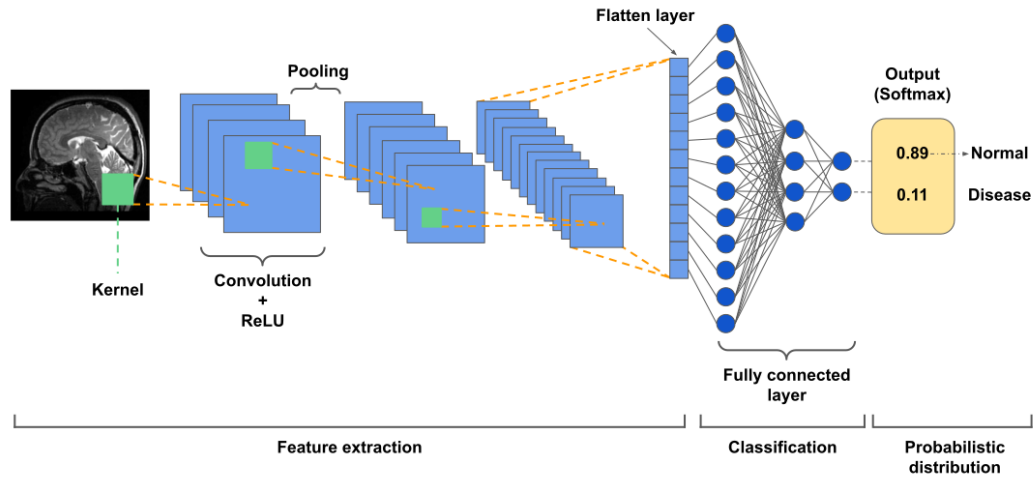


Figure 3.1: Overview of a CNN, outlining its hierarchical structure. The network comprises convolutional layers that extract spatial features, non-linear activation functions that introduce model flexibility, pooling layers that reduce dimensionality, fully connected layers to integrate features, and output heads to generate predictions.

imaging. This included the detection of diabetic retinopathy in fundus images and the classification of dermatological lesions, both of which attained diagnostic accuracy comparable to or exceeding that of board-certified clinicians [59, 61]. The training of classification networks involves the minimisation of a loss function $\mathcal{L}$ (e.g. binary, categorical, and sparse cross-entropy, hinge loss) between predicted class probabilities and ground-truth labels via stochastic gradient descent and back-propagation.

The transfer learning paradigm, which involves initialising the network weights from models pre-trained on extensive datasets of images (such as ImageNet) has been demonstrated to enhance the convergence process and improve the generalisation capability of the DL models when faced with small, labelled medical datasets [61].

Object detection extends classification to localisation, requiring the network to simultaneously predict the class and spatial extent (bounding box) of multiple objects within an image. Region-based CNN (R-CNN) and its descendants, Fast R-CNN and Faster R-CNN, decompose detection into region proposal and classification stages, while single-stage detectors (such as YOLO) predict bounding boxes and class scores from a single forward pass, reducing inference time at the cost of some accuracy in detecting small objects. In particular, it has been observed that although large-scale variants (e.g. YOLOv5x) can detect small objects more effectively, the number of missed detections tends to increase as the model scale decreases, highlighting the trade-off between model size and detection performance [63]. In medical imaging, this capability is employed for tasks such as pulmonary nodule detection in Computed Tomography (CT), polyp localisation in colonoscopy video, human osteosarcoma cell detection, and white blood cells detection [60, 61].

3.1.2 Semantic Segmentation

Semantic segmentation, defined as the process of assigning a class label to each pixel (or voxel, in three dimensions) of an image, represented a significant advancement in the field of medical image analysis. This task is considered to be one of the most clinically consequential and challenging in this domain. The applications of semantic segmentation include, for instance, the organ delineation for radiotherapy planning, tumour volume measurement, and lesion tracking [64].

The conceptual precursor to modern segmentation networks is the Fully Convolutional Network (FCN), which repurposed CNNs by replacing fully connected layers with convolutional layers, thereby enabling dense prediction at the pixel level. However, repeated pooling operations in the encoder path of the FCN cause progressive loss of spatial resolution, leading to coarse segmentation masks. The introduction of the encoder-decoder paradigm addressed this limitation: the encoder progressively downsamples the feature maps to capture high-level semantic context, while a symmetric decoder path gradually upsamples the feature maps back to the original image resolution to restore precise spatial localisation [64].

The U-Net architecture [65] became the gold standard for medical image segmentation [66]. Its original design consists of two principal branches (see Figure 3.2):

1. **Contracting path (encoder).** Each stage comprises two consecutive 3×3 convolutional layers, each followed by a RELU activation, succeeded by a 2×2 max pooling operation that halves the spatial dimensions. The number of feature channels is doubled at each down-sampling step. This results in the progressive encoding of increasingly abstract semantic information while spatial detail is being compressed.
2. **Expansive path (decoder).** Each stage applies a 2×2 transposed convolution (up-convolution) that doubles the spatial resolution and halves the number of channels. The resulting feature maps are concatenated with the corresponding feature maps from the contracting path via skip connections, which directly relay fine-grained spatial detail from encoder to decoder. Two subsequent 3×3 convolutions with RELU activations then refine the fused features.

The distinctive U-shaped topology and skip connections facilitate the integration of shallow, spatially rich encoder feature maps with deep, semantically rich decoder feature maps. This configuration enables U-Net to simultaneously capture global context and recover fine structural details, even when trained on limited datasets [65, 66].

Following its introduction, a significant number of derivatives of U-Net have been developed [66]. For instance, the UNet++ [68] redesigned the skip connections into nested, dense pathways. Each encoder feature map was enriched by a dense convolutional block before being fused with the corresponding decoder map, thereby reducing the semantic gap between encoder and decoder representations. The Attention U-Net incorporated attention gates on the skip connections, enabling the network to suppress irrelevant activations and focus on target structures [69]. Furthermore, 3D U-Net and V-Net extended the architecture

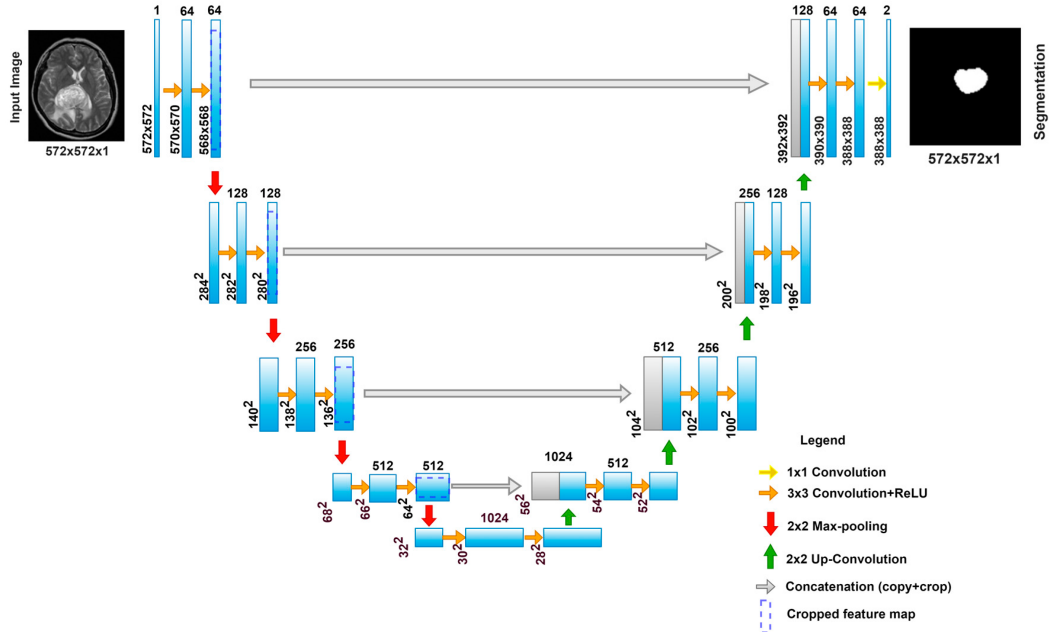


Figure 3.2: A basic U-Net architecture for brain tumour segmentation. Taken from Yousef et al. [67], licensed under the terms of the Creative Commons Attribution License (CC BY 4.0).

to volumetric data by replacing all two-dimensional (2D) operations with three-dimensional (3D) equivalents, which is particularly relevant for CT and Magnetic Resonance Imaging (MRI) segmentation, where inter-slice context may carry diagnostic significance [64]. In contrast, the nnU-Net framework focused on automating the design of the segmentation pipeline, systematically adapting preprocessing, network topology, training scheme, and post-processing to the properties of a given dataset. The robust performance of nnU-Net stems not from a novel network architecture, loss function or training scheme –hence the “nnU-Net”, an acronym for “no new net”– but rather, from the automation of the intricate process of manual workflow design through rule-based heuristics and empirical selection. Consequently, this facilitated immediate segmentation efficacy across a range of biomedical imaging tasks [70].

3.1.3 Image-to-Image Translation

In the medical context, Image-to-Image (I2I) translation denotes the generation of a target-domain image from a corresponding source-domain image, without necessitating additional acquisitions. Specifically, the formal problem consists of learning a mapping $G : X \rightarrow Y$ between a source domain X and a target domain Y , such that the synthesised image $\hat{y} = G(\mathbf{x})$ preserves the anatomical content of $\mathbf{x} \in X$ while reflecting the appearance and intensity statistics of the target modality [71].

Encoder–decoder architectures, particularly U-Net and its derivatives, have proven highly effective in this setting. The encoder progressively extracts hierarchical features from the input image, while the decoder reconstructs an output image with the desired intensity values. Skip connections remain crucial in this context, as they facilitate the transfer of spatially detailed information from the encoder to the decoder, thereby preserving fine anatomical structures during the synthesis process. For this reason, U-Net and its variants are frequently adopted as generator backbones in medical image synthesis frameworks [1, 3, 57].

While such architectures define the structural backbone of many synthesis models, generative frameworks are often required to learn the complex data distributions underlying medical images. Among these, the most influential paradigm is the Generative Adversarial Network (GAN) [1, 3, 57, 72].

A GAN consists of two neural networks trained simultaneously in an adversarial setting: a generator G , which produces synthetic samples intended to resemble real data, and a discriminator D , which attempts to distinguish between real and generated samples. The training process is characterised by the following objective [72]:

$$\min_G \max_D V(D, G) = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})} [\log(1 - D(G(\mathbf{z})))] \quad (3.3)$$

At convergence, the generator ideally approximates the underlying data distribution, preventing the discriminator from reliably distinguishing between real and generated samples [72, 73].

Several extensions have been proposed to address the training instabilities of the original GAN. The Wasserstein GAN (WGAN), for instance, replaced the Jensen–Shannon divergence with the Wasserstein distance to provide more stable gradients during training. Conditional GANs (Conditional Generative Adversarial Network (cGAN)) incorporate auxiliary information such as class labels, segmentation masks, or source-modality images, enabling geometry-aware image synthesis [74]. The Pix2Pix framework [73, 75] further specialised cGANs for supervised I2I translation, learning a mapping between paired images using a combination of adversarial and $\mathcal{L}_1$ reconstruction losses.

Beyond the adversarial objective, I2I translation models typically rely on additional loss functions to enforce structural consistency between the generated and reference images. In supervised settings with paired data, reconstruction losses such as the $\mathcal{L}_1$ or $\mathcal{L}_2$ distance between the generated image $\hat{\mathbf{y}}$ and the ground truth $\mathbf{y}$ are commonly employed to encourage pixel-wise fidelity [1, 57]. Furthermore, perceptual losses have been introduced to measure similarity in a deep feature space extracted from pre-trained CNN networks, enabling the preservation of higher-level structural characteristics that may not be fully captured by pixel-wise losses alone [53].

While the previously described approaches assume the availability of paired training data, in many clinical scenarios such aligned image pairs may be difficult to obtain. Therefore, CycleGAN [71] introduced a cycle-consistency constraint that allows learning mappings between two image domains without aligned image pairs. In this framework, two generators $G : X \rightarrow Y$ and $F : Y \rightarrow X$ are trained such that $F(G(\mathbf{x})) \approx \mathbf{x}$ and $G(F(\mathbf{y})) \approx \mathbf{y}$, preserving the structural content of the images through the translation cycle [71].

The combination of adversarial, reconstruction, and consistency losses therefore forms the training objective used to balance visual realism with anatomical fidelity in medical image synthesis.

Medical I2I translation gained considerable attention due to its direct clinical relevance [1, 73]. Typical applications include MRI-to-CT synthesis for MRI-only radiotherapy planning, CBCT-to-CT translation for adaptive radiotherapy, and intra-modality MRI translation (e.g. T1w-to-T2w), where the goal is to generate missing contrasts to reduce acquisition time [76]. Additional applications include PET synthesis from CT to improve accessibility and reduce acquisition costs [77], as well as super-resolution and motion correction tasks [78]. Further technical details concerning the specific I2I translation training techniques are provided in the article presented in Chapter 4 - Section 4.2 of this thesis.

Although CNNs remain central to medical image analysis and synthesis, more recent architectures have also been explored. The Vision Transformer (ViT) paradigm, for instance, adapt the self-attention mechanism originally developed for natural language processing to the image domain [79]. In this approach, an input image is divided into a sequence of fixed-size patches, each projected into a learnable embedding and processed through multi-head self-attention layers that model global contextual relationships between image regions. ViT-based models achieved competitive performance in classification, segmentation, and synthesis tasks, particularly when trained on large-scale datasets [79]. However, they typically require substantial computational resources and large amounts of training data [57].

More recently, Denoising Diffusion Probabilistic Models (DDPM) have emerged as a powerful alternative paradigm for generative modelling [57]. These models define a forward diffusion process that gradually corrupts an image $\mathbf{x}_0$ by adding Gaussian noise over T time steps until it becomes indistinguishable from random noise $\mathbf{x}_T \sim N(\mathbf{0}, \mathbf{I})$. A neural network, often based on a U-Net architecture, is then trained to iteratively reverse this process by predicting and removing the noise component at each step. Compared with GAN, diffusion models can produce images with higher diversity and fewer mode-collapse artefacts, albeit at the cost of increased computational complexity and cost during the training and inference [80]. However, their ability to generate high-fidelity images while preserving anatomical structures positions them as a promising frontier for future research in medical image translation and synthesis.

3.1.4 Evaluation Metrics

Quantitative evaluation plays a crucial role in assessing the performance and reliability of DL models in medical image analysis. Depending on the specific task (e.g. classification, segmentation, or synthesis), different evaluation metrics are employed to quantify the agreement between model predictions and reference data.

Classification metrics. In medical image classification, performance is commonly evaluated using metrics derived from the confusion matrix, which summarises the number of

true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). From these quantities, several indicators are computed, including accuracy, sensitivity (recall), specificity, and precision:

$$\begin{aligned} \text{Accuracy} &= \frac{TP + TN}{TP + TN + FP + FN}, & \text{Sensitivity} &= \frac{TP}{TP + FN}, \\ \text{Specificity} &= \frac{TN}{TN + FP}, & \text{Precision} &= \frac{TP}{TP + FP}. \end{aligned} \quad (3.4)$$

Sensitivity quantifies the proportion of correctly detected positive cases, whereas specificity measures the proportion of correctly identified negative cases. Precision indicates the proportion of predicted positive samples that are correctly classified. To balance the trade-off between precision and recall, the F1-score is often used, defined as the harmonic mean of precision and sensitivity:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}. \quad (3.5)$$

In clinical applications where class imbalance is common, the receiver operating characteristic (ROC) curve and its corresponding area under the curve (AUC) are frequently used to provide a threshold-independent assessment of classification performance [81].

Segmentation metrics. For segmentation tasks, the objective is to measure the spatial overlap between the predicted mask and the reference annotation. One of the most widely used metrics is the Dice Similarity Coefficient (DSC), defined as:

$$\text{DSC} = \frac{2|A \cap B|}{|A| + |B|}, \quad (3.6)$$

where A and B denote the sets of voxels belonging to the predicted and ground-truth regions, respectively. The Dice score ranges from 0 (no overlap) to 1 (perfect agreement) and is particularly suitable for evaluating medical segmentations where the target structures may occupy only a small fraction of the image volume.

Other commonly reported metrics include the Intersection over Union (IoU), also known as the Jaccard index, and distance-based measures such as the Hausdorff distance (HD), which quantifies the maximum boundary deviation between two segmentations. These complementary metrics provide insights into both volumetric overlap and boundary accuracy [81].

Image synthesis metrics. The quality of synthesised images in I2I translation tasks, where paired ground-truth data are typically available, is commonly evaluated using pixel- or voxel-wise image similarity metrics computed between the generated image and the corresponding reference image. Among the most widely adopted metrics in medical image synthesis are the Mean Absolute Error (MAE), the Root Mean Squared Error (RMSE), the Structural Similarity Index Measure (SSIM), and the Peak Signal-to-Noise Ratio (PSNR) [1, 57].

Error-based metrics such as MAE and RMSE quantify the absolute or squared differences between corresponding pixel or voxel intensities of the synthesised and ground-truth images,

thereby providing a direct measure of pixel- or voxel-wise intensity accuracy [1, 82]. While MAE measures the average absolute deviation, RMSE penalises larger errors more strongly due to the quadratic formulation:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (3.7)$$

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (3.8)$$

where y_i and $\hat{y}_i$ denote the pixel or voxel intensities of the ground-truth and synthesised images respectively, and N is the total number of voxels within the evaluated region of interest [1, 82].

The SSIM evaluates image similarity by jointly comparing luminance, contrast, and structural information between the two images. As a result, it is often considered more consistent with human visual perception and better suited to assess the preservation of anatomical structures [1, 82]:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (3.9)$$

where μ_x and μ_y denote the mean intensities of the two images, σ_x^2 and σ_y^2 their variances, and σ_{xy} their covariance.

The stabilisation constants are defined as:

$$C_1 = (k_1 L)^2, \quad C_2 = (k_2 L)^2 \quad (3.10)$$

where L represents the dynamic intensity range of the image, and $k_1 = 0.01$ and $k_2 = 0.03$ are small constants introduced to stabilise the division.

The PSNR, derived from the mean squared error between images, expresses the ratio between the maximum possible signal intensity and the reconstruction error in logarithmic decibel units. Higher PSNR values indicate lower reconstruction error and therefore higher image fidelity [1, 82]:

$$\text{PSNR} = 10 \log_{10} \left(\frac{\text{MAX}_I^2}{\text{MSE}} \right) \quad (3.11)$$

where MAX_I denotes the maximum possible voxel intensity value of the image.

Although these metrics provide useful quantitative indicators of synthesis quality, they remain generic similarity measures and may not fully capture the clinical suitability of the generated images. For clinical deployment, task-specific evaluation is increasingly recognised as the appropriate gold standard [1]. In radiotherapy, for instance, the quality of Synthetic Computed Tomography (sCT) images is not only assessed voxel-wise using metrics such as dose difference (DD) but also via more clinically relevant, task-specific metrics. These include the gamma index analysis, which combines dose and spatial agreement

criteria, dose-volume histogram (DVH) comparisons, and range shift (RS) evaluations in proton therapy. Notably, the choice of region-of-interest (ROI) and implementation details (2D vs 3D evaluation, dose thresholds) can influence metric outcomes, highlighting that clinical validation often requires context-specific, task-oriented assessments rather than solely generic image similarity measures [1].

Furthermore, the recent Metrics Reloaded framework [83] highlighted that the choice of evaluation metrics is itself a critical design decision and that inappropriate metric selection may lead to misleading conclusions about algorithm performance. In particular, the framework emphasises that validation metrics should act as proxies for the underlying domain objective and therefore must be selected in a problem-aware manner [83]. In particular, the framework introduces the concept of a *problem fingerprint*, a structured description of the task that includes aspects such as the clinical question, properties of the target structures, dataset characteristics, and the form of the algorithm output. Based on this fingerprint, appropriate metrics can be chosen while accounting for common pitfalls such as class imbalance, annotation uncertainty, or differences between pixel-, object-, and image-level predictions. This perspective highlights that reliable model validation often requires a combination of complementary metrics rather than relying on a single quantitative score [83].

3.1.5 Data Augmentation

DL models benefit substantially from diverse and representative training datasets in order to learn robust and generalisable feature representations. Nevertheless, the availability of labelled data in medical imaging is often limited due to several factors, such as the cost and time required for expert annotations, and the strict privacy regulations concerning the sharing of clinical data. Consequently, models trained on small datasets are particularly susceptible to overfitting, where the network learns dataset-specific characteristics rather than generalisable patterns [84–86]. Therefore, data augmentation became a widely adopted strategy for mitigating the effects of limited training data.

The main objective is to artificially expand the training dataset by applying label-preserving transformations to existing samples. Exposing the model to multiple variations of the same data, augmentation encourages the learning of invariant and more robust features, thereby improving model generalisation and reducing the risk of overfitting [84, 85].

In the domain of medical image analysis, augmentation techniques can broadly be categorised into two primary groups: (i) transformation-based and (ii) generative model-based. Transformation-based methods apply deterministic or stochastic operations directly to existing images. These include geometric transformations such as rotations, translations, scaling, flipping, and elastic deformations, as well as intensity-based transformations, such as noise injection, contrast adjustment, histogram perturbations, and blurring. These transformations aim to simulate plausible variations that may arise from differences in patient positioning, scanner hardware, acquisition protocols, or reconstruction parameters [85, 86].

Recently, generative approaches have been explored to produce synthetic training samples that reflect the underlying data distribution. Generative models such as GANs, and DDPM can be used to generate synthetic images that resemble real clinical data, thereby further increasing dataset diversity [84, 86].

Despite the benefits of augmentation pipelines, their design requires careful consideration. Some transformations may lead to anatomically implausible samples or distort clinically relevant structures. For instance, excessive geometric transformations or intensity perturbations may alter tissue characteristics or anatomical relationships in ways that are not physically meaningful. Moreover, if biases or systematic errors are present in the original dataset, augmentation procedures may unintentionally propagate or amplify these biases in the training data. For these reasons, augmentation strategies are typically designed with task-specific constraints and tailored to the imaging modality, anatomical region, and downstream learning objective [85, 86].

Thus, data augmentation represents an important practical technique for improving the robustness and generalisation capability of DL models in medical imaging, particularly in scenarios where data availability is limited. For this reason, a growing number of research efforts focused on the development of dedicated augmentation frameworks and tools designed to support reproducible and task-specific augmentation pipelines within medical imaging workflows [85].

3.1.6 Data Sharing and Privacy

In clinical practice, medical imaging data are typically generated and stored within individual healthcare institutions. Hospitals maintain local repositories, most commonly within Picture Archiving and Communication Systems (PACS), which are designed for clinical use rather than large-scale data sharing for Artificial Intelligence (AI) research.

As a consequence, imaging datasets are frequently fragmented across institutional silos. While combining these datasets could significantly improve the diversity and representativeness of training data, the aggregation of patient data across institutions is strongly constrained by privacy, ethical, and legal considerations. Indeed, regulatory frameworks such as the General Data Protection Regulation (GDPR) in Europe [5] impose strict requirements on the handling, processing, and sharing of personally identifiable health information [6–9].

Therefore, the external sharing of raw imaging data requires complex compliance procedures (e.g. Data Protection Impact Assessment) or extensive de-identification processes, such as anonymisation or de-facing. While these procedures aim to protect patient privacy, they may also reduce the informational value of the images or complicate downstream analysis workflows. This limitation poses a significant challenge for the development of robust AI models in medical real-world scenarios, as they often rely on large-scale and heterogeneous datasets to learn generalisable representations [87]. Indeed, models trained exclusively on data from a single institution often exhibit limited performance, particularly in terms of generalisation when evaluated on external datasets [88, 89]. This phenomenon is primarily caused by variations in imaging devices, acquisition protocols, reconstruction parameters,

and patient populations across institutions. Such variations introduce differences between training and deployment data distributions, a phenomenon commonly referred to as domain shift, which can significantly affect model performance during inference [88, 89]. Thus, training models exclusively on small, institution-specific datasets may lead to biased or incomplete representations of the underlying data distribution. At the same time, traditional centralised training often conflicts with legal restrictions and institutional policies governing patient data.

These limitations motivated the development of alternative collaborative training paradigms that enable multi-institutional model development without requiring direct data sharing. Among these approaches, Federated Learning (FL) emerged as one of the most promising solutions for privacy-aware AI in healthcare [18].

3.2 Principles and Applications of Federated Learning in Medical Imaging

FL is a distributed AI training paradigm designed to enable collaborative model training across multiple institutions without requiring the exchange of raw data. Instead of collecting data into a single repository, FL enables multiple institutions to collaboratively train a ML or DL model while retaining their datasets locally stored. Each participating client performs local optimisation and shares only model updates with a coordinating server, which aggregates these updates to construct a global model [19, 90]. Therefore, FL is particularly well suited to cross-organisational environments in which participating entities are autonomous institutions that retain full control over their data while contributing to a shared learning process.

However, this paradigm introduces several methodological challenges, which differ from conventional centralised training. In particular, federated systems must explicitly address heterogeneous data distributions, communication efficiency, partial client participation, and privacy-preserving aggregation. In the context of medical imaging applications, these challenges are particularly pronounced, as imaging data are often non-independent and identically distributed (non-IID) across institutions. This discrepancy can be attributed to variations in scanner vendors, acquisition protocols, patient demographics, and annotation practices. Consequently, these variations can significantly influence model convergence and generalisation in federated settings [19].

The categorisation of FL is determined by the distribution of data across participating clients. In horizontal FL, each client holds datasets with the same feature space but different samples. This is the most common configuration in medical imaging, where multiple hospitals collect the same imaging modality (e.g. CT, MRI, or X-ray) from different patient populations.

In contrast, vertical FL assumes that different institutions have complementary features describing the same individuals. For instance, one institution may store imaging data while

another holds genomic or clinical records for the same patients. Although vertical FL enables multimodal modelling, it is less common in practice because patient overlap between institutions is typically limited [15].

A second distinction concerns the scale and nature of participating clients. Indeed, FL is commonly divided into cross-device and cross-silo settings. Cross-device FL typically involves thousands of consumer devices, such as smartphones, each contributing small amounts of data. In contrast, medical imaging applications almost exclusively operate in a cross-silo FL setting, where a relatively small number of institutions (e.g. hospitals or research centres) collaborate to train models using local datasets [90].

In cross-silo environments, the training process typically follows a server–client architecture organised into synchronous communication rounds. During each round, a central server distributes the current global model to participating clients. Each client performs local training using its own dataset and sends the updated model parameters back to the server, which aggregates these updates to construct a new global model [90]. This workflow is presented in Figure 3.3.

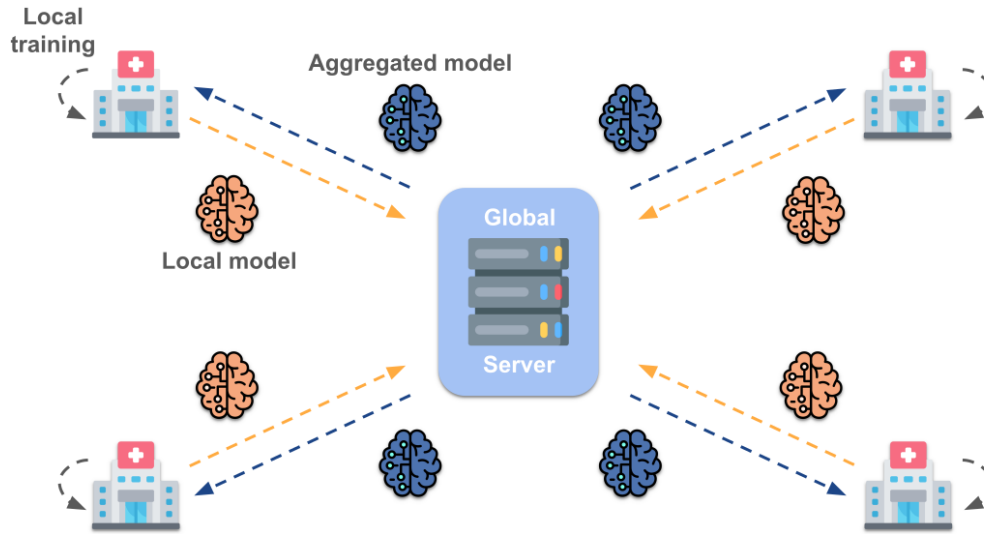


Figure 3.3: Representation of the cross-silo paradigm. Each client (data silo) performs local training on its private data. The resulting local model updates are transmitted to a central server, where they are aggregated to form an aggregated model. This model is subsequently redistributed to the participating institutions for the next training round.

The most widely used aggregation algorithm is the Federated Averaging (FedAvg) [91, 92], which computes a weighted average of client models based on dataset size. Formally, if client k holds n_k training samples and produces local model parameters w_k , the aggregated global model w_{t+1} can be expressed as

$$w_{t+1} = \sum_{k=1}^K \frac{n_k}{n} w_{t+1}^k \quad (3.12)$$

Although FedAvg performs effectively in various scenarios, its optimisation process may become unstable when data distributions differ substantially across clients (formally non-IID). Such heterogeneity can cause local model updates to diverge, thereby slowing convergence or degrading the quality of the global model [19]. In order to address this limitation, more advanced variants, such as the FedProx algorithm, have been proposed. This involves the introduction of a proximal regularisation term into the local optimisation objective [93]. Specifically, each client minimises a modified local objective function defined as:

$$h_k(w, w^t) = F_k(w) + \frac{\mu}{2} \|w - w^t\|^2, \quad (3.13)$$

where $F_k(w)$ denotes the original local objective function of client k , w^t represents the global model parameters at communication round t , and w denotes the local model parameters during optimisation. The coefficient $\mu \geq 0$ controls the strength of the proximal constraint. This proximal term penalises large deviations of local model parameters from the current global model, thereby limiting excessive client drift and improving optimisation stability in heterogeneous environments. When $\mu = 0$, the proximal term vanishes and FedProx reduces to the standard FedAvg algorithm.

FL has been increasingly explored across a broad range of medical imaging applications. Most existing federated medical imaging studies focused on predictive tasks, including image classification, lesion detection, and semantic segmentation [19]. Federated CNN have been successfully applied to brain tumour classification from MRI scans, dermatological lesion analysis, and lung disease detection from chest imaging. Several multi-institutional studies demonstrated that federated models can achieve performance comparable to centralised training while preserving patient data locally [15]. Segmentation tasks have also received significant attention. Federated implementations of architectures such as U-Net have been applied to tumour segmentation, organ delineation, and histopathological image analysis [19].

Despite the increasing interest in FL for predictive tasks, only a few studies explored generative applications, such as MRI reconstruction from undersampled or misaligned data [94–96].

Nevertheless, as discussed in Section 2.1.5, I2I translation remains a crucial task in medical imaging, particularly in radiotherapy planning and multimodal diagnostic workflows. Tasks such as MRI-to-CT, CBCT-to-CT translation, and intra-modality MRI translation can reduce scanning time, lower radiation exposure, and simplify clinical workflows. Indeed, centralised DL approaches based on U-Nets, GANs, or diffusion models have been extensively explored and achieved promising results for these tasks [1, 3, 57]. However, the application of the FL paradigm for I2I translation remained very limited.

Existing studies include FedMed-GAN and FedMed-ATL, both CycleGAN-based frameworks for intra-modality MRI translation [97, 98], as well as a personalised federated algo-

rithm for multi-contrast MRI synthesis employing site-specific latent representations [14]. In the context of CT synthesis, the author of this thesis investigated a federated approach for the MRI-to-CT synthesis using U-Nets, GANs and a Transformer architecture. The U-Net variants, which achieved superior performance, were evaluated against several aggregation strategies [99]. Furthermore, the author of this thesis proposed a cross-silo federated model based on the Pix2Pix architecture for Cone Beam Computed Tomography (CBCT)-to-CT synthesis in the head and neck region [100].

A notable aspect to emphasise is that although FL is frequently described as a privacy-preserving framework, concerns remain regarding its ability to offer absolute protection against information leakage. Given that model parameters or gradients are exchanged during the federated training process, potential malicious actors may seek to infer sensitive information from these updates. For instance, gradient inversion attacks attempt to reconstruct training samples directly from shared gradients, while membership inference attacks aim to determine whether a particular data point was included in the training dataset. Additional threats include model poisoning or Byzantine attacks, in which malicious clients manipulate local updates to degrade the global model or introduce targeted biases [19].

In order to mitigate the aforementioned risks, several mechanisms for enhancing privacy have been proposed. These include differential privacy techniques, which introduce controlled noise into model updates in order to prevent the reconstruction of sensitive data. Cryptographic approaches, such as secure multiparty computation and homomorphic encryption, allow model aggregation to be performed on encrypted parameters without revealing individual client updates [90, 101]. However, these protection mechanisms introduce important trade-offs. Differential privacy may reduce model accuracy, while cryptographic approaches increase computational overhead and communication costs. Designing secure yet efficient federated frameworks therefore remains an active research challenge, particularly in the context of computationally demanding generative models.

As discussed in the preceding paragraphs, FL represents a distinct paradigm shaped by constraints related to data ownership, privacy governance, and decentralised collaboration. While centralised DL focuses primarily on optimising model performance given access to a unified dataset, FL must balance several competing objectives simultaneously, including privacy preservation, robustness to heterogeneous data distributions, fairness across institutions, and communication efficiency. Thus, direct comparisons between federated and centralised models are not always straightforward. In particular, the performance of federated models is often influenced by non-IID data distributions across institutions or aggregation mechanisms that cannot be tailored to a single client. This emphasises the distinct challenges inherent in the two training paradigms. Consequently, the adoption of FL should be regarded as a complementary approach, rather than a direct substitute for centralised methodologies.

Overall, given the limited exploration, this thesis focused on advancing the state of the art in the field of federated I2I translation, with a particular focus on MRI-to-CT and CBCT-

to-CT synthesis. In addition, efficiency, sustainability, and security aspects of the federated frameworks for I2I translation and brain tumour segmentation have been investigated.

PART II

ENABLING ROBUST AND
PRIVACY-PRESERVING SYNTHETIC
CT GENERATION WITH
FEDERATED LEARNING

FedSynthCT-Brain: A Federated Brain MRI-to-CT Synthesis Framework

4.1 Introduction

The generation of sCT images from MRI represents a key enabling technology for MRI-only radiotherapy workflows as well as for attenuation correction in hybrid PET/MRI systems [1, 20, 102]. Despite the substantial progress achieved by DL-based translation methods, several fundamental limitations persist. As reported in Chapter 1 and 3, most existing Deep Learning (DL) models are trained on datasets that reflect narrowly defined patient populations, which significantly restricts their ability to generalise across heterogeneous clinical environments [10–13]. For example, training cohorts frequently exclude patients with metallic implants or uncommon anatomical configurations, leading to reduced reliability of sCT generation in these clinically relevant but underrepresented scenarios [103]. Moreover, models developed using single-centre data typically exhibit degraded performance when deployed on data acquired at external institutions, due to their susceptibility to domain shifts arising from variations in imaging protocols, scanner hardware, and patient demographics [14].

In an effort to enhance generalisability in MRI-to-sCT synthesis and to advance methodological robustness, several studies have leveraged cross-centre and multi-centre datasets to train models capable of generalising to multiple anatomical regions originating from diverse institutions [53, 88, 104–106]. Beyond individual studies, community-driven initiatives such as the SynthRAD challenges [33, 53] have been introduced to further promote the development of powerful DL-based sCT generation methods. These initiatives enabled models to be trained multi-centre datasets encompassing multiple anatomical regions, including brain, head and neck and abdomen imaging.

As discussed in Chapter 1 and 3, although the direct aggregation of medical images from multiple institutions could intuitively mitigate the limitations associated with small and homogeneous datasets, such an approach is often infeasible in practice due to strict data protection regulations. The combined effect of limited data diversity inherent to single-institution datasets and the practical and legal challenges associated with data centralisation highlights the need for collaborative learning paradigms that enable access to large and diverse datasets without requiring centralised data storage [17].

In this context, the study included in this Chapter, presents the development and validation of *FedSynthCT-Brain*, a federated learning framework for brain T1-weighted Magnetic Resonance Imaging (MRI)-to-Synthetic Computed Tomography (sCT) translation.

FedSynthCT-Brain addressed these challenges by adopting a cross-silo horizontal federated learning paradigm, which enabled multiple institutions to collaboratively train a DL model for MRI-to-sCT translation without exchanging raw patient data. The framework was explicitly designed to operate under realistic clinical conditions, accounting for heterogeneity in scanner vendors, acquisition protocols, and patient populations. The proposed approach was validated on real-world multi-centre data collected from four American and European institutions and was further evaluated on an independent external centre to assess its generalisation capability [99].

A comprehensive experimental analysis was conducted to assess both model performance and system efficiency under federated conditions. The evaluation demonstrated that the federated model generalised effectively to both unseen local clients data and the external dataset, confirming its robustness to heterogeneous data distributions. In particular, the analysis of multiple U-Net-based architectures showed no relevant differences in image similarity metrics across centres or on external data, suggesting that these architectures are sufficiently robust to handle inter-centre variability within the proposed federated framework. Consequently, the comparison naturally shifted towards computational efficiency, which emerged as a potential critical factor in federated environments where training time, resource availability, and communication overhead may directly impact feasibility [99].

The investigation of federated aggregation strategies further highlighted the importance of balancing convergence speed and stability. Among the evaluated methods, the combination of FedAvg with a proximal term (FedProx) achieved comparable image quality to more complex strategies while requiring fewer communication rounds, thereby reducing overall training time and mitigating synchronisation overhead. These findings reinforced previous observations in FL literature regarding the benefits of proximal regularisation in heterogeneous settings and underline the suitability of simple yet effective aggregation schemes for medical imaging applications.

From a qualitative perspective, the federated model demonstrated reliable performance across centres. Residual errors between the CT and the sCT were predominantly localised in anatomically complex or low-contrast areas, such as air–bone interfaces and small detailed

structures, reflecting known challenges in MRI-to-sCT translation. Importantly, the federation contributed to increased robustness in cases affected by acquisition-related artefacts or reduced spatial resolution, as the shared knowledge across institutions helped stabilise model behaviour.

Beyond implementing and validating the proposed framework, this work also introduced a novel and dedicated training strategy, called “Random-Multi2D”, aimed at improving training robustness and generalizability under federated settings. This establishes an applicational and methodological foundation for future developments in federated medical image synthesis.

The content of this chapter has been adapted from a manuscript previously published in the peer-reviewed journal “Computers in Biology and Medicine”. Portions of the text are reproduced verbatim from the published article [99]. The author of this thesis, as the first author of the aforementioned publication, made substantial and independent contributions to all stages of the research, as detailed in the author contributions section. Specifically, the author was responsible for the overall conceptualisation and methodological framework, the design and implementation of the FL pipeline, the experimental evaluation on multi-centre data and the interpretation and discussion of the results. The author was also responsible for developing most of the software code underlying the framework and for integrating the trained models into the open-source SlicerModalityConverter¹ extension for the 3D Slicer Imaging Computing Platform [107]. Furthermore, the author led the drafting and revision of the original manuscript in close collaboration with the co-authors.

4.2 Manuscript

¹SlicerModalityConverter, <https://github.com/ciroraggio/SlicerModalityConverter>. Last accessed: October 27, 2026



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

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

^a Institute of Biomedical Engineering, Karlsruhe Institute of Technology, Fritz-Haber-Weg 1, Karlsruhe, 76131, Baden-Württemberg, Germany

^b Department of Radiation Oncology, University Medical Center Schleswig-Holstein, Feldstrasse 21, Kiel, 24105, Schleswig-Holstein, Germany

^c Department of Experimental and Clinical Medicine, Magna Graecia University, Viale Europa, Catanzaro, 88100, Calabria, Italy

^d Department of Information Engineering, Università Politecnica delle Marche, Via Brecce Bianche 12, Ancona, 60131, Marche, Italy

ARTICLE INFO

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

Keywords:

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

ABSTRACT

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

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

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

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

1. Introduction

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

derivation of Synthetic CT (sCT) from MRI or Cone Beam CT (CBCT), to facilitate radiotherapy planning/adaptation or to improve attenuation correction in PET/MRI acquisition [2–4].

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

* Corresponding author.

E-mail address: ciro.raggio@kit.edu (C.B. Raggio).

<https://doi.org/10.1016/j.complbiomed.2025.110160>

Received 12 December 2024; Received in revised form 5 March 2025; Accepted 4 April 2025

Available online 22 April 2025

0010-4825/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

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

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

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

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

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

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

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

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

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

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

2. Related work

2.1. Centralised MRI-to-sCT

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

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

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

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

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

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

¹ <https://www.hhs.gov/hipaa/index.html>.

² <https://gdpr-info.eu/>.

2.2. Federated learning in medical imaging

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

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

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

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

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

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

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

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

3. Methods

3.1. Datasets and preprocessing

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

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

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

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

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

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

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

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

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

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

3.2. Federation setup

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

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

Table 1

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

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

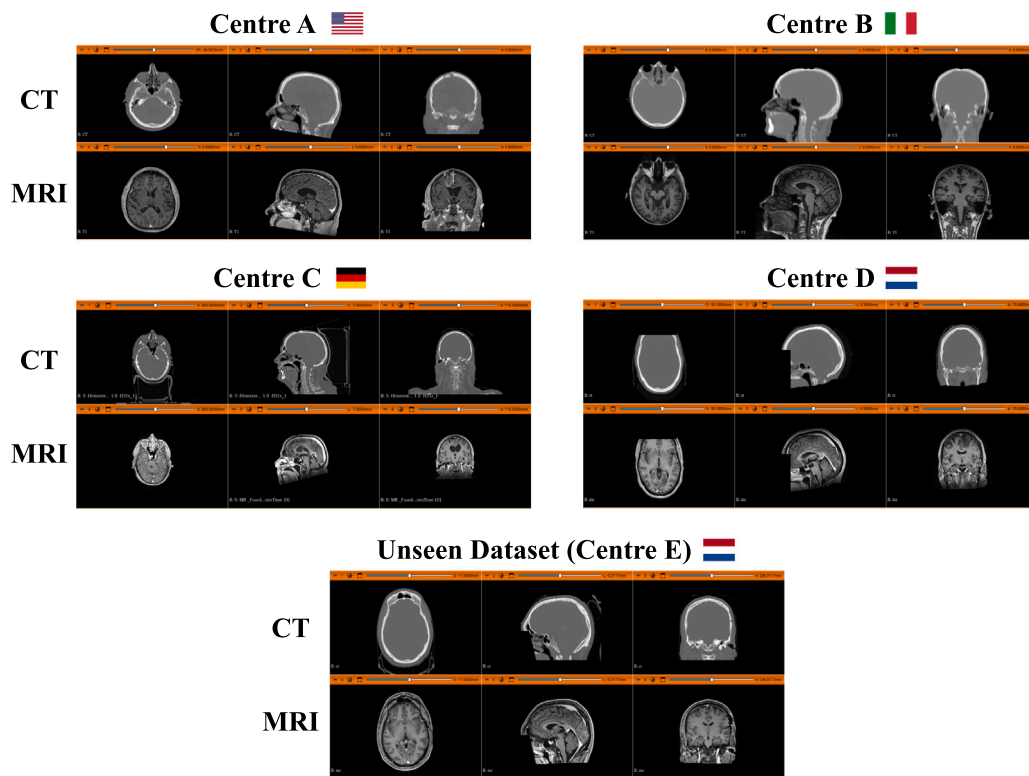


Fig. 1. Visual comparison of image appearance across datasets from multiple Centres. The FOV of the images from Centre A was limited to the level of the mouth and included partial cut-off of the ears. Images from centre B were acquired using a PET-MRI hybrid scanner. Furthermore the higher CT slice thickness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

3.3. Deep learning model

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

with a 7×7 convolutional layer for initial feature extraction, followed by multiple 3×3 convolutional layers.

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

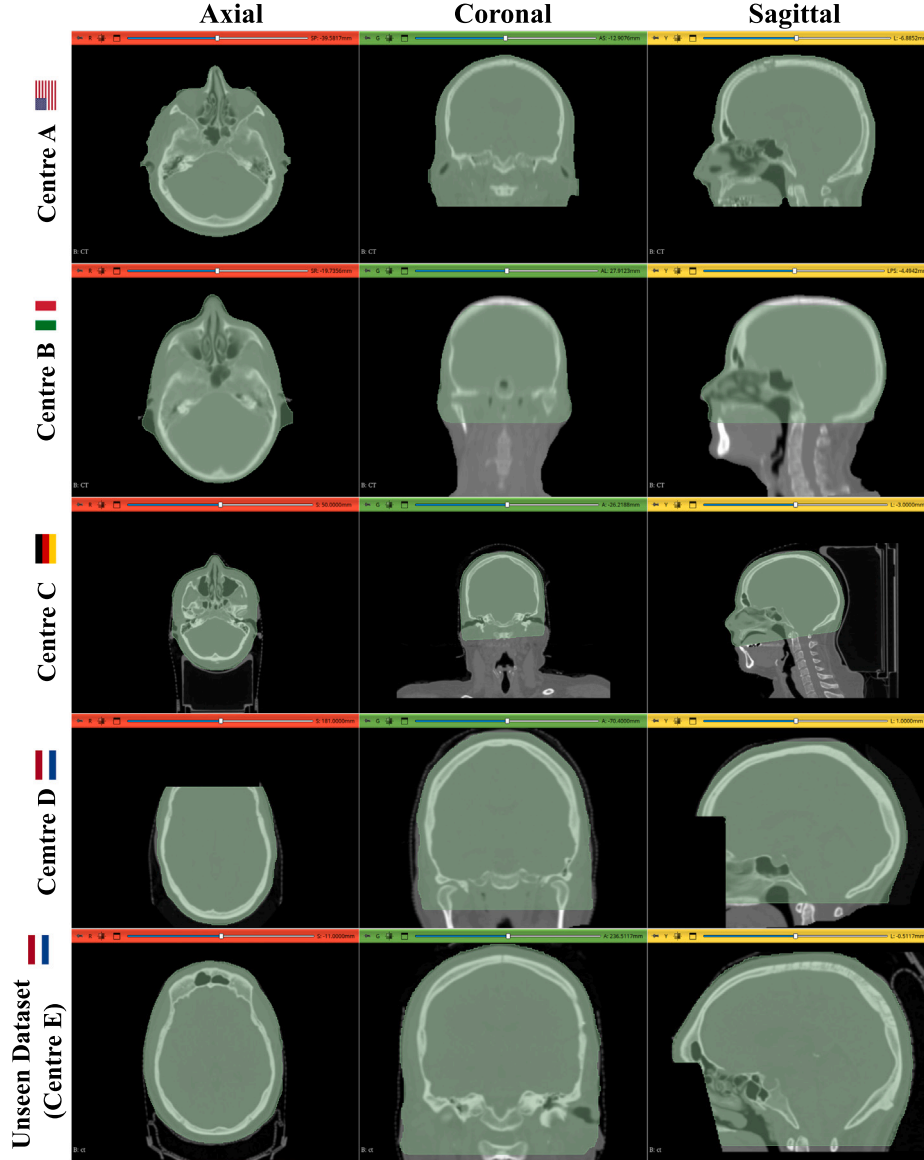


Fig. 2. The figure illustrates the background masking for each institution. By eliminating potential misleading features for the model, the masking approach limits the prediction of the model to areas of interest and facilitates the harmonisation of the model's knowledge within the federation. The masking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

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

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

3.4. Voting

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

$$sHU_{V_n} = \text{median}(sHU_{Cor_n}; sHU_{Ax_n}; sHU_{Sag_n}), \quad n \in \{1, \dots, N\} \quad (1)$$

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

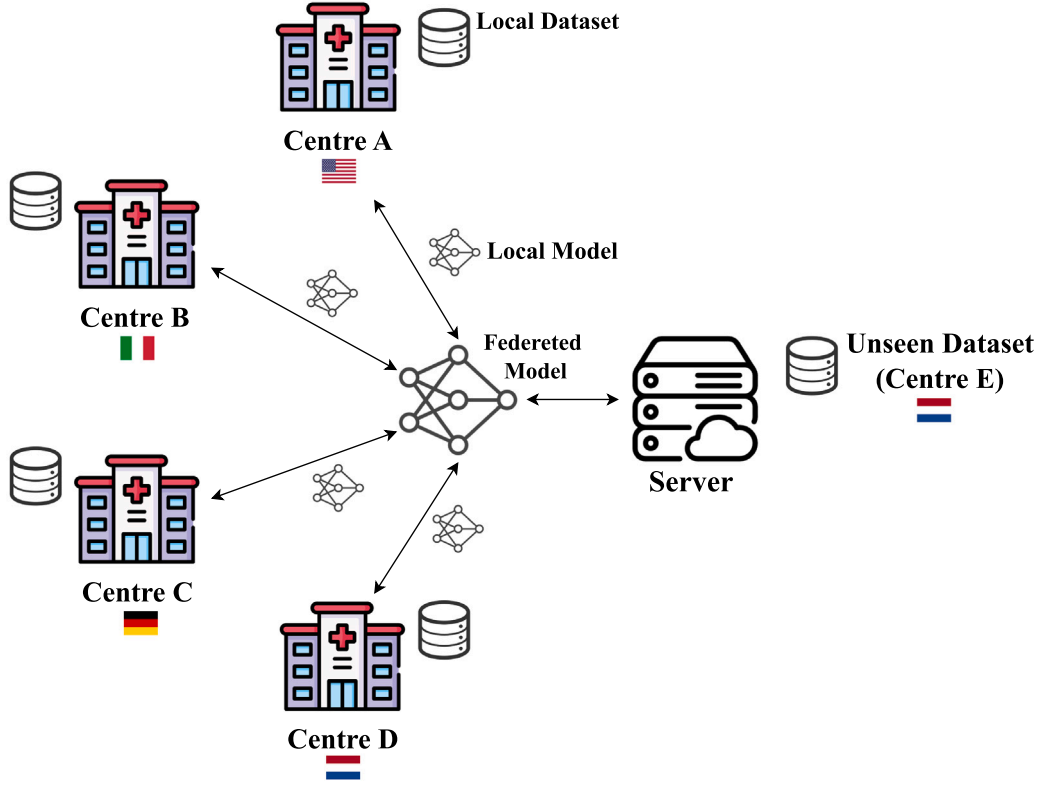


Fig. 3. Federated setup. Each centre (also called client) has its own local data and trains a model on the local dataset. The weights of the local model are exchanged with the server, which, after the aggregation process, returns a federated model to the clients. The server also hosts a dataset that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

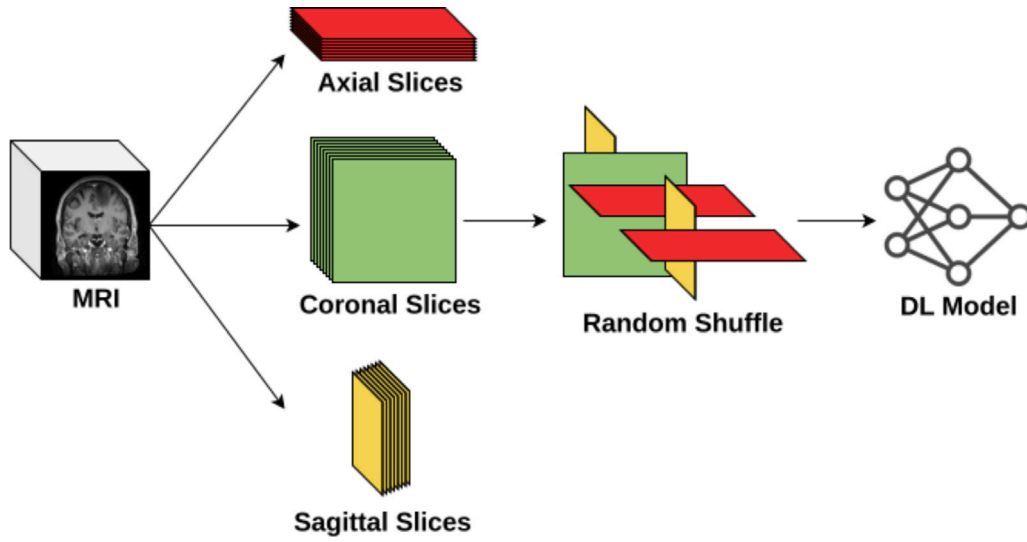


Fig. 4. Proposed training methodology. The pre-processed and augmented 3D (MRI) volume is subdivided into 2D slices, extracted along the three anatomical planes (sagittal, axial, coronal). These slices are then subjected to a random shuffling process, ensuring that the model is agnostic with respect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

3.5. Aggregation strategy

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

More in detail, the weights from each client are aggregated round-by-round using a weighted average (thus, FedAvg), with each client's

contribution proportional to the size of its dataset.

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

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

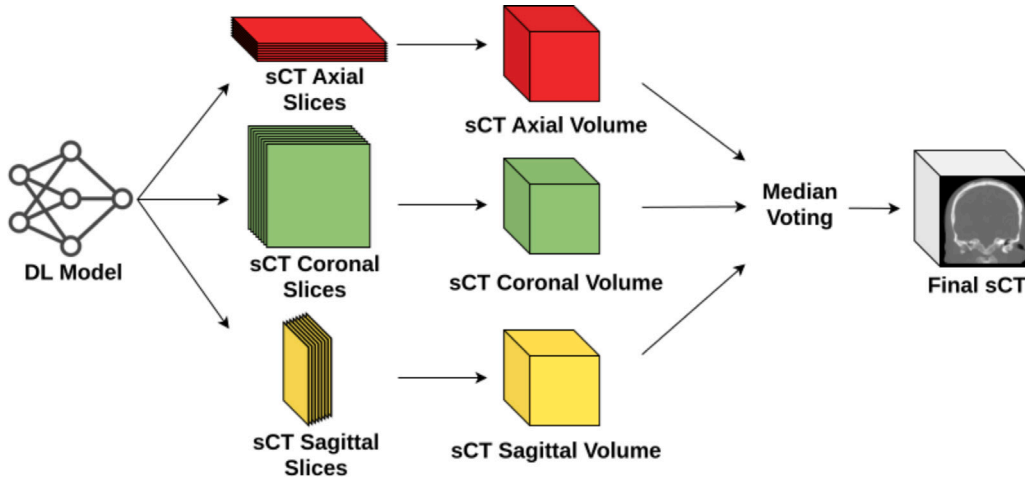


Fig. 5. Reconstruction of the sCT. The axial, coronal, and sagittal slice predictions are rearranged to construct three distinct volumes. Subsequently, a median voting approach is adopted using the three volumes in order to enhance the final result.

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

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

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

4. Experimental protocol

4.1. Data preprocessing details

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

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

```

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

```

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

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

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

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

4.2. Implementation and training settings

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

To maintain consistency and prevent overfitting, a custom-built data augmentation pipeline was implemented for each client. The AugmentedDataLoader library³ was used to perform optimised on-the-fly data augmentation during training. In order to address the considerable computational demands associated with this data augmentation approach, two distinct data augmentation pipelines were employed: (i) a minimal pipeline, comprising a flip and a fixed-angle single-axis rotation; (ii) an extended pipeline, comprising a flip, a random-angle three-axis rotation and a fixed spatial translation; based on the amount of data available for the individual site.

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

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

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

³ <https://github.com/ciroraggio/AugmentedDataLoader>.

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

4.3. Performance metrics

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

- Mean Absolute Error (MAE):

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

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

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

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

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

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

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

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

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

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

4.4. Deep learning architecture and aggregation strategy selection

4.4.1. Deep learning architecture selection

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

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

4.4.2. Aggregation strategy selection

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

Further optimisation was pursued through the incorporation of additional techniques, including FedBN [64], and the combination

of FedProx and FedBN, with the aim of refining and optimising the federated training process. In our study, the FedBN method was specifically designed by the exclusion of batch normalisation layers from communication between the server and client. This was done with a particular focus on ensuring that the global model could converge more rapidly and improve the robustness.

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

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

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

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

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

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

5. Results

5.1. Benchmarking deep learning architectures

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

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

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

5.2. Benchmarking aggregation strategies

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

Table 2

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

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

Table 3

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

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

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

5.3. Federated sCTs

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

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

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

The results obtained by the federated model on one random patient of each client test dataset are also presented for each centre in Figs. 7 and 8. The central slice of the three anatomical planes is provided for

the clients' data, along with the ground truth CT and sCT difference. The figures include rectangles of different colours to identify errors or artefacts observed in the images during the comparative analysis, that are not evident when considering only the similarity metrics.

6. Discussion

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

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

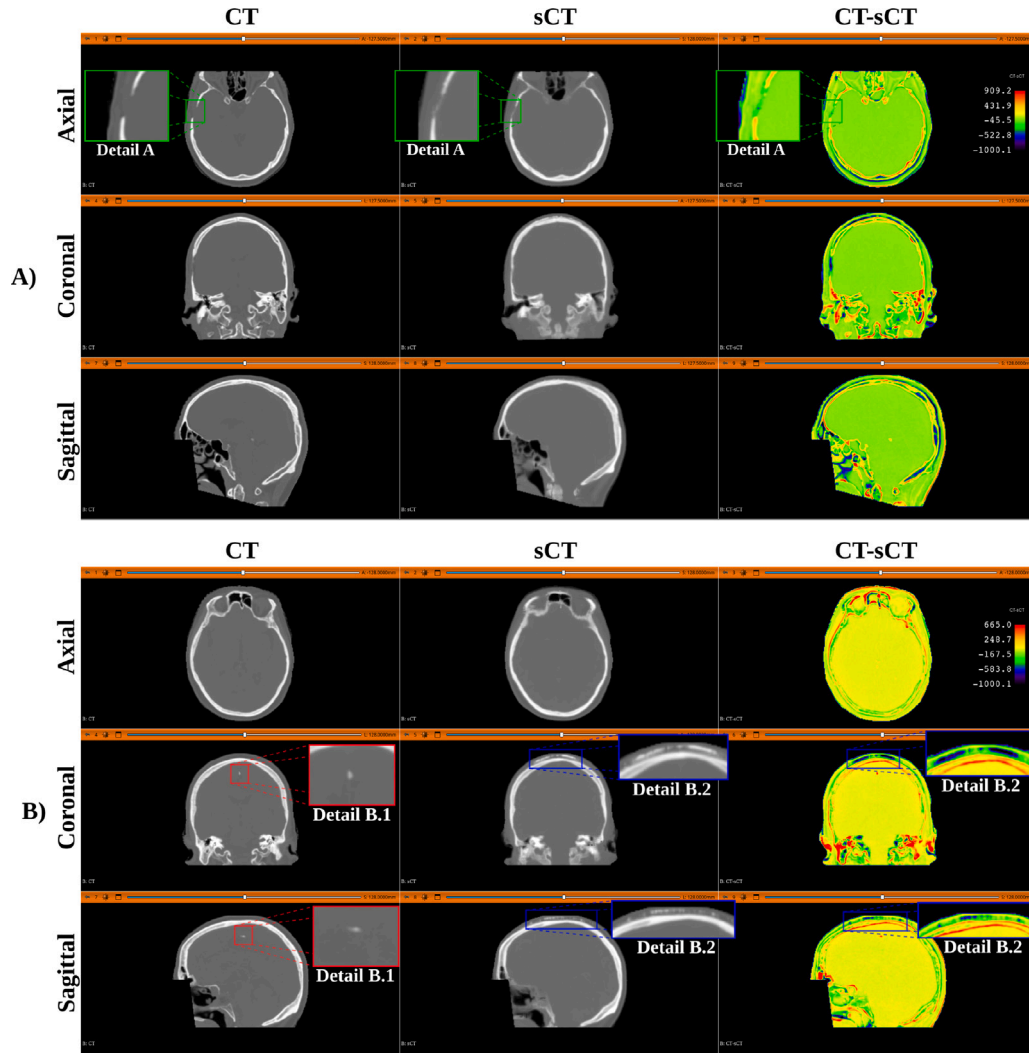


Fig. 6. The figure shows the central slice for the axial, coronal and sagittal planes, which is used for an evaluation of the worst and best – A and B respectively – predictions based on image similarity metrics of the federated model on the unseen dataset (Centre E). The Detail A (in green), shows the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

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

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

The employment of U-Net architectures was further justified by supplementary experiments comparing various models, including Conditional GAN, CycleGAN and Transformers. The results of these experiments, which are discussed in Appendix A, demonstrated that

the U-Net achieved the optimal overall performance, balancing image similarity metrics and computational efficiency, thus rendering them more suitable for FL MRI-to-sCT applications.

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

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

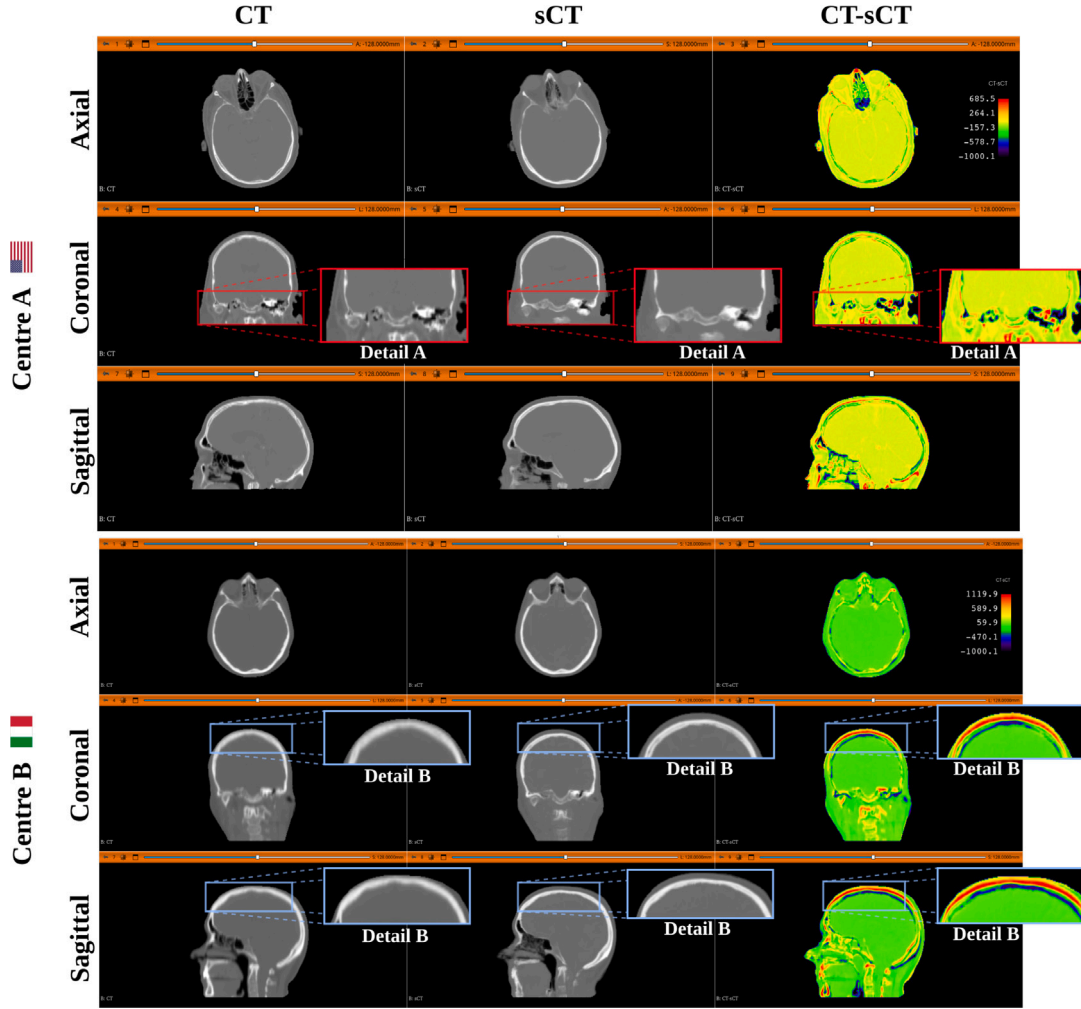


Fig. 7. The figure shows the central slice for the axial, coronal, and sagittal planes, which is used for an assessment of one of the two test cases of Centres A and B, respectively. The Detail A (in red) comparison highlight the approximation of the small structure performed by the model during the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

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

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

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

To confirm the benefit of the proximal term in this specific federated task, a combination of FedProx with FedAvg and FedBN (FedAvg + FedProx + FedBN) was investigated. It is evident from Table 3 that, despite the lack of improvement, the introduction of the proximal term (FedProx) resulted in a slight reduction in the average number of rounds required for convergence. Consequently, we concluded that

the chosen strategy, which involves the simple combination of FedAvg with FedProx, with a proximal coefficient $\mu = 3$, offers a comparable performance to more complex and optimised strategies, while requiring fewer rounds.

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

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

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

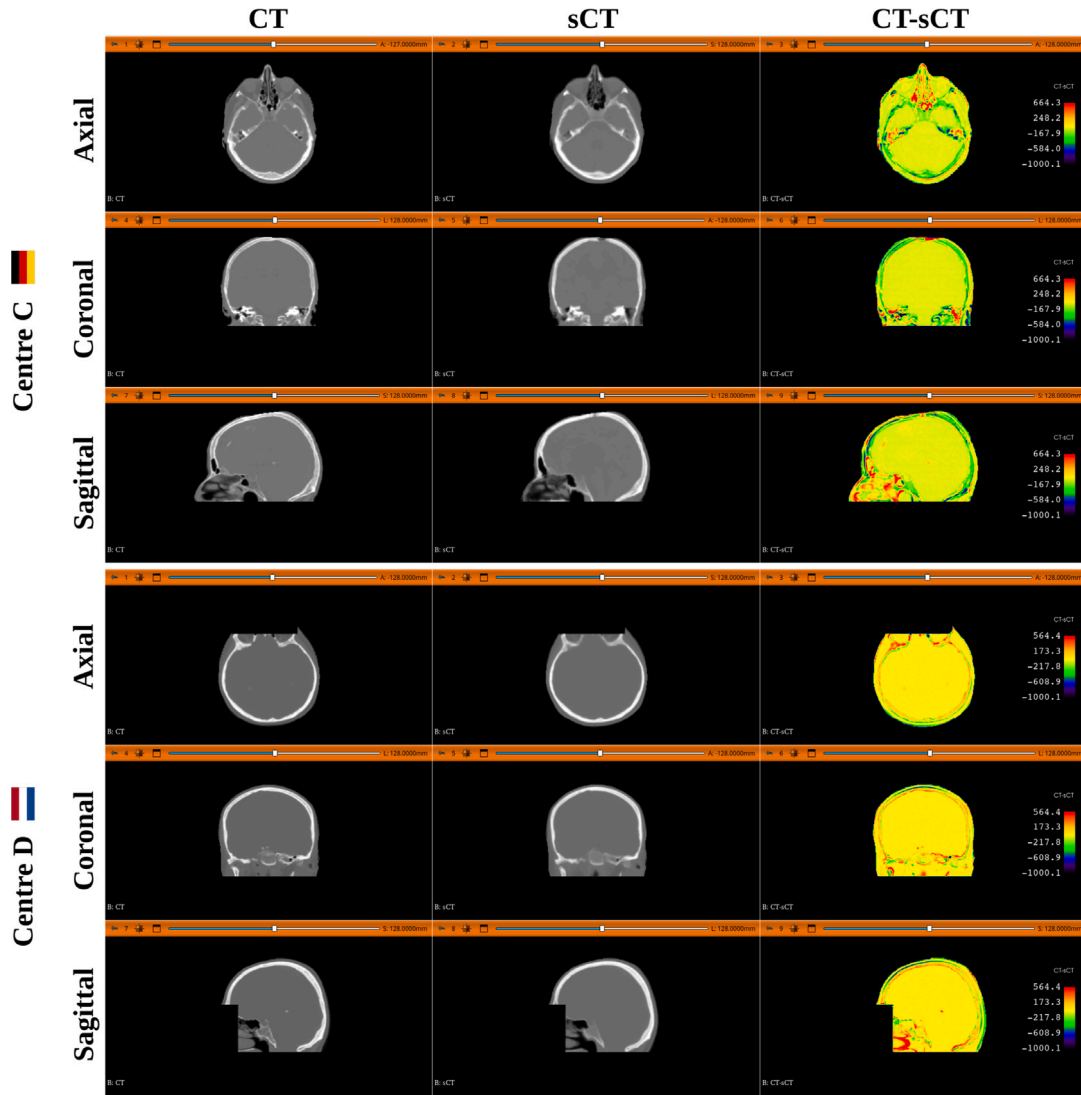


Fig. 8. The results obtained from the federated model for one of the two test cases of Centres C and D. As for the other Centres of the federation, conversion problems emerge in small and detailed areas, while for the majority of the image composed of soft tissue the error remains contained.

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

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

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

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

sCT appears more realistic and consistent with other examples. This establishes an advantage of the federation, as the influence given by the weights of the other Centres allows the federated model to remain robust in front of optimisation errors during training.

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

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

This work established the foundations for future developments. Several limitations of the employed architecture could be addressed by adopting more advanced medical image synthesis methodologies [24–26], optimising their efficiency, and ensuring compatibility with the FL environment. A possible improvement could also involve optimising the implementation of 3D architectures for federated scenarios by leveraging more efficient training strategies, such as model pruning,

knowledge distillation, or dimensionality reduction [67]. Additionally, exploring hybrid federated approaches, such as cross-silo horizontal FL with asynchronous updates or client clustering [68,69], could reduce synchronisation times.

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

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

7. Conclusion

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

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

CRedit authorship contribution statement

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

Ethics statement

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

Declaration of competing interest

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

Acknowledgement

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

Appendix A. Investigation of deep learning architectures

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

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

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

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

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

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

Table A.4

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

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

^a Time required to perform a single round.

Table B.5

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

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

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

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

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

Appendix B. Training methods investigation

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

The results presented in [Table B.5](#) further supported the superiority of the proposed Random Multi-2D approach over the alternative 2D methods evaluated. Specifically, the Random Multi-2D method achieved performance in terms of image similarity metrics comparable to the 2D+ approach, where each model was specifically trained on a single plane (axial, coronal and sagittal). Our training method required only a single model for all anatomical planes, thereby simplifying the federated training and inference processes. Additionally, the training

time was significantly reduced with the Random Multi-2D model completing training in 12 h and 34 min, compared to 16 h for the 2D+ models. In contrast, the 2D patch-based method requiring less training time but exhibited inferior performance across all image similarity metrics.

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

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

Appendix C. Centre B artefact measurement

See [Fig. C.9](#).

Appendix D. Aggregation strategies investigation

Convergence trends

See [Fig. D.10](#).

Generated sCTs

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

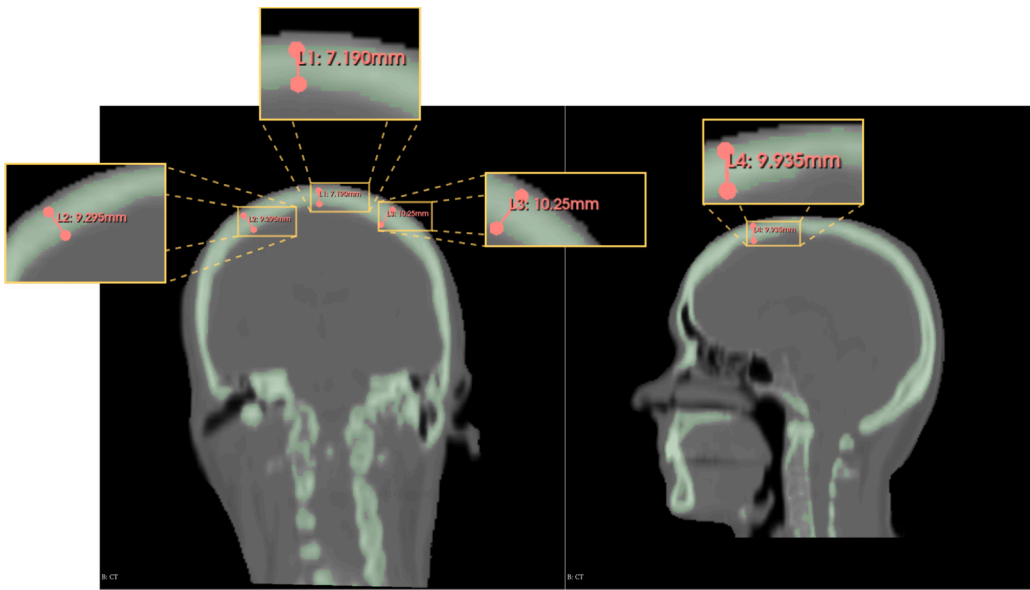


Fig. C.9. Measurement of a portion of the skull to highlight the partial volume artefact due to the higher slice thickness in Centre B CTs.

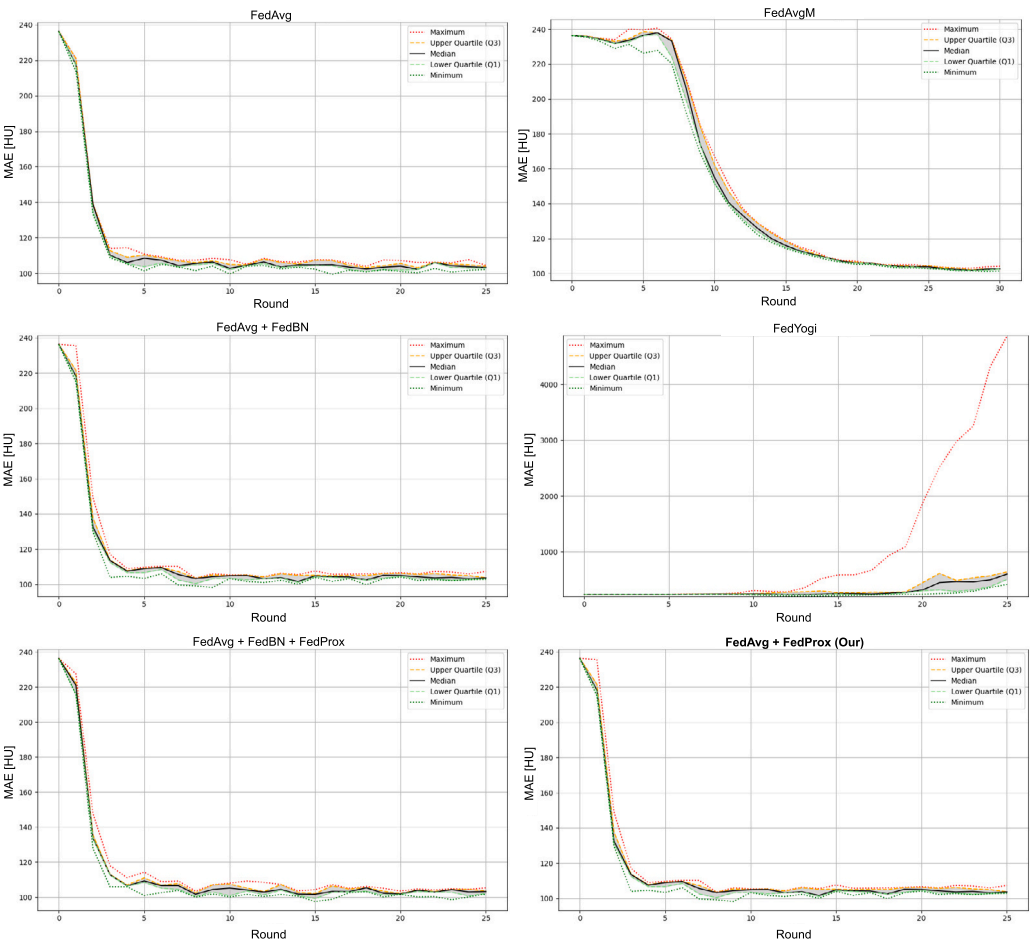


Fig. D.10. Comparison of convergence trends of aggregation strategies.

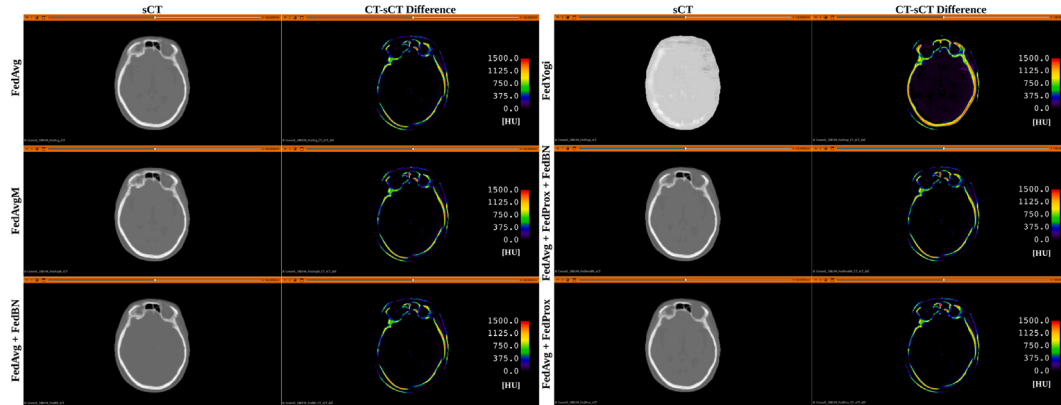


Fig. D.11. Visual comparison of sCT generated with different aggregation strategies on an axial slice from a patient of Centre E. The discrepancy between the ground truth CT and the sCT is highlighted on the right.

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

Data availability

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

References

- [1] S. Dayarathna, K.T. Islam, S. Uribe, G. Yang, M. Hayat, Z. Chen, Deep learning based synthesis of mri, ct and pet: Review and analysis, *Med. Image Anal.* 92 (2024) 103046, <http://dx.doi.org/10.1016/j.media.2023.103046>.
- [2] M.F. Spadea, M. Maspero, P. Zaffino, J. Seco, Deep learning based synthetic-ct generation in radiotherapy and pet: A review, *Med. Phys.* 48 (2021) 6537–6566, <http://dx.doi.org/10.1002/mp.15150>.
- [3] A. Thummerer, P. Zaffino, M.F. Spadea, A. Knopf, M. Maspero, Artificial intelligence to generate synthetic ct for adaptive particle therapy, in: *Imaging in Particle Therapy*, IOP Publishing, 2024, pp. 8–1 to 8–16, <http://dx.doi.org/10.1088/978-0-7503-5117-1ch8>, 2053-2563.
- [4] M. Boulanger, J.-C. Nunes, H. Chourak, A. Largent, S. Tahri, O. Acosta, R. De Crevoisier, C. Lafond, A. Barateau, Deep learning methods to generate synthetic ct from mri in radiotherapy: A literature review, *Phys. Medica* 89 (2021) 265–281, URL <http://dx.doi.org/10.1016/j.ejmp.2021.07.027>.
- [5] Y. Liu, A. Chen, H. Shi, S. Huang, W. Zheng, Z. Liu, Q. Zhang, X. Yang, Ct synthesis from mri using multi-cycle gan for head-and-neck radiation therapy, *Comput. Med. Imaging Graph.* 91 (2021) 101953, <http://dx.doi.org/10.1016/j.compmedimag.2021.101953>.
- [6] R. Graf, J. Schmitt, S. Schlaeger, H.K. Möller, V. Sideri-Lampretsa, A. Sekuboyina, S.M. Krieg, B. Wiestler, B. Menze, D. Rueckert, J.S. Kirschke, Denoising diffusion-based mri to ct image translation enables automated spinal segmentation, *Eur. Radiol. Exp.* 7 (2023) 70, <http://dx.doi.org/10.1186/s41747-023-00385-2>.
- [7] M. Özbey, O. Dalmaz, S.U. Dar, H.A. Bedel, Ş. Öztürk, A. Güngör, T. Çukur, Unsupervised medical image translation with adversarial diffusion models, *IEEE Trans. Med. Imaging* 42 (2023) 3524–3539, <http://dx.doi.org/10.1109/TMI.2023.3290149>.
- [8] M.A. Bahloul, S. Jabeen, S. Benoumhani, H.A. Alsaleh, Z. Belkhatir, A. Al-Wabil, Advancements in synthetic ct generation from mri: A review of techniques, and trends in radiation therapy planning, *J. Appl. Clin. Med. Phys.* 25 (2024) <http://dx.doi.org/10.1002/acm2.14499>.
- [9] J. Grigo, J. Szklisak, D. Höfler, R. Fietkau, F. Putz, C. Bert, Sct-feasibility - a feasibility study for deep learning-based mri-only brain radiotherapy, *Radiat. Oncol.* 19 (2024) 33, <http://dx.doi.org/10.1186/s13014-024-02428-3>.
- [10] O. Dalmaz, M.U. Mirza, G. Elmas, M. Özbey, S.U. Dar, E. Ceyani, K.K. Oguz, S. Avestimehr, T. Çukur, One model to unite them all: Personalized federated learning of multi-contrast mri synthesis, *Med. Image Anal.* 94 (2024) 103121, <http://dx.doi.org/10.1016/j.media.2024.103121>.
- [11] S. Tahri, B. Texier, J.-C. Nunes, C. Hemon, P. Lekieffre, E. Collot, H. Chourak, J. Le Guevelou, P. Greer, J. Dowling, O. Acosta, I. Bessieres, L. Marage, A. Boue-Rafle, R. De Crevoisier, C. Lafond, A. Barateau, A deep learning model to generate synthetic ct for prostate mr-only radiotherapy dose planning: a multicenter study, *Front. Oncol.* 13 (2023) URL <http://dx.doi.org/10.3389/fonc.2023.1279750>.
- [12] B. Texier, C. Hémon, P. Lekieffre, E. Collot, S. Tahri, H. Chourak, J. Dowling, P. Greer, I. Bessieres, O. Acosta, A. Boue-Rafle, J.L. Guevelou, R. de Crevoisier, C. Lafond, J. Castelli, A. Barateau, J.-C. Nunes, Computed tomography synthesis from magnetic resonance imaging using cycle generative adversarial networks with multicenter learning, *Phys. Imaging Radiat. Oncol.* 28 (2023) 100511, URL <http://dx.doi.org/10.1016/j.phro.2023.100511>.
- [13] S. Han, B. Texier, C. Hémon, Y. Kortli, A. Queffelec, R. de Crevoisier, J. Dowling, P. Greer, I. Bessières, A. Barateau, C. Lafond, J.-C. Nunes, Balancing data consistency and diversity: Preprocessing and online data augmentation for multi-center deep learning-based mr-to-ct synthesis, *Pattern Recognit. Lett.* (2025) URL <http://dx.doi.org/10.1016/j.patrec.2025.01.002>.
- [14] E.M. Huijben, M.L. Terpstra, S. Pai, A. Thummerer, P. Koopmans, M. Afonso, M. van Eijnatten, O. Gurney-Champion, Z. Chen, Y. Zhang, et al., Generating synthetic computed tomography for radiotherapy: Synthrad2023 challenge report, *Med. Image Anal.* 97 (2024) 103276.
- [15] E.M. Huijben, M.L. Terpstra, A.J. Galapon, S. Pai, A. Thummerer, P. Koopmans, M. Afonso, M. van Eijnatten, O. Gurney-Champion, Z. Chen, Y. Zhang, K. Zheng, C. Li, H. Pang, C. Ye, R. Wang, T. Song, F. Fan, J. Qiu, Y. Huang, J. Ha, J. Sung Park, A. Alain-Beaudoin, S. Bériault, P. Yu, H. Guo, Z. Huang, G. Li, X. Zhang, Y. Fan, H. Liu, B. Xin, A. Nicolson, L. Zhong, Z. Deng, G. Müller-Franzes, F. Khader, X. Li, Y. Zhang, C. Hémon, V. Boussot, Z. Zhang, L. Wang, L. Bai, S. Wang, D. Mus, B. Kooiman, C.A. Sargeant, E.G. Henderson, S. Kondo, S. Kasai, R. Karimzadeh, B. Ibragimov, T. Helfer, J. Dafflon, Z. Chen, E. Wang, Z. Perko, M. Maspero, Generating synthetic computed tomography for radiotherapy: Synthrad2023 challenge report, *Med. Image Anal.* 97 (2024) 103276, <http://dx.doi.org/10.1016/j.media.2024.103276>, URL <https://www.sciencedirect.com/science/article/pii/S1361841524002019>.
- [16] Z. Chen, K. Zheng, C. Li, Z. Yiwen, A hybrid network with multi-scale structure extraction and preservation for mr-to-ct synthesis in synthrad2023, 2023, SynthRAD2023.
- [17] G.A. Kaissis, M.R. Makowski, D. Rückert, R.F. Braren, Secure, privacy-preserving and federated machine learning in medical imaging, *Nat. Mach. Intell.* 2 (2020) 305–311.
- [18] H. Guan, P.-T. Yap, A. Bozoki, M. Liu, Federated learning for medical image analysis: A survey, *Pattern Recognit.* 151 (2024) 110424, <http://dx.doi.org/10.1016/j.patcog.2024.110424>, URL <https://www.sciencedirect.com/science/article/pii/S0031320324001754>.
- [19] M. Adnan, S. Kalra, J.C. Cresswell, G.W. Taylor, H.R. Tizhoosh, Federated learning and differential privacy for medical image analysis, *Sci. Rep.* 12 (2022) 1953.
- [20] N. Hernandez-Cruz, P. Saha, M.M.K. Sarker, J.A. Noble, Review of federated learning and machine learning-based methods for medical image analysis, *Big Data Cogn. Comput.* 8 (2024) <http://dx.doi.org/10.3390/bdcc8090099>.
- [21] Y. Lei, J. Harms, T. Wang, Y. Liu, H.-K. Shu, A.B. Jani, W.J. Curran, H. Mao, T. Liu, X. Yang, MRI-only based synthetic CT generation using dense cycle consistent generative adversarial networks, *Med. Phys.* 46 (2019) 3565–3581.
- [22] H. Koh, T.Y. Park, Y.A. Chung, J.-H. Lee, H. Kim, Acoustic simulation for transcranial focused ultrasound using gan-based synthetic ct, *IEEE J. Biomed. Heal. Inform.* 26 (2022) 161–171, <http://dx.doi.org/10.1109/JBHI.2021.3103387>.
- [23] R.S. Antunes, C. André da Costa, A. Küderle, I.A. Yari, B. Eskofier, Federated learning for healthcare: Systematic review and architecture proposal, *ACM Trans. Intell. Syst. Technol. (TIST)* 13 (2022) 1–23.

- [24] O.F. Atli, B. Kabas, F. Arslan, A.C. Demirtas, M. Yurt, O. Dalmaz, T. Çukur, I2i-mamba: Multi-modal medical image synthesis via selective state space modeling, 2024, URL <https://arxiv.org/abs/2405.14022>, arXiv:2405.14022.
- [25] B. Kabas, F. Arslan, V.A. Nezhad, S. Ozturk, E.U. Saritas, T. Çukur, Physics-driven autoregressive state space models for medical image reconstruction, 2024, URL <https://arxiv.org/abs/2412.09331>, arXiv:2412.09331.
- [26] F. Arslan, B. Kabas, O. Dalmaz, M. Ozbey, T. Çukur, Self-consistent recursive diffusion bridge for medical image translation, 2024, URL <https://arxiv.org/abs/2405.06789>, arXiv:2405.06789.
- [27] D. Saxena, J. Cao, Generative adversarial networks (gans): Challenges, solutions, and future directions, *ACM Comput. Surv.* 54 (2021) 1–42, URL <http://dx.doi.org/10.1145/3446374>.
- [28] B. Texier, C. Hémon, A. Queffelec, J. Dowling, I. Bessieres, P. Greer, O. Acosta, A. Boue-Rafle, R. de Crevoisier, C. Lafond, J. Castelli, A. Barateau, J.-C. Nunes, 3D unsupervised deep learning method for magnetic resonance imaging-to-computed tomography synthesis in prostate radiotherapy, *Phys. Imaging Radiat. Oncol.* 31 (2024) 100612, <http://dx.doi.org/10.1016/j.phro.2024.100612>, URL <https://www.sciencedirect.com/science/article/pii/S2405631624000824>.
- [29] G. Li, L. Bai, C. Zhu, E. Wu, R. Ma, A novel method of synthetic ct generation from mr images based on convolutional neural networks, in: 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics, CISP-BMEI, 2018, pp. 1–5, <http://dx.doi.org/10.1109/CISP-BMEI.2018.8633142>.
- [30] M.F. Spadea, G. Pileggi, P. Zaffino, P. Salome, C. Catana, D. Izquierdo-Garcia, F. Amato, J. Seco, Deep convolution neural network (dcnn) multiplane approach to synthetic ct generation from mr images—application in brain proton therapy, *Int. J. Radiat. Oncol. Biol. Phys.* 105 (2019) 495–503, <http://dx.doi.org/10.1016/j.ijrobp.2019.06.2535>.
- [31] S.-H. Hsu, Z. Han, J.E. Leeman, Y.-H. Hu, R.H. Mak, A. Sudhyadhom, Synthetic ct generation for mri-guided adaptive radiotherapy in prostate cancer, *Front. Oncol.* 12 (2022) <http://dx.doi.org/10.3389/fonc.2022.969463>.
- [32] M.J. Sheller, G.A. Reina, B. Edwards, J. Martin, S. Bakas, Multi-Institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation, *Brainlesion* 11383 (2019) 92–104.
- [33] W. Li, F. Milletari, D. Xu, N. Rieke, J. Hancox, W. Zhu, M. Baust, Y. Cheng, S. Ourselin, M.J. Cardoso, A. Feng, Privacy-preserving federated brain tumour segmentation, 2019, arXiv:1910.00962.
- [34] D. Li, A. Kar, N. Ravikumar, A.F. Frangi, S. Fidler, Fed-sim: Federated simulation for medical imaging, 2020, arXiv:2009.00668.
- [35] Q. Chang, H. Qu, Y. Zhang, M. Sabuncu, C. Chen, T. Zhang, D. Metaxas, Synthetic learning: Learn from distributed asynchronized discriminator gan without sharing medical image data, 2020, arXiv:2006.00080.
- [36] A. Xu, W. Li, P. Guo, D. Yang, H.R. Roth, A. Hatamizadeh, C. Zhao, D. Xu, H. Huang, Z. Xu, Closing the generalization gap of cross-silo federated medical image segmentation, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2022, pp. 20866–20875.
- [37] S.S. Sandhu, H.T. Gorji, P. Tavakolian, K. Tavakolian, A. Akhbardeh, Medical imaging applications of federated learning, *Diagnostics* 13 (2023) <http://dx.doi.org/10.3390/diagnostics13193140>.
- [38] I. Feki, S. Ammar, Y. Kessentini, K. Muhammad, Federated learning for covid-19 screening from chest x-ray images, *Appl. Soft Comput.* 106 (2021) 107330, <http://dx.doi.org/10.1016/j.asoc.2021.107330>, URL <https://www.sciencedirect.com/science/article/pii/S1568494621002532>.
- [39] D. Yang, Z. Xu, W. Li, A. Myronenko, H.R. Roth, S. Harmon, S. Xu, B. Turkbey, E. Turkbey, X. Wang, W. Zhu, G. Carrafiello, F. Patella, M. Cariati, H. Obinata, H. Mori, K. Tamura, P. An, B.J. Wood, D. Xu, Federated semi-supervised learning for covid region segmentation in chest ct using multi-national data from China, Italy, Japan, *Med. Image Anal.* 70 (2021) 101992, <http://dx.doi.org/10.1016/j.media.2021.101992>, URL <https://www.sciencedirect.com/science/article/pii/S1361841521000384>.
- [40] J. Wicaksana, Z. Yan, X. Yang, Y. Liu, L. Fan, K.-T. Cheng, Customized federated learning for multi-source decentralized medical image classification, *IEEE J. Biomed. Heal. Inform.* 26 (2022) 5596–5607, <http://dx.doi.org/10.1109/JBHI.2022.3198440>.
- [41] M.N. Hossen, V. Panneerselvam, D. Koundal, K. Ahmed, F.M. Bui, S.M. Ibrahim, Federated machine learning for detection of skin diseases and enhancement of internet of medical things (iomt) security, *IEEE J. Biomed. Heal. Inform.* 27 (2023) 835–841, <http://dx.doi.org/10.1109/JBHI.2022.3149288>.
- [42] Z. Yan, J. Wicaksana, Z. Wang, X. Yang, K.-T. Cheng, Variation-aware federated learning with multi-source decentralized medical image data, *IEEE J. Biomed. Heal. Inform.* 25 (2021) 2615–2628, <http://dx.doi.org/10.1109/JBHI.2020.3040015>.
- [43] J. Lo, T.T. Yu, D. Ma, P. Zang, J.P. Owen, Q. Zhang, R.K. Wang, M.F. Beg, A.Y. Lee, Y. Jia, M.V. Sarunic, Federated learning for microvasculature segmentation and diabetic retinopathy classification of oct data, *Ophthalmol. Sci.* 1 (2021) 100069, <http://dx.doi.org/10.1016/j.xops.2021.100069>, URL <https://www.sciencedirect.com/science/article/pii/S2666914521000671>.
- [44] A. Tuladhar, L. Tyagi, R. Souza, N.D. Forkert, Federated learning using variable local training for brain tumor segmentation, in: A. Crimi, S. Bakas (Eds.), *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Springer International Publishing, Cham, 2022, pp. 392–404.
- [45] E. Albalawi, M. T.R., V.V. Thakur, M. Kumar, S.B. Gupta, A. Khan, A. Al-musharraf, Integrated approach of federated learning with transfer learning for classification and diagnosis of brain tumor, *BMC Med. Imaging* (2024) <http://dx.doi.org/10.1186/s12880-024-01261-0>.
- [46] M. Zhu, Z. Chen, Y. Yuan, Feddm: Federated weakly supervised segmentation via annotation calibration and gradient de-conflicting, *IEEE Trans. Med. Imaging* 42 (2023) 1632–1643, <http://dx.doi.org/10.1109/TMI.2023.3235757>.
- [47] L. Qiu, J. Cheng, H. Gao, W. Xiong, H. Ren, Federated semi-supervised learning for medical image segmentation via pseudo-label denoising, *IEEE J. Biomed. Heal. Inform.* 27 (2023) 4672–4683, <http://dx.doi.org/10.1109/JBHI.2023.3274498>.
- [48] J. Wang, Y. Jin, D. Stoyanov, L. Wang, Feddp: Dual personalization in federated medical image segmentation, *IEEE Trans. Med. Imaging* 43 (2024) 297–308, <http://dx.doi.org/10.1109/TMI.2023.3299206>.
- [49] Y. Liu, G. Luo, Y. Zhu, Fedfms: Exploring federated foundation models for medical image segmentation, in: M.G. Linguraru, Q. Dou, A. Feragen, S. Giannarou, B. Glocker, K. Lekadir, J.A. Schnabel (Eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, Springer Nature Switzerland, Cham, 2024, pp. 283–293.
- [50] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A.C. Berg, W.-Y. Lo, P. Dollár, R. Girshick, Segment anything, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV*, 2023, pp. 4015–4026.
- [51] J. Wang, G. Xie, Y. Huang, J. Lyu, F. Zheng, Y. Zheng, Y. Jin, Fedmedgan: Federated domain translation on unsupervised cross-modality brain image synthesis, *Neurocomputing* 546 (2023) 126282, <http://dx.doi.org/10.1016/j.neucom.2023.126282>.
- [52] N.J. Tustison, B.B. Avants, P.A. Cook, Y. Zheng, A. Egan, P.A. Yushkevich, J.C. Gee, N4ITK: improved N3 bias correction, *IEEE Trans. Med. Imaging* 29 (2010) 1310–1320.
- [53] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, J. Finet, J.-C. Fillion-Robin, S. Pujol, C. Bauer, D. Jennings, F. Fennessy, M. Sonka, J. Buatti, S. Aylward, J.V. Miller, S. Pieper, R. Kikinis, 3D slicer as an image computing platform for the quantitative imaging network, *Magn. Reson. Imaging* 30 (2012) 1323–1341, <http://dx.doi.org/10.1016/j.mri.2012.05.001>, quantitative Imaging in Cancer.
- [54] Y. Li, W. Li, P. He, J. Xiong, J. Xia, Y. Xie, Ct synthesis from mri images based on deep learning methods for mri-only radiotherapy, in: 2019 International Conference on Medical Imaging Physics and Engineering, ICMPE, 2019, pp. 1–6, <http://dx.doi.org/10.1109/ICMPE47306.2019.9098190>.
- [55] K. Kamnitsas, C. Ledig, V.F. Newcombe, J.P. Simpson, A.D. Kane, D.K. Menon, D. Rueckert, B. Glocker, Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation, *Med. Image Anal.* 36 (2017) 61–78.
- [56] T. Li, A.K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, V. Smith, Federated optimization in heterogeneous networks, *Proc. Mach. Learn. Syst.* 2 (2020) 429–450.
- [57] X. Li, K. Huang, W. Yang, S. Wang, Z. Zhang, On the convergence of fedavg on non-iid data, 2019, arXiv preprint arXiv:1907.02189.
- [58] D.J. Beutel, T. Topal, A. Mathur, X. Qiu, J. Fernandez-Marques, Y. Gao, L. Sani, K.H. Li, T. Parcollet, P.P.B. de Gusmão, N.D. Lane, Flower: A friendly federated learning research framework, 2022, arXiv:2007.14390.
- [59] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An Imperative Style, High-Performance Deep Learning Library, Curran Associates Inc., 2019.
- [60] X. Chen, X. Zhou, H. Zhang, M. Sun, T. Zhao, Cost-effective federated learning: A unified approach to device and training scheduling, in: *ICC 2024-IEEE International Conference on Communications*, IEEE, 2024, pp. 3488–3493.
- [61] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W.M. Wells, A.F. Frangi (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Springer International Publishing, Cham, 2015, pp. 234–241.
- [62] J. Fu, Y. Yang, K. Singhrao, D. Ruan, F.-I. Chu, D.A. Low, J.H. Lewis, Deep learning approaches using 2d and 3d convolutional neural networks for generating male pelvic synthetic computed tomography from magnetic resonance imaging, *Med. Phys.* 46 (2019) 3788–3798, <http://dx.doi.org/10.1002/mp.13672>.
- [63] T. Li, M. Sanjabi, A. Beirami, V. Smith, Fair resource allocation in federated learning, 2020, arXiv:1905.10497.
- [64] X. Li, M. Jiang, X. Zhang, M. Kamp, Q. Dou, FedBN: Federated learning on non-iid features via local batch normalization, 2021, arXiv preprint arXiv:2102.07623.
- [65] Y. Li, W. Li, J. Xiong, J. Xia, Y. Xie, Comparison of supervised and unsupervised deep learning methods for medical image synthesis between computed tomography and magnetic resonance images, *BioMed Res. Int.* 2020 (2020) 5193707, <http://dx.doi.org/10.1155/2020/5193707>.
- [66] S. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečný, S. Kumar, H.B. McMahan, Adaptive federated optimization, 2020, arXiv preprint arXiv:2003.00295.
- [67] S. Soni, A. Saeed, Y.M. Asano, Federated learning with a single shared image, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7782–7790.

- [68] J. Nguyen, K. Malik, H. Zhan, A. Yousefpour, M. Rabbat, M. Malek, D. Huba, Federated learning with buffered asynchronous aggregation, in: *International Conference on Artificial Intelligence and Statistics*, PMLR, 2022, pp. 3581–3607.
- [69] Z. Wang, H. Xu, J. Liu, Y. Xu, H. Huang, Y. Zhao, Accelerating federated learning with cluster construction and hierarchical aggregation, *IEEE Trans. Mob. Comput.* 22 (2022) 3805–3822.
- [70] P. Zaffino, C.B. Raggio, A. Thummerer, G.G. Marmitt, J.A. Langendijk, A. Procopio, C. Cosentino, J. Seco, A.C. Knopf, S. Both, M.F. Spadea, Toward closing the loop in image-to-image conversion in radiotherapy: A quality control tool to predict synthetic computed tomography hounsfield unit accuracy, *J. Imaging* 10 (2024) <http://dx.doi.org/10.3390/jimaging10120316>, URL <https://www.mdpi.com/2313-433X/10/12/316>.
- [71] D. Jhunjhunwala, S. Wang, G. Joshi, Fedfisher: Leveraging fisher information for one-shot federated learning, 2024, URL <https://arxiv.org/abs/2403.12329>, arXiv:2403.12329.
- [72] Q. Liu, C. Chen, J. Qin, Q. Dou, P.-A. Heng, Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space, 2021, arXiv:2103.06030.
- [73] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, 2018, URL <https://arxiv.org/abs/1611.07004>, arXiv:1611.07004.
- [74] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: *2017 IEEE International Conference on Computer Vision, ICCV*, 2017, pp. 2242–2251, <http://dx.doi.org/10.1109/ICCV.2017.244>.

Privacy-Preserving and Generalizable Federated Head and Neck CBCT-to-CT Translation

5.1 Introduction

Cone Beam Computed Tomography (CBCT) has become a cornerstone of modern Image-guided radiotherapy (IGRT), enabling volumetric imaging directly in the treatment room. It supports daily patient positioning, treatment plan verification, margin reduction as well as dose escalation strategies. The integration of CBCT systems into linear accelerators has significantly improved treatment precision and facilitated the development of online adaptive radiotherapy paradigms. However, despite its widespread adoption, CBCT imaging remains intrinsically limited by increased noise, reduced soft-tissue contrast and scatter artefacts. These factors lead to inaccurate Hounsfield Unit (HU) values, preventing its direct use for electron density estimation and dose calculation. To overcome these limitations, synthetic Computed Tomography (CT) generation has emerged as a key enabling technology for CBCT-based adaptive radiotherapy, allowing the translation of CBCT volumes into CT-equivalent images with improved dosimetric reliability [1, 57, 108]. In particular, Deep Learning (DL)-based Image-to-Image (I2I) translation methods have demonstrated strong capability in learning complex, non-linear mappings between degraded CBCT inputs and high-quality CT targets, producing synthetic images suitable for adaptive planning workflows [1, 53, 57, 109]. Convolutional architectures such as U-Net [65] and adversarial frameworks based on conditional Generative Adversarial Networks (GANs) have shown promising performance in preserving anatomical structures while restoring realistic intensity distributions.

Despite the effectiveness of DL-based CBCT-to-sCT translation methods, their clinical scalability remains limited by institutional heterogeneity, scanner-dependent variability, and protocol-specific domain shifts, which typically confine model training to single-centre datasets. At the same time (as reported in Chapter 1 and 3) multi-centre data aggregation is

constrained by data protection regulations, making centralised training paradigms suboptimal in real-world clinical settings [18, 57, 99, 101, 110].

Building upon the methodological foundations established in the FedSynthCT-Brain study [99] (presented in Chapter 4), this study extended the FedSynthCT framework to CBCT imaging and head and neck IGRT potential application. Compared to MRI-to-CT translation, CBCT-to-CT synthesis requires the restoration of physically meaningful HU values from degraded, scatter-affected scans. Moreover, the head and neck anatomical site introduces increased structural variability, high-density dental materials, and air cavities, making the task particularly sensitive to misalignment and residual artefacts.

To address these challenges, a cross-silo horizontal federated strategy was adopted to collaboratively train a Conditional Generative Adversarial Network (cGAN) using the SynthRAD2025 dataset [33], enabling decentralised training without data sharing. The proposed framework was validated both on the three federated centres and on an independent external dataset, demonstrating robust cross-centre generalisation without site-specific fine-tuning.

Overall, the results of the study presented in this Chapter, demonstrated the technical feasibility of federated CBCT-to-Synthetic Computed Tomography (sCT) synthesis as a privacy-preserving and scalable solution for developing centre-agnostic models in the adaptive radiotherapy context.

The content of this chapter has been adapted from a manuscript previously published in the peer-reviewed journal “Frontiers in Digital Health”. Portions of the text are reproduced verbatim from the published article [100]. The author of this thesis, as first author of the manuscript, led the conceptualisation, methodological design, implementation, experimental evaluation, and interpretation of the results. The author was also responsible for integrating the trained model into the open-source SlicerModalityConverter extension for the 3D Slicer Imaging Computing Platform [107].

5.2 Manuscript



OPEN ACCESS

EDITED BY
Francesca Angelone,
Università del Sannio, Italy

REVIEWED BY
Federica Kiyomi Ciliberti,
Reykjavik University, Iceland
Sara Spaziani,
University of Sannio, Italy

*CORRESPONDENCE
Ciro Benito Raggio
✉ ciro.raggio@kit.edu

RECEIVED 16 February 2026
REVISED 22 May 2026
ACCEPTED 29 May 2026
PUBLISHED 15 June 2026

CITATION
Raggio CB, Zaffino P and Spadea MF
(2026) A privacy-preserving federated
learning framework for generalizable
CBCT to synthetic CT translation in head
and neck.
Front. Digit. Health 8:1812254.
doi: 10.3389/fdgth.2026.1812254

COPYRIGHT
© 2026 Raggio, Zaffino and Spadea. This
is an open-access article distributed
under the terms of the [Creative
Commons Attribution License \(CC BY\)](#).
The use, distribution or reproduction in
other forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution
or reproduction is permitted which does
not comply with these terms.

A privacy-preserving federated learning framework for generalizable CBCT to synthetic CT translation in head and neck

Ciro Benito Raggio^{1*}, Paolo Zaffino² and
Maria Francesca Spadea¹

¹Institute of Biomedical Engineering, Karlsruhe Institute of Technology, Karlsruhe, Germany,

²Department of Experimental and Clinical Medicine, Magna Graecia University, Catanzaro, Italy

Background: Cone-beam computed tomography (CBCT) has become a widely adopted modality for image-guided radiotherapy (IGRT). However, CBCT is characterized by increased noise, limited soft-tissue contrast, and artifacts. These issues result in unreliable Hounsfield unit (HU) values, which limits electron density estimation for direct dose calculation. These issues have been addressed by deriving synthetic CT (sCT) from CBCT, particularly by adopting deep learning (DL) methods. However, existing DL approaches are hindered by institutional heterogeneity, scanner-dependent variations, and data privacy regulations that prevented multi-center data sharing.

Methods: To overcome these challenges, we propose a cross-silo federated learning (FL) approach for CBCT-to-sCT synthesis in the head and neck region. This approach extends the original FedSynthCT framework to a different image modality and anatomical region. A conditional generative adversarial network (cGAN) was trained using data from three European medical centers within the SynthRAD2025 public challenge dataset while maintaining data privacy at each institution. A combination of the FedAvg and FedProx aggregation strategies, alongside a standardized preprocessing pipeline, was adopted to federate the DL model.

Results: The federated model effectively generalized across participating centers, as evidenced by the mean absolute error (MAE) ranging from 64.38 ± 13.63 to 85.90 ± 7.10 HU, the structural similarity index (SSIM) ranging from 0.88 ± 0.02 to 0.92 ± 0.04 , and the peak signal-to-noise ratio (PSNR) ranging from 32.86 ± 0.94 to 34.91 ± 1.04 dB. Notably, performance on an external validation dataset of 60 patients yielded comparable metrics: a MAE of 75.22 ± 11.81 HU, an SSIM of 0.90 ± 0.03 and a PSNR of 33.52 ± 2.06 , confirming robust cross-center generalization despite differences in imaging protocols and scanner types, without additional training. Furthermore, a visual analysis of the results revealed that the obtained metrics were influenced by registration errors.

Conclusions: Our findings demonstrated the technical feasibility of FL for CBCT-to-sCT synthesis task while preserving data privacy, offering a collaborative solution for developing generalizable models across institutions without requiring data sharing or center-specific models.

KEYWORDS

CBCT, data privacy, data sharing, deep learning, federated learning, head and neck cancer, image-to-image translation, synthetic computed tomography

1 Introduction

Radiotherapy (RT) has seen significant advancements in recent decades, particularly with the integration of image-guided radiotherapy (IGRT). IGRT involves the use of frequent imaging during treatment to account for anatomical changes and improve the accuracy of radiation delivery. Cone-beam computed tomography (CBCT) scanners are often incorporated into the gantry of linear accelerators, facilitating their integration into clinical practice. Consequently, CBCT emerged as one of the most widely adopted IGRT modalities (1–3). The advent of CBCT addressed the limitations of planar imaging, previously dominant in early image guidance, by enabling volumetric imaging directly within the treatment suite, significantly improving treatment precision and clinical workflows (2, 4, 5). This technological evolution enabled improved alignment, visualization of patient anatomy, and real-time verification of target localization, allowing clinicians to manage both inter- and intra-fractional anatomical variability. As a result, CBCT became a pivotal technology in the development of online image-guided adaptive radiotherapy (IGART) approaches (4–7). Despite these advantages, CBCT is hindered by several drawbacks, including increased image noise, reduced soft-tissue contrast, and reconstruction artifacts, which lead to unreliable Hounsfield Unit (HU) values and limit its applicability in electron density estimation for direct dose calculation (1, 5, 6). To overcome these limitations, the concept of synthetic CT (sCT) generation was introduced, where CBCT images are translated into equivalent synthetic CT images with improved dosimetric fidelity. This approach avoids unnecessary radiation exposure and reduces clinical workload by eliminating the need for additional scans (3, 5, 8).

The CBCT-to-sCT conversion task has evolved considerably, with applications across diverse anatomical regions, including head and neck, thorax, pelvis, prostate, abdomen, and pancreas, highlighting both its broad clinical relevance and intrinsic methodological challenges (3, 9). Among the strategies developed for sCT generation, deep learning (DL) approaches have proven especially effective, demonstrating strong capabilities in learning complex, non-linear mappings between image domains and producing high-quality synthetic images suitable for adaptive RT applications (1, 3, 5, 8).

Convolutional Neural Networks (CNNs) have been widely adopted in medical image analysis due to their ability to extract both local and global features, making them well suited for enhancing CBCT image quality in CBCT-to-sCT synthesis tasks. In this context, the U-Net architecture (10), has demonstrated notable efficacy in image-to-image translation tasks by enabling pixel-level predictions while preserving spatial and contextual information (3, 9). Generative Adversarial Networks (GANs) have also been extensively employed, particularly conditional GANs such as Pix2Pix (11), which utilize a generator-discriminator framework and a PatchGAN discriminator to promote local structural consistency and realistic texture synthesis. Cycle-GAN architectures further introduced cycle consistency constraints to preserve anatomical structures during CBCT-CT translation, although training instability has been reported (3, 9). More recent approaches have explored Transformer-based architectures and Denoising Diffusion

Probabilistic Models (DDPMs) for sCT generation. Transformer-based models, such as TransCBCT (12), employ self-attention mechanisms to capture long-range spatial dependencies and complex anatomical relationships, but require large training datasets, substantial computational resources, and are sensitive to image resolution (3, 12). DDPMs, which learn to reverse a progressive noising process, have demonstrated high-fidelity sCT generation when conditioned on CBCT inputs, but are similarly computationally intensive and characterized by slow inference and high hardware requirements (3). Despite their architectural differences, all of the aforementioned models have demonstrated promising performance in producing clinically viable sCTs for radiotherapy planning (3, 9).

The evaluation of sCT generation DL models in predicting accurate images comparable to gold standard planning CTs, also named ground-truth CT, was assessed by computing metrics such as mean absolute error (MAE), mean squared error (MSE), peak signal-to-noise ratio (PSNR), and structural similarity index (SSIM), computed within patient-specific masks to exclude background regions. Among these metrics, the MAE was often used as loss function, and was referred to as masked MAE or L1 Loss (3, 9). Gradient descent-based optimization has been extensively adopted for the iterative update of model weights. Concurrently, pre-processing methodologies such as image registration, image normalization and data augmentation have been demonstrated to be critical for model performance and stability (3, 8, 9, 13).

Despite the increasing adoption of DL for the sCT generation, its application remains limited by institutional heterogeneity, scanner-dependent variations, and heterogeneous imaging protocols, often necessitating single-site model training and constraining clinical scalability (3, 7, 14). Although multi-center data aggregation could partially address these limitations, it is severely constrained by strict data privacy regulations, such as the General Data Protection Regulation (GDPR)¹ in Europe, and the Health Insurance Portability and Accountability Act (HIPAA)² in the United States, which restrict the sharing and centralization of sensitive medical data (14–16). These limitations emphasize the necessity for collaborative solutions that facilitate access to diverse and extensive datasets without necessitating direct data sharing or centralized storage.

In this context, federated learning (FL) has emerged as a decentralized training paradigm that enables institutions to collaboratively train DL models while keeping data local, thereby maintaining compliance with privacy regulations and enhancing cross-center collaboration (17–19). FL has gained increasing attention in medical imaging, with applications in classification and segmentation tasks (16, 17, 20–23). Beyond the domains of classification and segmentation, federated approaches recently emerged in the context of medical image synthesis. For instance, FedMed-GAN has been developed to synthesize cross-modality brain magnetic resonance images while leveraging simple federated aggregation techniques such as the federated averaging (FedAvg), which averaging the locally updated model weights

¹<https://gdpr-info.eu/>.

²<https://www.hhs.gov/hipaa/index.html>.

from multiple clients to obtain a global model without sharing raw data (24, 25).

FedSynthCT-Brain was the first study to demonstrate the successful application of FL to the sCT synthesis task, specifically from brain T1-weighted magnetic resonance imaging (MRI) data, without compromising performance on external datasets (14). This finding suggested potential applications for extending the federated concept to the context of CBCT-to-sCT translation, where the development of robust, center-agnostic and generalizable models is essential for clinical adoption (8).

More in details, while MRI-to-CT translation involves inferring HU values from images which cannot inherently encode them, CBCT measures x-ray attenuation but suffers from severe quality degradation. Thus, the main challenge in the CBCT-to-CT translation task shifts from learning a cross-modality mapping to restoring accurate HU values from noisy, scattered and vendor-dependent reconstructed images. Furthermore, the transition of the model working region, from the brain (introduced by FedSynthCT-Brain) to the head and neck, introduces greater anatomical variability and high-contrast structures, such as dental implants and extensive air cavities. Consequently, this transition results in a task that is substantially different and more sensitive to misalignment and residual registration errors. These factors make the successful deployment of FL in this setting non-straightforward and highlight the need for a dedicated framework to determine whether federated training can remain robust and viable under such adverse and heterogeneous conditions.

Therefore, in this study, we proposed a cross-silo horizontal FL approach for CBCT-to-sCT in the head and neck region, extending the FedSynthCT (14) framework to a different imaging modality and a larger anatomical region. A federated DL model was collaboratively trained using data from different

European centers of the public SynthRAD2025 challenge dataset (26). The approach was validated not only on each federated center but also on an independent dataset outside the federation. This validation process was used to assess model generalization capabilities, thereby demonstrating that the proposed approach can result in more robust and generalizable models while preserving data privacy.

2 Materials and methods

2.1 Datasets

The datasets employed in this work were from the publicly available SynthRAD2025 challenge repository (26, 27). The institutions (also referred to as centers or clients) focus on head and neck (HN) region and were labeled as centers A, B, C, and E. Center D was deliberately excluded from this study due to the limited availability. For consistency, we adopted the same naming convention for the centers as used in the challenge.

As reported, a few thorax scans were included in the HN cohort (26); these outlier cases were excluded from the study. The final number of subjects included per center is reported in Table 1. For each subject, the dataset included a CBCT scan, a CT scan, and a binary segmentation mask delineating the patient body.

Variability in imaging protocols and scanner types among relevant centers has been identified. Notably, CBCT scans from Center E were acquired using a different system, in contrast to the devices employed in centers A, B and C. Center E also exhibited markedly longer exposure times and higher in-plane resolution variability. Regarding the CT images, it was observed that all centers employed standard 120 kVp acquisitions; however, exposure time, slice thickness and pixel spacing varied.

TABLE 1 Overview of the key CBCT and CT original acquisition parameters used at Centers A, B, C, and E according to the dataset documentation (26).

Parameter	Center A	Center B	Center C	Center E
CBCT				
Patient #	60	65	63	63
Scanner	Elekta XVI v5.x	Elekta XVI v5.52	Elekta XVI v5.x	Varian TrueBeam OBI
kVp	100–120	100	120	100–125
Exposure Time [ms]	10–32	10	22	7,500–18,060
Rows/Columns	270	270	270–512 × 270–512	512 × 512
Pixel spacing [mm]	1 × 1	1 × 1	1 × 1	0.5–0.9 × 0.5–0.9
Slice thickness [mm]	1	1	1	2
CT				
Patient #	60	65	63	63
Scanner	Philips Big Bore (90), Siemens Biograph40 (10)	Toshiba Aquilion/LB	Philips Brilliance Big Bore (93), Siemens Biograph40 (7)	Toshiba Aquilion/LB, Siemens Biograph128
kV	120	120	120	120
Exposure Time [ms]	615–1,000	500–1,000	922–1457	500–1,000
Rows/Columns	512	512	512	512
Pixel spacing [mm]	0.7–1.4 × 0.7–1.4	1.1–1.4 × 1.1–1.4	1–1.2 × 1–1.2	1–1.5 × 1–1.5
Slice thickness [mm]	2–3	1–3	2–3	3–5

2.2 Federation setup and pre-processing

The federated training setup included a central server associated with Center A dataset and three clients from Centers B, C, and E. An overview of the proposed FL framework is illustrated in Figure 1. The federated infrastructure was implemented using Flower (28), PyTorch (29) and MONAI (30). The federated training followed a cross-silo, synchronous, and full-participation scheme. At each global round, all clients performed local updates simultaneously, and the server aggregated the global model parameters only after receiving updates from every client. Experiments were conducted on a single-node machine with an NVIDIA A100 (80GB), 16 CPU cores, and 128GB RAM. While all experiments were conducted within a single-node environment for prototyping purposes, the FL framework incorporated all the necessary components to replicate a real-world federated deployment, including encrypted communication using TLS certificates, client reconnection strategies, and the ability to resume training from saved checkpoints.

The Center A dataset was reserved exclusively for external validation, enabling an evaluation of the federated model's generalization performance. Within each participating client, four validation patients and four test patients were held out, while the remaining cases were used to train local models. This allocation reflects a trade-off between maximizing client training data and maintaining a diagnostic set for monitoring local fitting behavior, with the primary generalization assessment assigned to the external cohort from Center A. Overall, of the 251 patients included in the study, 167 contributed to federated

training ($\approx 66.5\%$ of the total data), 24 were allocated to local validation and test sets across the three clients ($\approx 9.5\%$ of the total data), and 60 served as the independent external test cohort, representing $\approx 24\%$ of the total data.

Since data were released for a public challenge, several pre-processing operations were applied by the authors prior to the publication, such as isotropic resampling, defacing, and rigid registration between modalities, resulting in residual anatomical misalignments between CBCTs and CTs (26). To mitigate the impact of these residual misalignments on model evaluation, an additional deformable registration step was applied exclusively to the test patients.

To ensure consistent spatial resolution across the federation, each client applied a standardized pre-processing pipeline to locally adapt data to a $256 \times 256 \times 256$ image grid dimension. This decentralized approach—which avoided any inter-client data exchange—involved conditional cropping followed by resizing and constant-padding as needed. Furthermore, an intensity processing was applied. A local normalization process was performed on each client for each CBCT scan, employing a LUT-based normalization method proposed by Vicario et al. (31) with a fixed intensity range of $[-800, 2,000]$. Subsequently, CT scans were clipped to the range $[-1,000, 1,500]$ to mitigate the impact of extreme values introduced by metal implants. These ranges were empirically chosen by analyzing the CBCTs and CTs intensity distributions available at each client site, without exchanging any images.

This preliminary harmonization step allowed for the approximation of a common intensity range across clients while preserving the constraints of the federated paradigm.

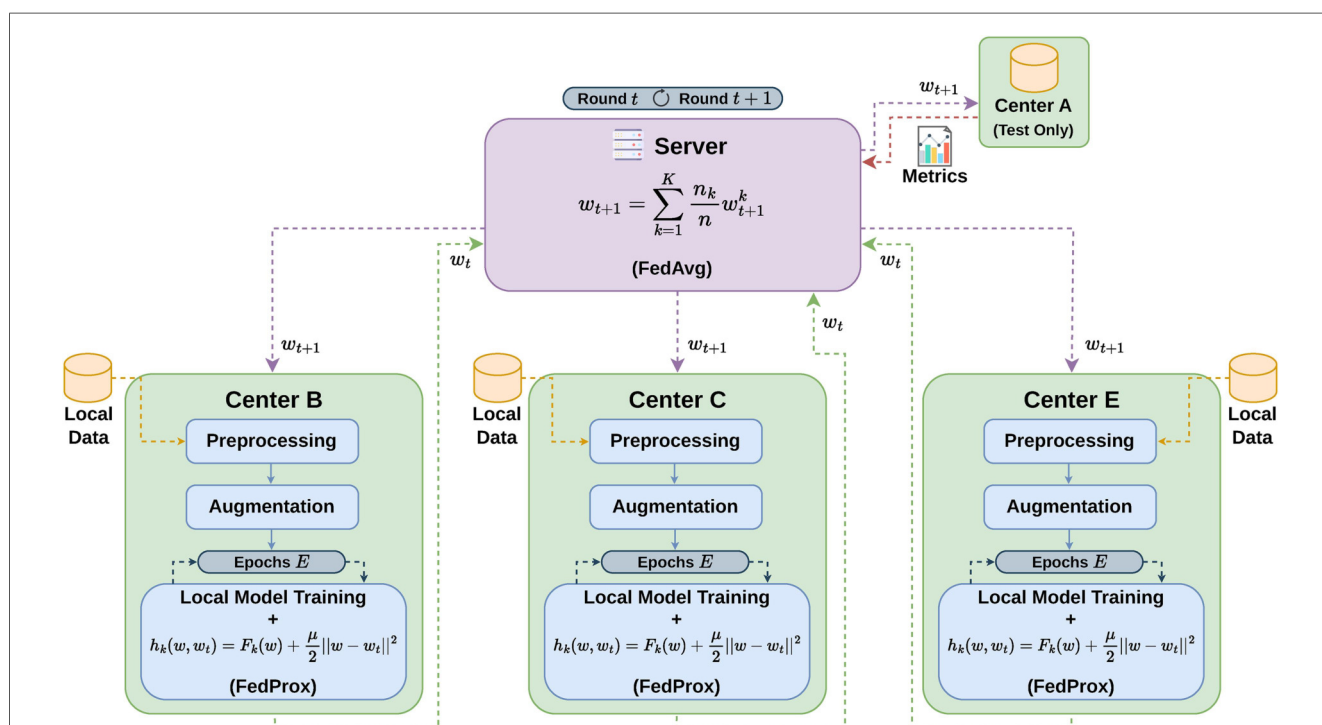


FIGURE 1

Schematic overview of the proposed FL framework. Three clients (Centers B, C, and E) implement a standardized preprocessing and data augmentation pipeline for their local datasets. Each client k performs local model training at round t , introducing the FedProx regularization. The locally updated weights (w_t) are subsequently transmitted to a central server, where they are aggregated via FedAvg. The resulting global model (w_{t+1}) is subsequently redistributed to all clients for the next round. The Center A dataset is exclusively reserved for external validation, thereby enabling an evaluation of the federated model's generalization performance.

2.3 Deep learning model

The DL architecture employed relied on the cGAN paradigm, and more specifically, on the original Pix2Pix (11) architecture.³ Pix2Pix has been identified as one of the most widely adopted architectures for the CBCT-to-sCT translation task, as indicated in Section 1. The adversarial paradigm exhibited a remarkable capacity in managing imperfectly paired or partially aligned data, a prevalent scenario in clinical practice where precise voxel-wise correspondence between images cannot be assured, even after the registration process. This has led to the identification of Pix2Pix as a suitable and practical solution for our application setting.

The generator was structured as a 2D U-Net with eight encoder-decoder blocks connected via skip connections, and included dropout layers to mitigate overfitting. The discriminator was implemented as a 2D PatchGAN, composed of convolutional blocks combining 2D convolutions, instance normalization, and LeakyReLU activations to assess local realism at the patch level.

The training of the DL model was implemented using the Random-Multi2D sampling approach, to enhance the robustness and generalisability. This approach consisted of presenting to the model different axial, sagittal, and coronal slices in a randomized order, ensuring that the DL model received non-sequential input from multiple anatomical planes (14).

In detail, the discriminator's objective (Equation 1) was achieved by minimizing the sum of two sigmoid cross-entropy losses:

$$\mathcal{L}_D = \frac{1}{2} \times [\text{BCE}(\mathcal{D}(CT, CBCT), 1) + \text{BCE}(\mathcal{D}(sCT, CBCT), 0)] \quad (1)$$

where $\text{BCE}(\cdot)$ was the cross-entropy loss output, $\mathcal{D}(\cdot)$ was the discriminator output, and 1 and 0 were images of ones and zeros indicating real and fake labels, respectively. The first term penalized misclassifications of CTs as fake, while the second term penalized misclassifications of sCTs as real.

The adversarial loss function employed for the optimization of the generator (Equation 2) was combined by two components. The first was a sigmoid cross-entropy loss that enabled the discriminator to classify the generated images as real or fake:

$$\mathcal{L}_{GAN} = \text{BCE}(\mathcal{D}(sCT, CBCT), 1). \quad (2)$$

The second term was the L1 loss (Equation 3), specifically the Masked MAE (see Equation 7)—typically employed for this task as reported in Section 1—between the generated sCT and the ground-truth CT, which enforced structural similarity and penalized large deviations. This loss was also referred to as pixel loss:

$$\mathcal{L}_{\text{pixel}} = \text{MAE}(CT - sCT). \quad (3)$$

The total generator loss (Equation 4) was subsequently computed as:

$$\mathcal{L}_G = \mathcal{L}_{GAN} + \lambda_{\text{pixel}} \cdot \mathcal{L}_{\text{pixel}}, \quad (4)$$

where $\lambda_{\text{pixel}} = 100$ as reported in the original Pix2Pix study (11).

Inference was based on the median voxel-voting mechanism. Specifically, following the generation of the slice-wise predictions along each anatomical plane, the final sCT volume was computed by taking the median across the three orientations (1, 14). Both training and inference strategies were previously validated in the federated scenario within the FedSynthCT-Brain framework (14).

In each local client, the model was trained using a batch size of 32 slices before participating in a federated aggregation round. The Adam optimizer with a learning rate of 10^{-4} was used. An additional investigation was conducted to establish the optimal number of local epochs, comparing 1, 2 and 3 epochs per round.

To further improve robustness and cross-site generalization, a consistent data augmentation pipeline was applied at the client level. Each client independently introduced random geometric and intensity transforms during the training phase, including flipping, rotations, translations, and random rescaling or shifting of intensity values.

2.4 Aggregation strategy and sensitivity analysis

The combination of FedAvg and FedProx (32) was employed as aggregation strategy for this study. The aggregation was equivalent with the approach used in the FedSynthCT-Brain study, as it was identified as the optimal aggregation strategy, yielding results that are equivalent to results obtained by alternative strategies while requiring a reduced number of rounds (14). Nevertheless, in order to determine the optimal proximal coefficient μ and to verify that the hyperparameter choice originally established in the FedSynthCT-Brain study (14) remains appropriate for the CBCT-to-sCT synthesis task, a sensitivity analysis was conducted.

The FedAvg algorithm was implemented on the server side to aggregate the clients' weights, ensuring that the contribution of each client was proportional to the size of its dataset. Therefore, in the Equation 5, K represents the total amount of clients; n_k denotes the number of samples for a single client k ; n is the total amount of samples across the federation, and w_{t+1}^k represents the post-training weights of the client k .

$$w_{t+1} = \sum_{k=1}^K \frac{n_k}{n} w_{t+1}^k \quad (5)$$

The FedProx method was subsequently implemented on the client side, thereby introducing a proximal term, denoted by $\frac{\mu}{2} \|w - w_t\|^2$ to the local objective function (thus, the Masked MAE loss). This term penalized large deviations from the global model weights in order to control local update variations. The Equation 6 provides a formal explanation of the impact of the proximal term on the client model. The original local objective

³<https://github.com/phillipi/pix2pix>.

is denoted by $F_k(w)$, the pre-training global model weight by w_t , the weights during local training by w , and the proximal coefficient by μ . Consequently, FedProx mitigates update divergence and approximates FedAvg when $\mu = 0$, as evidenced in the original FedProx study (32).

$$h_k(w, w_t) = F_k(w) + \frac{\mu}{2} \|w - w_t\|^2 \quad (6)$$

To assess the sensitivity of the aggregation strategy to the proximal coefficient μ , four configurations were evaluated: (i) $\mu = 0$ (FedAvg), (ii) $\mu = 1$, (iii) $\mu = 3$, and (iv) $\mu = 5$. All other hyperparameters (presented in Section 2.3) were held constant across configurations. Each experiment was run for a total of 50 global rounds, ensuring sufficient training time for convergence and enabling fair comparisons across different μ configurations.

2.5 Evaluation

The performance of the federated model was evaluated using standard image similarity metrics—presented in Section 1—on the external test dataset (Center A) and the client test datasets (Centers B, C, and E). Specifically, the following metrics were employed:

- **Masked Mean Absolute Error:**

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |CT_i - sCT_i|, \quad (7)$$

where n is the number of voxels within the region of interest (ROI).

- **Masked Structural Similarity Index Measure (SSIM):**

$$\text{SSIM} = \frac{(2\mu_{sCT}\mu_{CT} + C_1)(2\sigma_{sCT,CT} + C_2)}{(\mu_{sCT}^2 + \mu_{CT}^2 + C_1)(\sigma_{sCT}^2 + \sigma_{CT}^2 + C_2)}, \quad (8)$$

where the mean (μ), variance/covariance (σ), and constants $C_1 = (k_1 L)^2$, $C_2 = (k_2 L)^2$ are computed over the ROI. Here, L is the dynamic range of the CT image, with $k_1 = 0.01$ and $k_2 = 0.03$.

- **Masked Peak Signal-to-Noise Ratio (PSNR):**

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}_{CT}^2}{\text{MSE}} \right), \quad (9)$$

where MAX_{CT} is the maximum intensity value in the CT image and MSE is the mean squared error between CT and sCT within the ROI.

In order to assess the impact of registration errors on image similarity metrics, a registration quality analysis was conducted on the external test cohort from Center A, which was sufficiently large to enable a correlation analysis. For each patient, the Normalized Mutual Information (33) (NMI) between the CT and CBCT volumes was computed both before and after the additional deformable registration step. Thus, the

NMI was used as an intensity-based surrogate for registration quality. The patient-specific variation in NMI was defined as:

$$\Delta \text{NMI} = \text{NMI}_{\text{after}} - \text{NMI}_{\text{before}} \quad (10)$$

This variation was then correlated with the corresponding change in generation error:

$$\Delta \text{MAE} = \text{MAE}_{\text{after}} - \text{MAE}_{\text{before}} \quad (11)$$

where $\text{MAE}_{\text{after}}$ and $\text{MAE}_{\text{before}}$ denote the MAE (see Equation 7), obtained by using as input the CBCT volumes after and before the deformable registration step, respectively.

3 Results

The performance of the proposed federated cGAN model across multiple clinical centers was assessed through both quantitative and qualitative analyses, as reported in Table 2 and Figures 3–6, respectively.

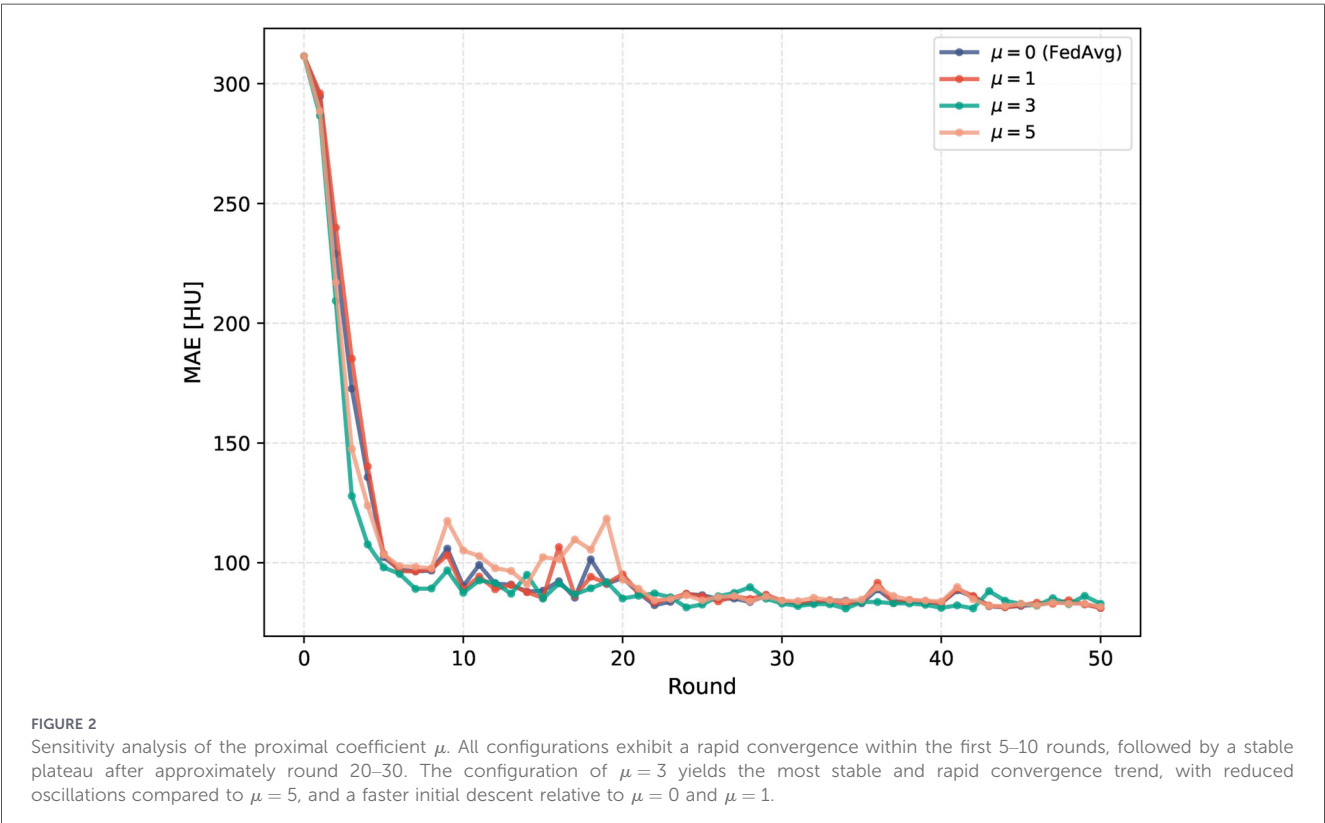
The convergence curves obtained across different μ configurations ($\mu = 0, 1, 3, 5$) are presented in Figure 2. In all configurations, the global MAE demonstrated a rapid decrease during the initial 5–10 global rounds, from ≈ 310 HU to values in the range of 90–130 HU, followed by a stable plateau beyond round 30–35, with round-to-round fluctuations remaining below ± 5 HU. The quantitative metrics across all configurations demonstrated high consistency (Table 2), with marginal variations in MAE, SSIM (Equation 8) and PSNR (Equation 9). Among the tested configurations, $\mu = 3$ exhibited the most stable convergence trajectory, with reduced oscillations during the intermediate rounds compared to $\mu = 5$, and a faster initial descent relative to $\mu = 0$ (FedAvg) and $\mu = 1$. Therefore, the $\mu = 3$ configuration was selected on the basis of its convergence properties, rather than on marginal differences in image quality metrics.

Subsequently, an additional investigation on the optimal number of local epochs was conducted (see Section 2.3). The convergence curves and quantitative metrics are provided in Supplementary Figure S1 and Supplementary Table S1, respectively. The federated training with a single local epoch led to a notable decrease in the stability of the global model convergence, exhibiting increased oscillations across rounds. As the number of local epochs per round increased (2 and 3), a proportional increase in computation time was observed, along with a progressively faster early descent and smaller oscillations in the intermediate rounds. Furthermore, a comparable outcome was observed in configurations with 2 and 3 local epochs, where convergence plateau was reached around round 30. However, no substantial differences were observed in the image similarity metrics. Indeed, MAE differences across centers ranged between 1–4 HU, with no single configuration consistently outperforming the others across all centers (see Supplementary Table S1). Consequently, given that increasing the number of local epochs failed to yield a substantial improvement in performance and instead led to an increase in the computational load on local clients, while a single epoch introduced instability in global convergence, adopting two local epochs was

TABLE 2 Quantitative performance across all centers for the four proximal coefficient μ configurations evaluated in the sensitivity analysis.

Center	Configuration	μ	MAE [HU]	SSIM	PSNR [dB]
Center A (external)	FedAvg	0	75.17 $\pm$ 12.54	0.90 $\pm$ 0.04	33.63 $\pm$ 2.18
	FedAvg + FedProx	1	75.57 $\pm$ 12.61	0.90 $\pm$ 0.04	33.61 $\pm$ 2.14
	FedAvg + FedProx	3	75.22 $\pm$ 11.81	0.90 $\pm$ 0.03	33.52 $\pm$ 2.06
	FedAvg + FedProx	5	75.80 $\pm$ 12.51	0.90 $\pm$ 0.04	33.59 $\pm$ 2.13
Center B	FedAvg	0	81.92 $\pm$ 5.89	0.88 $\pm$ 0.02	33.09 $\pm$ 0.93
	FedAvg + FedProx	1	82.22 $\pm$ 6.11	0.88 $\pm$ 0.02	33.07 $\pm$ 0.93
	FedAvg + FedProx	3	85.90 $\pm$ 7.10	0.88 $\pm$ 0.02	32.86 $\pm$ 0.94
	FedAvg + FedProx	5	82.15 $\pm$ 6.30	0.89 $\pm$ 0.02	33.06 $\pm$ 0.88
Center C	FedAvg	0	64.75 $\pm$ 14.38	0.92 $\pm$ 0.04	34.38 $\pm$ 1.38
	FedAvg + FedProx	1	65.02 $\pm$ 14.43	0.92 $\pm$ 0.04	34.38 $\pm$ 1.35
	FedAvg + FedProx	3	64.38 $\pm$ 13.63	0.92 $\pm$ 0.04	34.48 $\pm$ 1.39
	FedAvg + FedProx	5	65.86 $\pm$ 14.68	0.92 $\pm$ 0.04	34.27 $\pm$ 1.37
Center E	FedAvg	0	75.55 $\pm$ 13.40	0.91 $\pm$ 0.02	34.74 $\pm$ 1.03
	FedAvg + FedProx	1	74.71 $\pm$ 12.93	0.92 $\pm$ 0.01	34.81 $\pm$ 0.98
	FedAvg + FedProx	3	74.22 $\pm$ 13.23	0.92 $\pm$ 0.01	34.91 $\pm$ 1.04
	FedAvg + FedProx	5	74.14 $\pm$ 12.49	0.92 $\pm$ 0.01	34.90 $\pm$ 0.98

Results are reported as mean $\pm$ standard deviation. Center A was used exclusively for external testing, while Centers B, C, and E participated in the federated training.



determined to be an optimal trade-off, balancing the stability of the global convergence trend, the performance of the federated model, and the computational cost on clients.

Therefore, all the subsequent per-center results reported herein refer to the $\mu = 3$ and 2 local epochs configuration.

As detailed in [Section 2.2](#), Center A dataset was excluded from the federated training process and served as an external test set, including 60 test cases that provided a more robust and less biased estimation of the model’s generalization capabilities. Centers B, C, and E contributed data to train the federated

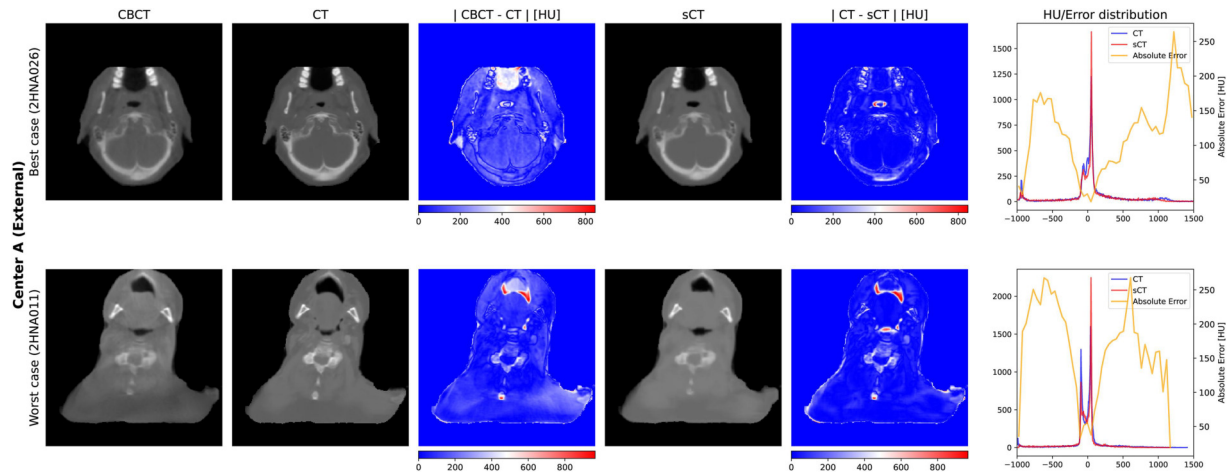


FIGURE 3

Visual evaluation of the best and worst sCT cases from Center A, used exclusively as an external test set. For each case, the mid-axial slice of the original CBCT, the ground-truth CT, the generated sCT, and the absolute difference maps for CBCT-to-CT and sCT-to-CT (error map) are shown. Additionally, intensity distributions of the sCT and CT for the displayed slice are presented, overlaid with the pixel-wise MAE. These distributions highlight the conversion accuracy across different tissue density ranges. Notably, regions with very high absolute differences appear in both CBCT-to-CT and error maps, suggesting the presence of residual misregistration between CBCT and CT scans.

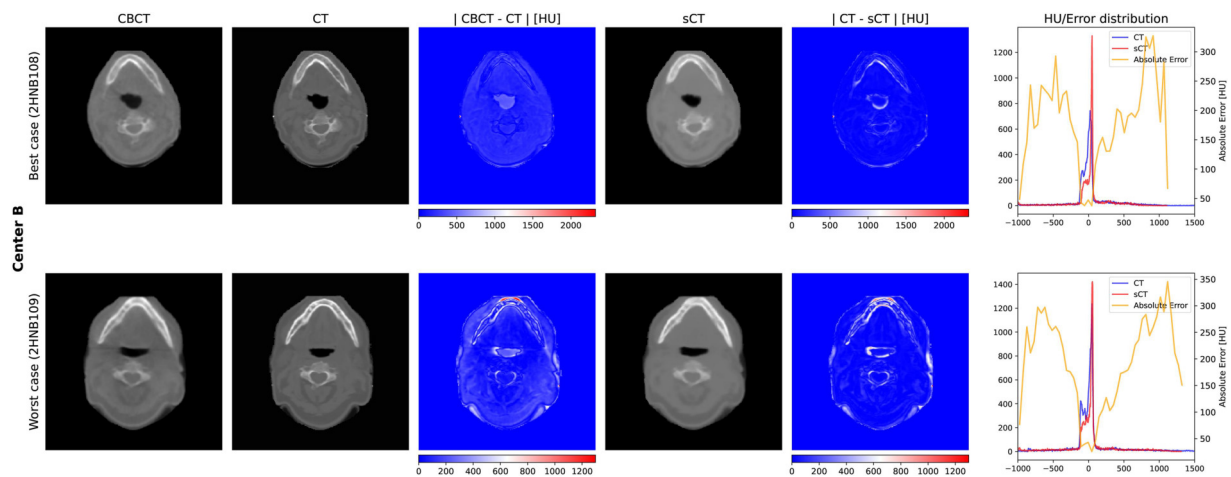


FIGURE 4

Qualitative results for the best and worst sCT cases from Center B, which showed lower image similarity metrics among the training centers. Widespread registration artifacts between CBCT and CT scans contributed to the visible errors. This was especially evident in the oral cavity of the 2HNB109 patient, and are reflected in both $|CBCT - CT|$ figure and error map.

model, with each center reserving four patients for local evaluation. As shown in Table 2 (for $\mu = 3$), despite not participating in the federated training, Center A achieved performance metrics comparable to those of the training centers, with a MAE of 75.22 ± 11.81 HU, an SSIM of 0.90 ± 0.03 , and a PSNR of 33.52 ± 2.06 dB. A qualitative evaluation of the best and worst test cases from Center A (Figure 3) showed that the generated sCTs aligned closely with the ground-truth CTs. The error maps, computed as $|CT - sCT|$, indicated that most deviations were within ± 100 HU. The absolute difference between the original CBCT and the ground-truth CT, reported as $|CBCT - CT|$ in Figure 3, highlighted spatial discrepancies mainly related to anatomical variations and registration mismatches.

Among the federated clients, Center B obtained a MAE = 85.90 ± 7.10 HU, an SSIM = 0.88 ± 0.02 , and a PSNR = 32.86 ± 0.94 dB. Qualitative results (Figure 4) demonstrated that the generated sCTs were anatomically consistent with the ground-truth CTs, particularly in the representation of major structures.

Center C yielded the best overall quantitative performance, with the lowest MAE (64.38 ± 13.63 HU), highest SSIM (0.92 ± 0.04), and highest PSNR (34.48 ± 1.39 dB). The worst-case example (2HNC123) shown in Figure 5 showed that the dominant discrepancies were associated with misalignments between CBCT and CT rather than pure generation errors.

Center E achieved results consistent with the overall federation despite images being acquired using a different scanner and

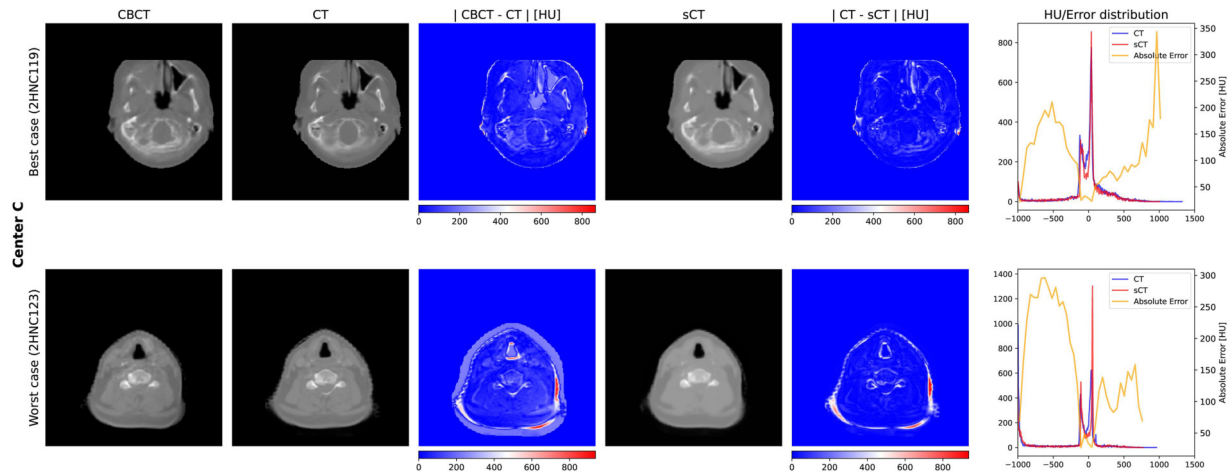


FIGURE 5

Visual comparison of the best and worst sCT cases from Center C. Compared to the other centers, Center C showed fewer registration-related errors and better image similarity, as evidenced by both the error maps ($|CT - sCT|$) and the distribution plots. However, case 2HNC123, identified as the worst-performing example, exhibited evident misalignments prior to the sCT generation despite the additional registration process.

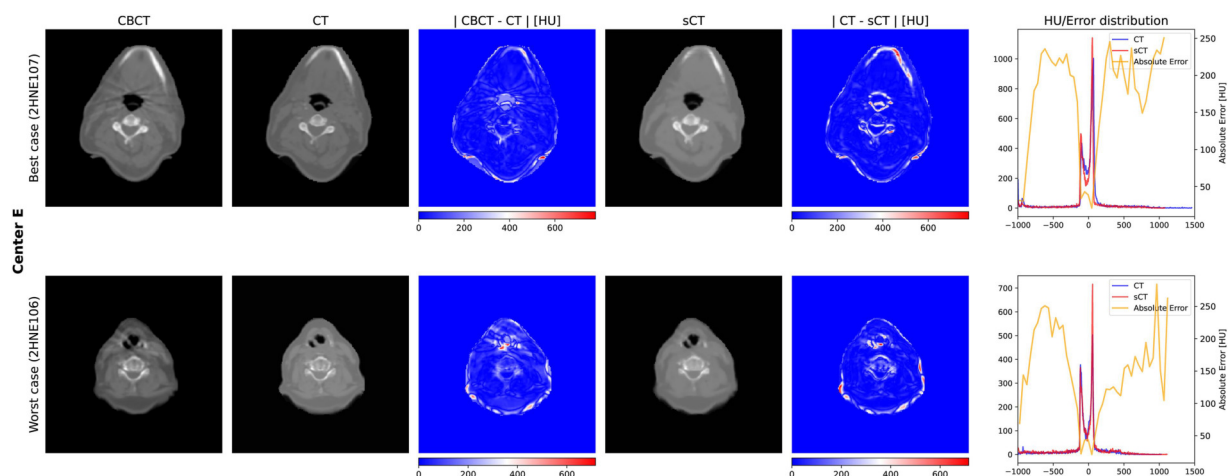


FIGURE 6

Visual evaluation of the best and worst sCT cases from Center E. Despite the use of a different scanner and acquisition protocol, the model maintained strong performance, with distribution plots indicating consistent HU alignment and localized discrepancies, mostly in air and bone regions due to inter-scan misalignments.

acquisition protocol (Table 1). The federated model achieved a MAE = 74.22 ± 13.23 HU, an SSIM = 0.92 ± 0.01 , and a PSNR = 34.91 ± 1.04 dB (Table 2). Furthermore, quantitative results confirmed the quality of the generated sCTs, as shown in Figure 6.

The results of the registration quality analysis described in Section 2.5 are presented in Figure 7. A significant negative correlation between ΔNMI (Equation 10) and ΔMAE (Equation 11) was observed in the Center A cohort, with a Pearson coefficient of $r = -0.63$ ($p = 5.95 \times 10^{-8}$) and a Spearman coefficient of $\rho = -0.61$ (2.35×10^{-7}).

4 Discussion

In this work, we investigated whether a cross-silo FL framework can support robust CBCT-to-sCT synthesis in the

head and neck region under realistic multi-center constraints. Specifically, we extended the FedSynthCT (14) study to a different imaging modality and a more anatomically heterogeneous region, evaluating its performance across three federated European centers and an independent external validation cohort from the SynthRAD2025 dataset (27). By analyzing both quantitative metrics and qualitative image characteristics, we aimed to determine whether federated training can preserve generalization capability despite inter-institutional variability in scanners, acquisition protocols, and anatomical presentation, without requiring data sharing or center-specific retraining.

A sensitivity analysis was conducted on the proximal coefficient μ , which confirmed the suitability of $\mu = 3$ for the current task in accordance with the findings of FedSynthCT-Brain (14), despite the acknowledged differences in imaging modality, anatomical region, and scanner heterogeneity. As

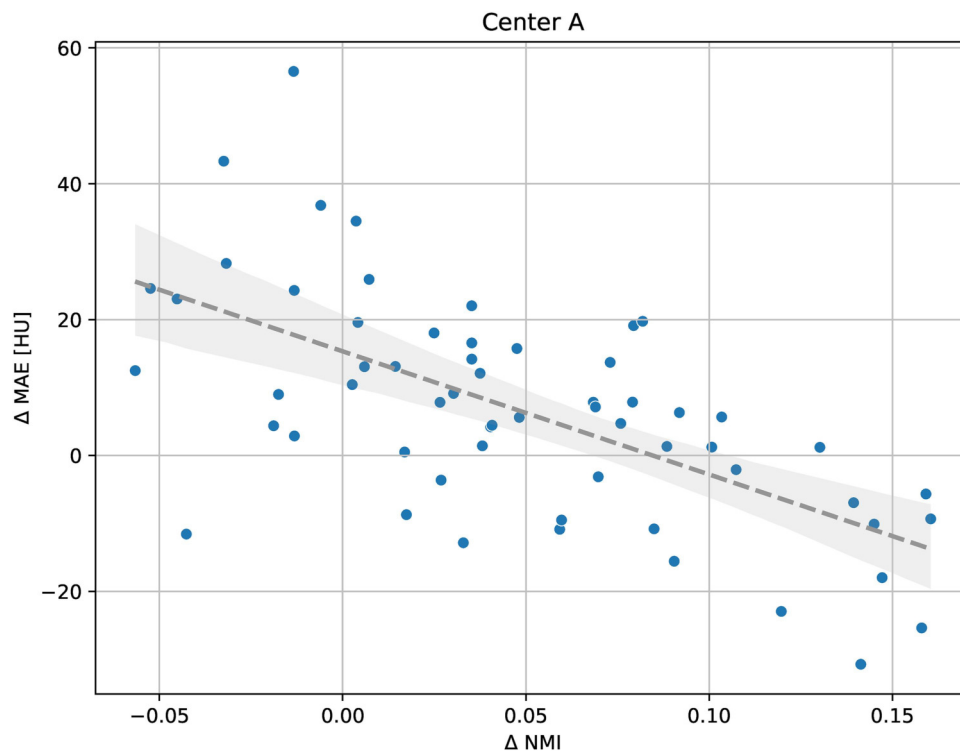


FIGURE 7

Correlation between Δ NMI and Δ MAE (HU) for the Center A test cohort before and after the deformable image registration. A significant negative correlation was observed (Pearson $r = -0.63$, 5.95×10^{-8} ; Spearman $\rho = -0.61$, $p = 2.35 \times 10^{-7}$). This finding indicates that enhancements in deformable registration, consequently leading to enhanced intensity correspondence between CT and CBCT, are associated with a decrease in MAE. The deterioration of NMI corresponds to MAE increases of ≈ 20 – 60 HU (upper-left area). In contrast, cases for which NMI improved substantially Δ NMI $\approx 10\%$ – 15% exhibit a consistent MAE reduction (lower-right area).

demonstrated in Figure 2, the convergence behavior exhibited that moderate regularization ($\mu = 3$) accelerated and stabilized convergence in comparison to FedAvg ($\mu = 0$), $\mu = 1$ and $\mu = 5$. Indeed, for $\mu < 3$ the convergence was slower and exhibited greater fluctuations in the intermediate rounds (approximately rounds 9–20), while more severe regularization ($\mu = 5$) introduced instability in the intermediate rounds, delaying optimization. Consequently, this finding indicates that proximal regularization is beneficial, but leads to over-constraining local updates and hinders global convergence when severe. Notably, the minimal variations in the final image quality metrics across all μ configurations (Table 2), suggest that the convergence benefit of $\mu = 3$ occurs without compromising final performance. Furthermore, the consistent plateau behavior observed across all configurations demonstrated that 50 rounds were sufficient to ensure stable convergence of the federated model.

The local epochs analysis revealed that a single local epoch introduced instability and slowed convergence (see Supplementary Figure S1). The configurations with 2 and 3 local epochs yielded fewer oscillations across rounds. However, no substantial improvement in image quality metrics was observed, with MAE differences ranging from 1 to 4 HU across centers and no consistent hierarchy among configurations. Moreover, minimizing local updates in the context of FL is typically advantageous to mitigate client drift (24). Although FedProx partially constrains local divergence, its regularizing effect could

diminish as the number of local epochs increases (32). Consequently, 2 local epochs were selected as the configuration that optimally balanced convergence stability, model performance, and computational cost on local clients.

External validation on Center A enabled further investigation into the generalization capability of the federated model in a fully independent setting. The observed performance metrics were comparable to those obtained across the participating centers, despite the absence of additional training or fine-tuning. These findings suggest stable model behavior on previously unseen data. Furthermore, qualitative analysis indicated that residual spatial discrepancies (Figure 3) between CBCT and planning CT were predominantly associated with registration inaccuracies and anatomical variations occurring during treatment. For instance, anatomical differences in the oral cavity were observed in patient 2HNA011 (Figure 3); these misalignments were propagated to the generated sCT, affecting its apparent visual quality and negatively influencing the quantitative metrics reported in Table 2.

Across the federated centers, performance differences were primarily attributed to image misalignments, despite the inclusion of an additional deformable registration step for the evaluation data. The lower performance at Center B relative to Centers C, E, and the external Center A was mainly attributable to a higher prevalence of misalignments. In contrast, the superior performance at Center C reflected a reduced influence of registration errors, resulting in higher image similarity

metrics. Taken together, these observations suggest that registration quality, rather than the inherent capability of the generative model, represented a primary limiting factor in the evaluation. To provide further analytical insights, an additional investigation was conducted on the Center A test cohort using NMI as an intensity-based surrogate for registration quality (see [Section 2.5](#)). Although NMI lacks direct geometric quantification of spatial misalignment, it measures intensity correspondence between CT and CBCT volumes. A significant negative correlation was identified between ΔNMI and ΔMAE . Furthermore, [Figure 7](#) supports the hypothesis that residual misalignment systematically affects sCT evaluation. In detail, when deformable registration substantially improved intensity correspondence ($\Delta\text{NMI} \approx 10\%–15\%$), a corresponding MAE reduction was observed. In contrast, when NMI deteriorated, MAE degraded by approximately 20–60 HU (upper-left area of [Figure 7](#)).

However, the robustness of the federated model was further reflected by the results obtained at Center E, where data were acquired using a different scanner and acquisition protocol. Despite this domain shift, the model maintained performance levels consistent with the federation, highlighting its capacity to adapt across heterogeneous scanners. A global analysis across all centers, focusing on anatomical structures and HU distributions, demonstrated that the model preserved tissue-specific intensity patterns ([Figures 3–6](#)). The largest errors were consistently observed in tissues occupying smaller volumes in the head and neck region, such as air cavities and dense bone, which were underrepresented in the training data relative to soft tissues. These regions are intrinsically more challenging to predict accurately and, due to their extreme HU values, disproportionately affected image similarity metrics.

A limitation of this study concerns the quantitative evaluation strategy. As demonstrated by the NMI-MAE correlation analysis, residual misalignment systematically affected the MAE. However, its independent contribution could not be isolated precisely. Furthermore, the local test sets comprised four patients per client, with the purpose of assessing the federated model fitting at the client level. This methodological approach inherently introduces higher variability in performance estimates. Nevertheless, the findings in [Table 2](#) and [Figures 3–6](#) demonstrated consistent performance of the federated model across both local and external testing, indicating its capacity to maintain comparable outcomes for the clients and the independent center.

In addition, although the integration of FL-generated sCTs into RT treatment planning represents a potential clinical application of this work, its direct evaluation was not feasible in this study due to the absence of dosimetric and planning data in the SynthRAD2025 dataset, which contains imaging data only. This scenario also reflects a realistic clinical setting in which sensitive RT data are not shared, further underlining the relevance of FL as a framework for privacy-preserving multi-institutional model development. Thus, future research may focus on evaluating the clinical applicability of sCTs generated through FL. Furthermore, FL does not inherently guarantee complete privacy protection. Potential information leakage from shared model updates has been reported in the literature, including through gradient and model inversion attacks ([34, 35](#)). Mitigation strategies such as secure aggregation or differential

privacy can further reduce these risks ([34, 35](#)). While these strategies have not been explicitly examined in this work, they remain an important area for future research in federated sCT generation.

From a methodological perspective, federated and centralized DL represent distinct optimization paradigms rather than interchangeable training strategies. As previously observed in the FedSynthCT-Brain ([14](#)) study, methodological configurations that are effective in centralized settings may not necessarily translate directly to federated environments, where statistical heterogeneity, client-specific data distributions, and aggregation constraints fundamentally shape the learning dynamics. Centralized training optimizes performance under full data availability and implicitly benefits from pooled distributions, whereas FL aims to achieve stable cross-institutional performance while preserving data privacy and respecting structural data-sharing constraints. FL also introduces additional communication and synchronization overhead compared to centralized training, which may increase overall training time, although this cost is often offset by the practical benefits of decentralized data access and improved feasibility in multi-institutional settings. Within this framework, moderate center-specific performance trade-offs should be interpreted in light of the federated objective: promoting generalization across heterogeneous institutions rather than maximizing performance on any single site. Federated models may also serve as robust shared initializations that can be further adapted through optional local fine-tuning when institution-specific optimization is required. Accordingly, FL should be regarded not as a substitute for centralized training, but as a complementary collaborative framework tailored to clinical scenarios in which data centralization is impractical or restricted.

Our findings suggest potential directions for future research. Within the GAN paradigm, more targeted training strategies may improve the synthesis of anatomically underrepresented structures. The multi-discriminator framework proposed by Kumar et al. ([36](#)) could represent a viable approach. In parallel, task-specific loss formulations, such as those introduced by Liang et al. ([37](#)) to improve HU fidelity, noise suppression, and dosimetric accuracy, could be integrated into the FL aggregation scheme. However, this would introduce additional challenges related to client-level heterogeneity and personalization. Furthermore, beyond GAN-based architectures, the investigation of diffusion-based models and hybrid CNN-transformer approaches ([38](#)) within a federated framework represents a promising direction for improving both image quality and structural consistency in FL-generated sCTs.

In conclusion, this study demonstrated the feasibility and effectiveness of a cross-silo FL approach for CBCT-to-sCT synthesis in the head and neck region, using the publicly available SynthRAD2025 challenge dataset ([26](#)). The federated model generalized effectively across the participating centers and on an external validation cohort, maintaining consistent performance despite differences in imaging protocols, scanner types, and acquisition parameters. This mitigates the impact of institutional heterogeneity, a common limitation in single-center model deployments. Furthermore, these results align with requirements outlined in recent related work ([8](#)), which

emphasized that a model should ideally be applicable across different centers without the need for conditional tuning.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: <https://zenodo.org/doi/10.5281/zenodo.15373853>.

Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

CBR: Software, Visualization, Data curation, Writing – original draft, Investigation, Conceptualization, Validation, Methodology, Writing – review & editing, Formal analysis. PZ: Visualization, Data curation, Writing – review & editing. MS: Writing – review & editing, Data curation, Visualization, Conceptualization, Validation.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

References

- Spadea MF, Maspero M, Zaffino P, Seco J. Deep learning based synthetic-CT generation in radiotherapy and PET: a review. *Med Phys.* (2021) 48:6537–66. doi: 10.1002/mp.15150
- Boldrini L, D'Aviero A, De Felice F, Desideri I, Grassi R, Greco C, et al. Artificial intelligence applied to image-guided radiation therapy (IGRT): a systematic review by the Young Group of the Italian Association of Radiotherapy and Clinical Oncology (yAIRO). *Radiol Med.* (2023) 129:133–51. doi: 10.1007/s11547-023-01708-4
- Altalib A, McGregor S, Li C, Perelli A. Synthetic CT image generation from CBCT: a systematic review. *IEEE Trans Radiat Plasma Med Sci.* (2025) 9(6): 1. doi: 10.1109/TRPMS.2025.3533749
- Landry G, Hua C. Current state and future applications of radiological image guidance for particle therapy. *Med Phys.* (2018) 45:1086–95. doi: 10.1002/mp.12744
- Rabe M, Kurz C, Thummerer A, Landry G. Artificial intelligence for treatment delivery: image-guided radiotherapy. *Strahlenther Onkol.* (2025) 201:283–97. doi: 10.1007/s00066-024-02277-9
- Riou O, Prunaretty J, Michalet M. Personalizing radiotherapy with adaptive radiotherapy: Interest and challenges. *Cancer/Radiothérapie.* (2024) 28:603–9. doi: 10.1016/j.canrad.2024.07.007
- Mastella E, Calderoni F, Manco L, Ferioli M, Medoro S, Turra A, et al. A systematic review of the role of artificial intelligence in automating computed tomography-based adaptive radiotherapy for head and neck cancer. *Phys Imaging Radiat Oncol.* (2025) 33:100731. doi: 10.1016/j.phro.2025.100731
- Huijben EM, Terpstra ML, Galapon AJ, Pai S, Thummerer A, Koopmans P, et al. Generating synthetic computed tomography for radiotherapy: SynthRAD2023 challenge report. *Med Image Anal.* (2024) 97:103276. doi: 10.1016/j.media.2024.103276
- Rusanov B, Hassan GM, Reynolds M, Sabet M, Kendrick J, Rowshanfarzad P, et al. Deep learning methods for enhancing cone-beam CT image quality toward adaptive radiation therapy: a systematic review. *Med Phys.* (2022) 49:6019–54. doi: 10.1002/mp.15840
- Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Cham: Springer International Publishing (2015). p. 234–41.
- Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. *arXiv [Preprint]* (2018). doi: 10.48550/arXiv.1611.07004
- Chen X, Liu Y, Yang B, Zhu J, Yuan S, Xie X, et al. A more effective CT synthesizer using transformers for cone-beam CT-guided adaptive radiotherapy. *Front Oncol.* (2022) 12:988800. doi: 10.3389/fonc.2022.988800
- La Greca Saint-Estevan A, Dal Bello R, Lapaeva M, Fankhauser L, Pouymayou B, Konukoglu E, et al. Synthetic computed tomography for low-field magnetic resonance-only radiotherapy in head-and-neck cancer using residual vision transformers. *Phys Imaging Radiat Oncol.* (2023) 27:100471. doi: 10.1016/j.phro.2023.100471
- Raggio CB, Zabaleta MK, Skupien N, Blanck O, Cicone F, Cascini GL, et al. FedSynthCT-Brain: a federated learning framework for multi-institutional brain MRI-to-CT synthesis. *Comput Biol Med.* (2025) 192:110160. doi: 10.1016/j.compbiomed.2025.110160

Conflict of interest

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

Generative AI statement

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

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

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fdgth.2026.1812254/full#supplementary-material>

15. Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat Mach Intell.* (2020) 2:305–11. doi: 10.1038/s42256-020-0186-1
16. Guan H, Yap PT, Bozoki A, Liu M. Federated learning for medical image analysis: a survey. *Pattern Recognit.* (2024) 151:110424. doi: 10.1016/j.patcog.2024.110424
17. Sandhu SS, Gorji HT, Tavakolian P, Tavakolian K, Akhbardeh A. Medical imaging applications of federated learning. *Diagnostics.* (2023) 13:3140. doi: 10.3390/diagnostics13193140
18. Dalmaz O, Mirza MU, Elmas G, Ozbey M, Dar SU, Ceyani E, et al. One model to unite them all: personalized federated learning of multi-contrast MRI synthesis. *Med Image Anal.* (2024) 94:103121. doi: 10.1016/j.media.2024.103121
19. Hernandez-Cruz N, Saha P, Sarker MMK, Noble JA. Review of federated learning and machine learning-based methods for medical image analysis. *Big Data Cogn Comput.* (2024) 8:99. doi: 10.3390/bdcc8090099
20. Sheller MJ, Reina GA, Edwards B, Martin J, Bakas S. Multi-Institutional deep learning modeling without sharing patient data: a feasibility study on brain tumor segmentation. *Brainlesion.* (2019) 11383:92–104. doi: 10.1007/978-3-030-11723-8_9
21. Li W, Milletari F, Xu D, Rieke N, Hancox J, Zhu W, et al. Privacy-preserving federated brain tumour segmentation. In: Suk HI, Liu M, Yan P, Lian C, editors. *Machine Learning in Medical Imaging*. Cham: Springer International Publishing (2019). p. 133–41.
22. Li D, Kar A, Ravikumar N, Frangi AF, Fidler S. Fed-Sim: federated simulation for medical imaging. *arXiv [Preprint]* (2020). doi: 10.48550/arXiv.2009.00668
23. Chang Q, Qu H, Zhang Y, Sabuncu M, Chen C, Zhang T, et al. Synthetic learning: learn from distributed asynchronous discriminator GAN without sharing medical image data. *arXiv [Preprint]* (2020). doi: 10.48550/arXiv.2006.00080
24. McMahan B, Moore E, Ramage D, Hampson S, Arcas BA. Communication-efficient learning of deep networks from decentralized data. In: Singh A, Zhu J, editors. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research*, Vol. 54. PMLR (2017). p. 1273–82.
25. Wang J, Xie G, Huang Y, Lyu J, Zheng F, Zheng Y, et al. FedMed-GAN: federated domain translation on unsupervised cross-modality brain image synthesis. *Neurocomputing.* (2023) 546:126282. doi: 10.1016/j.neucom.2023.126282
26. Thummerer A, van der Bijl E, Galapon AJ, Kamp F, Savenije M, Muijs C, et al. SynthRAD2025 grand challenge dataset: generating synthetic CTs for radiotherapy from head to abdomen. *Med Phys.* (2025a) 52:e17981. doi: 10.1002/mp.17981
27. Thummerer A, van der Bijl E, Galapon AJ, Kamp F, Maspero M. Data from: SynthRAD2025 grand challenge dataset: training (2025). doi: 10.5281/ZENODO.15373853
28. Beutel DJ, Topal T, Mathur A, Qiu X, Fernandez-Marques J, Gao Y, et al. Flower: a friendly federated learning research framework. *arXiv [Preprint]* (2022). doi: 10.48550/arXiv.2007.14390
29. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. Red Hook: Curran Associates Inc. (2019).
30. Cardoso MJ, Li W, Brown R, Ma N, Kerfoot E, Wang Y, et al. MONAI: an open-source framework for deep learning in healthcare. *arXiv [Preprint]* (2022). doi: 10.48550/arXiv.2211.02701
31. Vicario CM, Kordon F, Denzinger F, Barbari JSE, Privalov M, Franke J, et al. Normalization techniques for CNN based analysis of surgical cone beam CT volumes. In: *Medical Imaging 2022: Image Processing* (2022). p. 85. doi: 10.1117/12.2606973
32. Li T, Sahu AK, Zaheer M, Sanjabi M, Talwalkar A, Smith V. Federated optimization in heterogeneous networks. *arXiv [Preprint]* (2020). doi: 10.48550/arXiv.1812.06127
33. Studholme C, Hill D, Hawkes D. An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recognit.* (1999) 32:71–86. doi: 10.1016/S0031-3203(98)00091-0
34. Hu K, Gong S, Zhang Q, Seng C, Xia M, Jiang S. An overview of implementing security and privacy in federated learning. *Artif Intell Rev.* (2024) 57:204. doi: 10.1007/s10462-024-10846-8
35. Feng Y, Guo Y, Hou Y, Wu Y, Lao M, Yu T, et al. A survey of security threats in federated learning. *Complex Intell Syst.* (2025) 11:165. doi: 10.1007/s40747-024-01664-0
36. Kumar V, Sharma M, Jehadeesan R, Venkatraman B, Sheet D. Simulating cross-modal medical images using multi-task adversarial learning of a deep convolutional neural network. *Int J Imaging Syst Technol.* (2024) 34:e23113. doi: 10.1002/ima.23113
37. Liang Z, Wei H, Liu G, Cheng M, Gao J, Li S, et al. Leveraging GAN-based CBCT-to-CT translation models for enhanced image quality and accurate photon and proton dose calculation in adaptive radiotherapy. *J Radiat Res Appl Sci.* (2024) 17:100809. doi: 10.1016/j.jrras.2023.100809
38. Hu C, Cao N, Li X, He Y, Zhou H. CBCT-to-CT synthesis using a hybrid U-Net diffusion model based on transformers and information bottleneck theory. *Sci Rep.* (2025) 15:10816. doi: 10.1038/s41598-025-92094-6

5.2.1 Supplementary Material

A Privacy-Preserving Federated Learning Framework for Generalizable CBCT to Synthetic CT Translation in Head and Neck - Supplementary Material

1 LOCAL EPOCHS INVESTIGATION

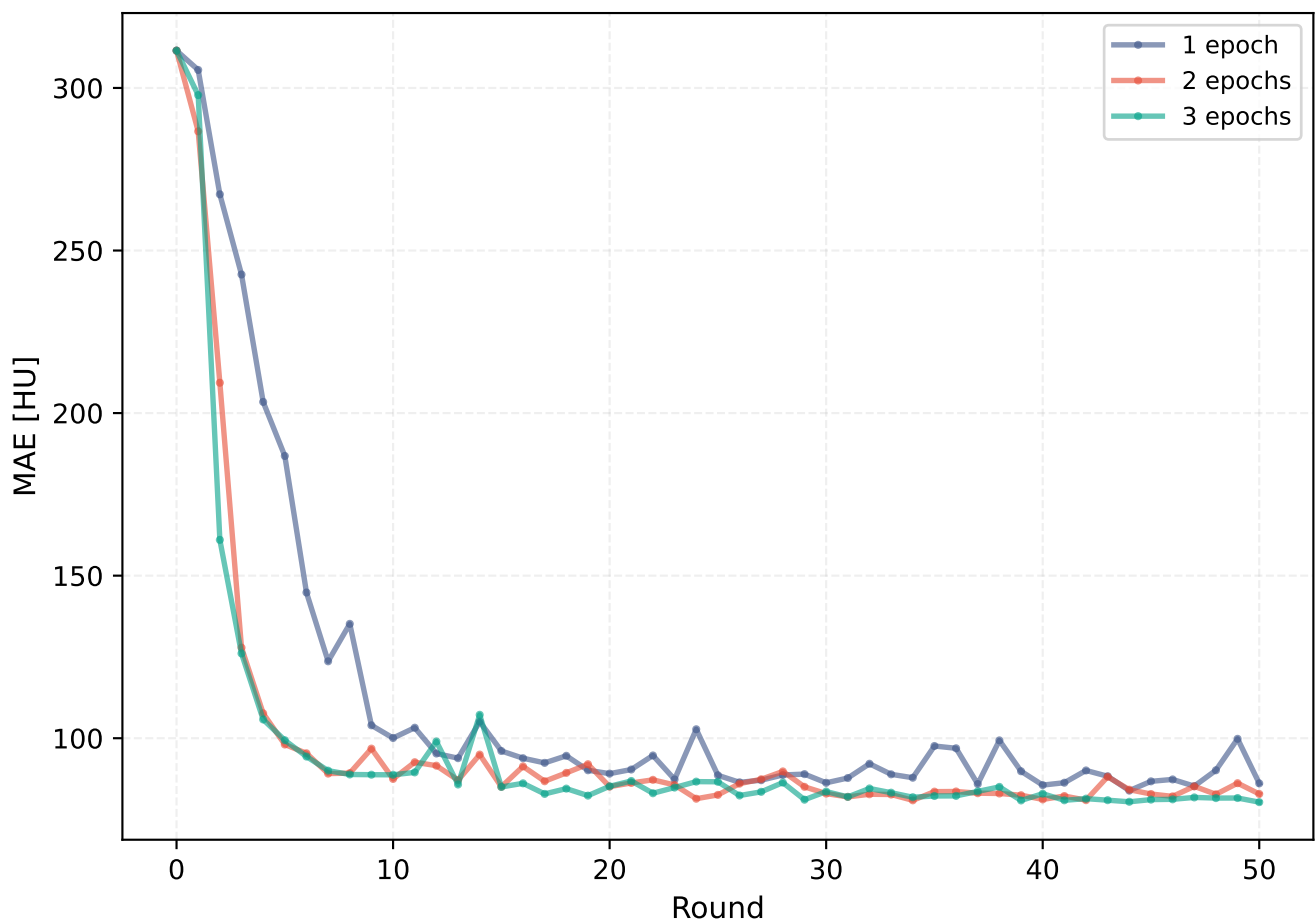


Figure S1. Convergence trend of the global model across federated rounds for different local training configurations (1, 2, and 3 local epochs per round). The MAE decreased substantially in the initial 10 rounds and stabilized at approximately round 30 for configurations with 2 and 3 local epochs. These configurations demonstrated faster and more stable convergence compared to the configuration with a single epoch.

Center	Local Epochs	MAE [HU]	SSIM	PSNR [dB]
Center A (external)	1	77.63 ± 13.29	0.90 ± 0.04	33.58 ± 2.28
	2	75.22 ± 11.81	0.90 ± 0.03	33.52 ± 2.06
	3	74.22 ± 12.34	0.90 ± 0.04	33.74 ± 2.20
Center B	1	83.07 ± 7.65	0.89 ± 0.02	33.24 ± 0.67
	2	85.90 ± 7.10	0.88 ± 0.02	32.86 ± 0.94
	3	81.43 ± 6.44	0.89 ± 0.02	33.15 ± 1.07
Center C	1	65.80 ± 16.17	0.92 ± 0.04	34.42 ± 1.69
	2	64.38 ± 13.63	0.92 ± 0.04	34.48 ± 1.39
	3	64.37 ± 13.91	0.92 ± 0.04	34.47 ± 1.46
Center E	1	73.80 ± 14.66	0.92 ± 0.02	34.89 ± 1.17
	2	74.22 ± 13.23	0.92 ± 0.01	34.91 ± 1.04
	3	76.15 ± 13.86	0.91 ± 0.02	34.69 ± 1.04

Table S1. Quantitative performance across all centers for different local epochs configurations. Results are reported as mean $\pm$ standard deviation. Center A was used exclusively for external testing, while Centers B, C, and E participated in the federated training.

PART III

EFFICIENCY, SUSTAINABILITY
AND RELIABILITY OF FEDERATED
LEARNING IN HEALTHCARE

Energy-Efficient Federated Learning via Adaptive Encoder Freezing for MRI-to-CT Translation

6.1 Introduction

Although the potential of Artificial Intelligence (AI), and in particular Deep Learning (DL), to enhance diagnostic precision and improve patient outcomes has been demonstrated, its development and clinical integration is often associated with significant computational demands [111]. The increasing complexity of contemporary DL architectures, characterised by an ever-growing number of parameters and floating-point operations, has been shown to entail considerable energy consumption and infrastructural requirements, including water resources for data centre cooling [112, 113]. Such resource-intensive approaches have been described as *red AI* [114], reflecting a paradigm primarily driven by maximising predictive performance, frequently without proportional consideration of computational cost or environmental impact. This trajectory poses a considerable risk of exacerbating inequalities in access to advanced digital health technologies, as only institutions with substantial financial resources will be able to sustain the financial and infrastructural burden associated with large-scale DL systems. This may consequently result in a widening of disparities in healthcare delivery.

In contrast, the concept of *green AI* or *green-in AI* [111] promotes the incorporation of sustainability principles directly into the design, training, and deployment of DL models. Rather than focusing exclusively on accuracy metrics, this perspective advocates for evaluating models also in terms of computational efficiency, energy consumption, and overall resource utilisation during both training and inference phases [114, 115]. In the context of increasing ethical concerns and environmental awareness, the scientific community has called for a transition towards more sustainable, transparent, and socially responsible AI

development practices. This reorientation emphasises not only environmental stewardship but also social equity, encouraging research frameworks that balance performance gains with long-term ecological and social considerations [116–118].

While Federated Learning (FL) is commonly framed as a solution to data privacy and data governance constraints, its computational footprint in medical imaging applications remains overlooked. Federated workflows require repeated distributed training cycles, synchronisation, and aggregation processes of large-scale DL models, resulting in substantial computational overhead, energy consumption, and infrastructural demands. This structural dependency on high-performance computing resources risks excluding resource-constrained institutions from meaningful participation, thereby reproducing inequalities that FL was originally intended to mitigate [119, 120].

In this context, the study introduced in this Chapter, extended the focus to the sustainability and optimisation of the FL process itself, adopting a Green AI-by-design [111, 114] perspective. A novel adaptive training methodology was introduced, based on a patience-driven encoder layer-freezing strategy that dynamically reduces the computational complexity of federated training. The method leverages the stabilisation dynamics of federated optimisation to selectively freeze network components, thereby decreasing training time, energy consumption, and carbon emissions without degrading model performance [121].

The proposed strategy was evaluated within the context of FedSynthCT-Brain [99] (presented in Chapter 4), which served as a controlled experimental environment for validating the underlying methodological principle. The results demonstrated that sustainability-oriented optimisation can be structurally integrated into FL pipelines, enabling substantial efficiency gains while preserving model accuracy and generalisation. This work therefore establishes a new methodological research direction at the intersection of FL, Green AI, and medical image synthesis, positioning sustainability, equity, and accessibility as first-class design objectives rather than secondary constraints.

The content of this chapter has been adapted from a manuscript previously published in the peer-reviewed journal “Sustainable Computing: Informatics and Systems”. Portions of the text are reproduced verbatim from the published article. The author of this thesis, as first author of the manuscript, led the conceptualisation of the proposed methodology, the design of the adaptive training strategy, the implementation of the federated pipeline, the experimental evaluation, and the analysis of the results. Furthermore, the author led the drafting and revision of the original manuscript in close collaboration with the co-authors. This work has contributed to fostering more inclusive and sustainability-aware AI research within the scientific community.

6.2 Manuscript



Energy-efficient federated learning via adaptive encoder freezing for MRI-to-CT conversion: A Green AI-guided research

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

^a Institute of Biomedical Engineering, Karlsruhe Institute of Technology, Fritz-Haber-Weg 1, Karlsruhe, 76131, Baden-Württemberg, Germany

^b Department of Political Science, Università Degli Studi Di Teramo, Via Renato Balzarini 1, Teramo, 64100, Abruzzo, Italy

^c Department of Experimental and Clinical Medicine, Magna Graecia University, Viale Europa, Catanzaro, 88100, Calabria, Italy

^d Department of Radiation Oncology, University Medical Center Schleswig-Holstein, Arnold-Heller-Strasse 3, Kiel, 24105, Schleswig-Holstein, Germany

ARTICLE INFO

Dataset link: <https://doi.org/10.5281/zenodo.7260705>

Keywords:

Federated learning
Green AI
Medical image synthesis
Synthetic computed tomography
Image-to-image translation
Equity
Fairness
Scalability

ABSTRACT

Background and Objective: Federated Learning (FL) holds the potential to advance health equity by enabling diverse institutions to collaboratively train deep learning (DL) models, even with limited data. However, the significant resource requirements of FL often exclude centres with limited computational infrastructure, further widening existing healthcare disparities. To address this issue, we propose a Green AI-oriented adaptive layer-freezing strategy designed to reduce energy consumption and computational load while maintaining model performance.

Methods: We tested our approach using different federated architectures for Magnetic Resonance Imaging (MRI)-to-Computed Tomography (CT) conversion. The proposed adaptive strategy optimises the federated training by selectively freezing the encoder weights based on the monitored relative difference of the encoder weights from round to round. A patience-based mechanism ensures that freezing only occurs when updates remain consistently minimal. The energy consumption and CO_{2eq} emissions of the federation were tracked using the CodeCarbon library.

Results: Compared to equivalent non-frozen counterparts, our approach reduced training time, total energy consumption and CO_{2eq} emissions by up to 23%. At the same time, the MRI-to-CT conversion performance was maintained, with only small variations in the Mean Absolute Error (MAE). Notably, for three out of the five evaluated architectures, no statistically significant differences were observed, while two architectures exhibited statistically significant reductions.

Conclusions: By incorporating computational sustainability into the design of FL for MRI-to-CT conversion, this work contributes to a Green AI-oriented and equity-aware approach to federated clinical AI, where reducing unnecessary computation is framed as both an environmental objective and a means to lower infrastructural barriers for resource-constrained clinical centres.

1. Introduction

Although Artificial Intelligence (AI) and particularly Deep Learning (DL) have shown great promise for supporting and improving diagnostic accuracy and patient's outcomes, their training and integration into routine care may require considerable computing power [1]. DL models, with an increasing number of parameters and operations, require large amounts of energy, in addition to water for cooling the data centres [2,3]. Such resource-intensive approaches – termed *red AI* [4] – prioritise ever-higher performance at the cost of substantial

computational and energy usage. In healthcare, this dynamic is not only technical or ecological, but also social: if AI systems can only be trained and deployed by institutions with sufficient computational capacity, digital health innovation may become less accessible to smaller or resource-constrained centres.

On the other side, *green AI* (or, more recently, *green-in AI* [1]) emphasises the integration of sustainable practices and techniques by *design*, to train and deploy DL models in clinics while reducing the environmental costs and the financial burden [4]. This paradigm advocates for assessing DL models not only on their quantitative performance, but

* Corresponding author.

E-mail address: ciro.raggio@kit.edu (C.B. Raggio).

<https://doi.org/10.1016/j.suscom.2026.101433>

Received 11 May 2026; Received in revised form 21 July 2026; Accepted 2 August 2026

Available online 8 August 2026

2210-5379/© 2026 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

also for their efficiency – both in training and testing – and therefore sustainability [5]. Consequently, in response to growing ethical and environmental concerns, the literature calls for more sustainable and socially responsible DL practices, requiring a reorientation of current research towards environmental and social sustainability as well as equity [6–8].

Complementing this, Federated Learning (FL) enables multiple institutions to collaboratively train DL models without centralising data, preserving patient’s privacy and broadening access to DL-based algorithms by potentially allowing even smaller centres to advance their diagnostic and research capabilities [9].

This paradigm has already been adopted in several medical imaging applications, including distributed lung cancer detection from CT images [10] and brain tumour segmentation [11], demonstrating the practical feasibility of collaborative model development across institutions.

Indeed, by participating in a federation, these centres can contribute with their data, thereby increasing data diversity and benefiting from a DL model that is potentially less prone to bias and designed to foster the development of more robust and generalisable models.

However, the effectiveness of FL is contingent upon the participating institutions’ computational resources and infrastructure [9]. The benefits of FL are not automatically equitable. Indeed, centres with limited computational capabilities may struggle to engage in FL initiatives, potentially leading to a concentration of AI benefits among better-equipped institutions. This could result in a scenario where only well-resourced centres can fully leverage FL [12].

Ensuring that FL does not exacerbate existing disparities or concentrate decision-making power within a few well-resourced institutions or countries, calls for a focus on equity, which entails actively reducing barriers that prevent resource-constrained institutions from fully participating [13]. In response, inspired by recent contributions in the literature [14], this work aims to lower FL’s computational demand by proposing an adaptive layers freezing method, to dynamically freeze the layers involved in training and aggregation, thereby mitigating computational overhead and making FL more accessible to resource-limited institutions.

Although prior studies have explored layer-freezing strategies to address heterogeneity in FL, most existing approaches have primarily focused on reducing communication overhead and memory usage. For instance, [15,16] proposed parameter-freezing methods to reduce communication costs, while [17] introduced SmartFreeze to improve memory efficiency. However, the relationship between these strategies and broader Green AI principles, clinical deployability, and healthcare-equity implications remains underexplored. In this work, we argue that reducing unnecessary computation during FL training can contribute not only to lower energy consumption and CO₂e emissions, but also to reducing the infrastructural barriers that may prevent resource-constrained centres from participating in multi-centre AI initiatives.

For our goal, we elected federated inter-modality medical Image-to-Image (I2I) translation as a case study. In particular, we address Magnetic Resonance Imaging (MRI)-to-Computed Tomography (CT) conversion, where synthetic CT (sCT) images are generated from MRI scans. This task has a clinical relevance, CT provides electron-density information required for applications such as radiotherapy and Positron emission tomography (PET) attenuation correction, whereas MRI provides soft-tissue contrast without ionising radiation. By enabling CT-like information to be derived from MRI, synthetic CT generation may reduce the need for additional CT acquisitions, thereby limiting registration issues and patient radiation exposure [18–20].

A recent study introduced FedSynthCT-Brain [21], a FL framework built upon encoder–decoder architectures, such as U-Net variants. The work dealt with a benchmark analysis aimed at identifying architectures that are both effective and computationally efficient for the

MRI-to-CT synthesis task. It emphasised the need to further optimise computational efficiency as a natural direction for future research.

This research builds on FedSynthCT-Brain [21] and is guided by research paradigms that advocate for DL-based clinical decision support systems (CDSS) that, *by design*, meet clinical requirements while upholding a commitment to optimisation, sustainability and equity [7]. We propose a patience-based mechanism, also inspired by early stopping strategies commonly used in DL [22], combined with the overall distance between consecutive encoder weights update. The patience mechanism is crucial in ensuring that freezing occurs exclusively when the designated network components (i.e., the encoder) achieved the stability. This facilitates the comprehension of the client consensus, as the aggregated model weights’ distance undergoes a natural reduction over successive rounds, culminating in a synchronised approach across all participants.

The innovative contribution of the work are summarised hereafter:

- We introduce a patience-based adaptive encoder-freezing strategy for federated encoder–decoder architectures applied to MRI-to-CT conversion.
- Following the guidelines in [4], we evaluate the proposed strategy in terms of MRI-to-CT synthesis performance, training time, energy consumption, and CO₂e emissions.
- We frame computational efficiency in FL as both a technical and equity-related requirement for AI in the healthcare domain, particularly when resource-constrained centres are expected to participate in multi-institutional training.

2. Related work

The application of FL to medical I2I translation is a recent development, with only a few studies specifically addressing this task [23,24]. FedSynthCT-Brain [21] tackles MRI-to-CT translation in a federated setting using a cross-silo horizontal FL approach. The authors observed that the task’s complexity and the high resolution of medical images pose a significant computational burden on individual clients, leading to extended training times.

A common initial optimisation strategy to address the aforementioned issue is to reduce communication overhead, which primarily arises from frequent synchronisation between the central server and clients during weight exchanges. A typical solution is to increase the number of local training epochs, thereby reducing synchronisation frequency. While this reduces communication costs, it may come at the expense of model accuracy, as more frequent synchronisations have been shown to improve knowledge sharing among clients and thus model performance [25]. The trade-off analysis in [25] suggests that a moderate increase in local epochs can reduce communication costs with only marginal accuracy degradation.

The implementation of alternative optimisation strategies is focused on the objective of reducing bandwidth usage, through the application of model compression techniques, such as pruning, quantisation and their combinations, such as FedWSOcomp [26–28]. However, these approaches result in a substantial increase in resource consumption, due to the additional computational load associated with the model compression and decompression on both client and server sides [16].

Several studies have also focused on developing customised aggregation functions to mitigate the adverse effects of non-IID and imbalanced data, thereby improving convergence and reducing the number of required communication rounds. One potential solution is the incorporation of a regularisation term within FedProx [29], which serves to penalise local updates that deviate from the global model. In the original study, the authors demonstrated that the employment of FedAvg in combination with FedProx achieves optimal performance, with FedProx typically requiring fewer federated rounds when compared with other strategies such as FedAvg, FedAvgM, FedBN and FedYogi [21]. However, this optimisation proved insufficient to reduce

the local computational burden or training time to a noteworthy extent, indicating that further optimisation of local training is still necessary.

In the context of federated I2I translation tasks, where encoder-decoder networks (such as U-Net, residual U-Net, or as generators in GAN architectures) are frequently employed [18,30,31], the methodology of freezing layers emerged as a particularly relevant area of interest. This approach demonstrates to be an effective training optimisation technique, both in centralised and federated settings [16,32–34]. In the context of centralised training, layer freezing demonstrates to accelerate the training process by circumventing the back-propagation procedure for the frozen layers. In federated settings, the use of freezing may reduce communication costs and the computational load on individual clients.

The most common freezing strategies involve “layer-wise” freezing, where specific layers are held constant throughout training. Over time, this reduces the computational load by skipping updates for the frozen parameters.

Both adaptive and permanent layer freezing approaches have been explored. In adaptive freezing, layers may be frozen or unfrozen at different training iterations, based either on random selection or guided by proxies, such as the magnitude of weight updates across epochs or classification accuracy on a calibration set. In permanent freezing, once a layer is frozen, it remains unchanged until training concludes [16].

Several studies have applied adaptive layer freezing to reduce local computational costs and encourage client participation [16,32–34]. FedPT [35] proposes reducing communication costs by statically selecting a subset of layers to train throughout the entire learning process, while keeping the remaining layers fixed with randomly initialised weights. However, this manual selection of trainable layers cannot guarantee optimality and may lead to substantial accuracy losses [16]. Furthermore, other works propose incremental freezing schedules based on the topological order of network layers, following predefined manual schedules or introduce an automated procedure for selecting layers to freeze by monitoring the evolution of client-model updates [16,34].

The method proposed in this paper leverages the benefits of freezing in federated training, while adapting it to different encoder-decoder architectures used in I2I translation tasks [18,30,31].

None of the state-of-the-art works have explored freezing strategies in the context of federated medical I2I translation (such as the MRI-to-CT task tackled in FedSynthCT-Brain [21]). Furthermore, our method addresses key limitations highlighted in previous work, including the study by Malan et al. [16]. Although their approach has shown promising results in cross-device federated classification using shallow CNNs, applying it to deeper models poses scalability challenges.

In contrast, our method introduces a patience-based adaptive freezing strategy that dynamically determines when to freeze encoder layers based on a quantitative measure of inter-round convergence in aggregated encoder weights. The mechanism incorporates user-controlled stability criteria – such as the patience rounds – to distinguish between transient fluctuations and stable convergence, ensuring robustness and scalability to complex encoder-decoder architectures. Our methodology is designed in line with the principles of Green AI [4], prioritising computational efficiency as a way to reach federated sustainability. By reducing client-side training loads via selective freezing, we aim to limit energy consumption and promote more responsible AI practices.

3. Methods

3.1. Federated learning setup and architectures

We designed a FL environment to simulate a realistic collaboration among four distinct clinical centres, in accordance with the setup employed in the original study [21]. Each centre trained a local DL model on its own data, while a central server coordinated the federated model by aggregating key parameters (e.g., weights, biases, batch

normalisation) from all clients after each training round. Following the findings outlined in [21], we selected a set of encoder-decoder DL architectures for the proposed FL framework:

- The original U-Net [36], referred to as Simple U-Net, featuring an encoder-decoder structure with concatenation-based long skip connections;
- The architecture proposed by Li et al. [37], designed specifically for the MRI-to-CT translation task, characterised by an encoder-decoder structure with residual connections instead of traditional skip connections, and identified as the best trade-off between effectiveness and efficiency according to [21];
- A deep encoder-decoder network inspired by U-Net, proposed by Spadea, Pileggi et al. [38], which is distinguished by its increased depth, extensive use of multi-layer convolutional blocks at each level, and regularisation via dropout;
- The architecture introduced by Fu et al. [39] for synthetic CT generation from MRI in the male pelvic region, which adopts a U-Net-like structure, but differs in its deep bottleneck, the use of instance normalisation, residual additions instead of concatenations, and the exclusive reliance on transposed convolutions for upsampling;
- A lightweight U-Net variant proposed by Li et al. [40] for the MRI-to-CT translation task, characterised by the use of Leaky ReLU activations and concatenation-based skip connections.

All the architectures predicted 256×256 sCT slices from 256×256 MRI inputs.

3.2. Proposed patience-based adaptive layer freezing

The concept of the proposed method is to selectively freeze the encoder weights of encoder-decoder architectures. This strategy aims to lessen the computational requirements within the FL framework, while preserving the performance akin to those of the unfrozen architectures (i.e., encoder-decoder based models without layer freezing).

As presented in Fig. 1, the proposed method quantified the alignment of clients’ models during federated training by tracking incremental changes in encoder weights across consecutive training rounds. The key metric of the proposed method was the *relative percentage difference* ($\rho_{\%}$) of encoder weights, computed at each round R as:

$$\rho_{\%R} = \left| \frac{f(W_R) - f(W_{R-1})}{f(W_R) + \epsilon} \right| \times 100 \quad (1)$$

Where: W_R represented the mean encoder weights at the current training round, W_{R-1} represented the mean encoder weights from the previous training round and $f(\cdot)$ computed the average Mean Absolute Error of encoder weights (MAE_{enc}) between consecutive communication rounds. The constant $\epsilon = 10^{-8}$ was added to the denominator to ensure numerical stability when $f(W_R)$ approached zero.

The MAE_{enc} is defined as:

$$MAE_{enc} = \frac{1}{n} \sum_{i=1}^n |L_R^{(i)} - L_{R-1}^{(i)}| \quad (2)$$

Where $L_R^{(i)}$ and $L_{R-1}^{(i)}$ denote the weights of the i th encoder layer for the current and previous training rounds, respectively, and n was the total number of layers. The average MAE across layers provided a measure of weight divergence.

As clients optimised the global objective rather than their local objectives round after round, the $\rho_{\%}$ decreased over successive training rounds (see Fig. 1), serving as an indicator of clients’ consensus in the federated task. A *patience mechanism* was then introduced to manage encoder freezing based on $\rho_{\%}$. A minimum threshold τ was established to distinguish values of $\rho_{\%}$ from minor fluctuations. The patience parameter $\mathcal{N}$ was defined to specify the number of consecutive rounds

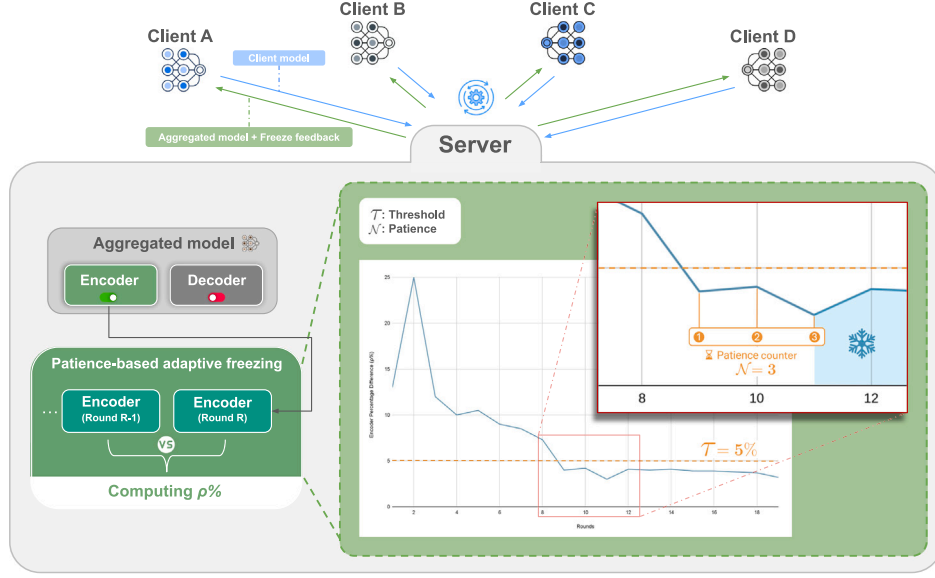


Fig. 1. Representation of the proposed patient-based adaptive layer freezing. The server is responsible for calculating the percentage difference in the aggregated encoder ($\rho_{\%}$) round by round. The encoder freezing condition is triggered when $\rho_{\%}$ remains consistently below the threshold (τ) established by the final user for a minimum of $\mathcal{N}$ (patience parameter) consecutive rounds.

during which $\rho_{\%}$ must remain below τ to trigger the mechanism. The patience mechanism is formalised as:

$$\text{Freeze Federated Encoder if: } \sum_{t=R-\mathcal{N}+1}^R \mathbb{I}(\rho_{\%}^{(t)} < \tau) = \mathcal{N} \quad (3)$$

Where, R is the current round, $\mathcal{N}$ is the patience parameter, defining the number of consecutive rounds where $\rho_{\%} < \tau$, $\mathbb{I}(\cdot)$ is the indicator function, returning 1 if the condition is true, and 0 otherwise, $\rho_{\%}^{(t)}$ is the percentage difference at round t , τ is the minimum threshold for $\rho_{\%}$ to be considered converged. The t was defined as $R - \mathcal{N} + 1$. Thus, the condition was only satisfied in the last $\mathcal{N}$ consecutive rounds. Once $\rho_{\%}$ was below the threshold τ for the defined $\mathcal{N}$ consecutive rounds, the encoder was frozen, which indicates sufficient convergence for the encoder.

Fig. 1 provides a conceptual illustration of the aforementioned patience-based freezing mechanism, with a specified threshold of $\tau = 5\%$ and a patience of $\mathcal{N} = 3$ rounds. The server monitors $\rho_{\%}$ after each aggregation step, and once the freezing condition is satisfied, a global freezing signal is generated and broadcast to all participating clients.

The server-side and client-side procedures are reported in Algorithms 1 and 2, respectively.

3.3. Datasets

As for [21], the experiments were conducted using the aforementioned FL setup with 4 clients (Centre A, B, C, and D) and 1 server (Centre E) on brain images.

Centres A, B, and C contained private datasets from US, Italy and Germany respectively, acquired in compliance with the ethical standards of the 1964 Declaration of Helsinki and its later amendments, with written informed consent obtained from all patients. Centres D and E were single institutions selected from the SynthRAD Grand Challenge 2023 [41]. The data distribution across clients was non-independent and identically distributed (non-IID), reflecting the heterogeneous nature of a realistic multi-site federation. Each centre used different MRI scanners, acquisition protocols, voxel spacings, and image resolutions, resulting in distinct local data distributions. The dimensions of the images and the number of samples in each dataset are detailed in Table 1. Further details on the scanners used for image acquisition at each centre were reported in the original study [21].

Algorithm 1 Server-Side Adaptive Freezing Strategy

Require: Threshold τ , patience $\mathcal{N}$, total rounds R

```

1: Initialize: encoder_frozen ← False,
   patience_counter ← 0,  $H \leftarrow []$ 
2: for each round  $r = 1, \dots, R$  do
3:   Server aggregates client weights  $\rightarrow W_r$ 
4:   Broadcast config = {encoder_frozen} to all clients
5:   if  $r = 1$  then
6:     Compute  $\text{MAE}_{\text{enc}}^{(1)} \leftarrow \text{MAE}_{\text{enc}}(W_1, W_0)$  { $W_0$  is the initial server
       model state; store for next round}
7:     Append  $\text{MAE}_{\text{enc}}^{(1)}$  to  $H$ 
8:   else
9:     Compute  $\text{MAE}_{\text{enc}}^{(r)} \leftarrow \text{MAE}_{\text{enc}}(W_r, W_{r-1})$ 
10:     $\rho_{\%} \leftarrow \left| \frac{\text{MAE}_{\text{enc}}^{(r)} - \text{MAE}_{\text{enc}}^{(r-1)}}{\text{MAE}_{\text{enc}}^{(r)} + \epsilon} \right| \times 100$ 
11:    if  $\rho_{\%} < \tau$  then
12:      patience_counter ← patience_counter + 1
13:    else
14:      patience_counter ← 0
15:    end if
16:    if patience_counter ≥  $\mathcal{N}$  and not encoder_frozen
       then
17:      encoder_frozen ← True
18:    end if
19:    Append  $\text{MAE}_{\text{enc}}^{(r)}$  to  $H$ 
20:  end if
21: end for

```

Each client represented a distinct hospital (or silo) with unique and non-overlapping datasets. The federated model was further validated on the unseen dataset of the server (Centre E), which was not included in any client's training and validation process. This setup ensured that the server-side evaluation process tested the generalisation capabilities of the federated model on an unseen dataset with different characteristics.

Following recent findings [21,42] a pre-preprocessing pipeline was applied to harmonise the datasets without exchanging data. Firstly, N4 bias field correction was applied with different parameters for each

Table 1
Dataset key characteristics across participating centres.

Centre	Data	MR Voxel [mm ³]	CT Voxel [mm ³]	Size (Axial × Sagittal × Coronal)
A	15	1 × 1 × 1	0.49–0.67 × 0.49–0.67 × 2.5	256 × 176 × 256
B	14	1 × 1 × 1	0.98 × 0.98 × 3.27	256 × 176 × 248
C	21	0.78 × 0.78 × 1	0.78 × 0.78 × 1	226–357 × 512 × 512
D	29	0.98–1.12 × 0.98–1.12 × 0.98–1.12	0.69–0.78 × 0.69–0.79 × 1–3	167–213 × 216–262 × 250–277
E	23	0.98 × 0.98 × 3.27	0.59–1.27 × 0.59–1.27 × 1–2	167–262 × 200–225 × 225–248

Algorithm 2 Client-Side Adaptive Training

Require: Global weights W_r , flag `encoder_frozen` from the server, local dataset D_k , local epochs E , loss $\mathcal{L}$, encoder layer keys F

- 1: $\theta \leftarrow W_r$ {load aggregated global weights}
- 2: **if** `encoder_frozen` **then**
- 3: **for** each layer $\ell \in F$ **do**
- 4: **Detach** θ_ℓ from the computational graph
- 5: **end for**
- 6: **end if**
- 7: **for** each local epoch $e = 1, \dots, E$ **do**
- 8: **for** each mini-batch $(x, y, m) \sim D_k$ **do**
- 9: $\hat{y} \leftarrow f_\theta(x)$ {forward pass}
- 10: $\mathcal{L}_{\text{batch}} \leftarrow \mathcal{L}(\hat{y}, y, m) + \frac{\mu}{2} \|\theta - W_r\|^2$ {FedProx term}
- 11: **Compute** $\nabla_\theta \mathcal{L}_{\text{batch}}$ {backprop on non-frozen layers only}
- 12: **Update** θ via Adam
- 13: **end for**
- 14: **end for**
- 15: **return** θ to server

client. All MRIs were then rescaled and normalised to the [0,1] range. To ensure uniform input dimensions for the networks, all images were resized to $256 \times 256 \times 256$ using a combination of cropping, resizing, and padding techniques. Additionally, each client applied data augmentation using spatial transformations, such as rotations, translations, and flipping.

3.4. Experimental settings and performance metrics

For our purposes, we used the experimental setup proposed in [21], which used PyTorch [43] and MONAI [44] for the DL implementation, and the Flower framework [45] for the FL implementation.

All experiments were conducted in a containerised environment equipped with 32 GB of RAM, 1 NVIDIA A100 80 GB GPU, and 16 CPU cores (AMD EPYC 7713). These computational resources were required to support the load of the entire federation (see Section 3.1) within a single physical node. Each client was nevertheless treated as an independent remote centre, and no individual client would require comparable resources in an actual distributed deployment.

The data augmentation process was implemented using the AugmentedDataLoader library¹ [21].

Energy consumption and CO₂ equivalent emissions were computed using the CodeCarbon² [1,46] library operating in offline mode. CodeCarbon was configured using process-based tracking with a sampling

interval of 30 s and Germany as the reference country for the electricity carbon-intensity factor. Independent emission trackers were used for both the federated server and each client process, and the reported CO₂ equivalent values correspond to CodeCarbon estimates.

To ensure reproducibility, each experiment was repeated five times using an identical setup for each client: a batch size of 8, a learning rate of 10^{-4} , 1 epoch of training for each local client for each round and the Adam optimiser. Following the experimental protocol and findings of the original study [21], models were trained using the Random Multi-2D approach, and a voting strategy to reach a consensus on the predictions made for each anatomical plane was applied before the test; the best performing aggregation strategy, FedAvg combined with FedProx (with a proximal term of 3), was used for each experiment. Finally, as convergence and stability were consistently achieved within 25 communication rounds and no substantial performance improvements were observed beyond this [21], a fixed number of 25 rounds was used in all experiments. Therefore, a fixed number of 25 rounds was used in all experiments.

To validate our approach, we computed a set of metrics, including time, emissions, and energy consumption, both with and without the proposed methodology, in an isolated environment.

To assess the impact of the methodology on image similarity, we employed four datasets (Centres A, B, C, and D) during the federation phase and an additional external dataset (Centre E) to evaluate the federated model's generalisation performance on previously unseen data. Image similarity was assessed using well-established metrics for the MRI-to-CT translation task [18,21,30], such as: (i) the MAE; (ii) the Peak Signal-to-Noise Ratio (PSNR); and (iii) the Structural Similarity Index Measure (SSIM), computed between the ground truth CT and the predicted CT.

3.5. Observational study and adaptive freezing evaluation

In the first set of experiments, a preliminary study was performed. Models were trained without the encoder freezing in the FL setup. The aim was to observe updates of the encoder weights and loss trends across training rounds to establish the values of τ for subsequent freezing. The two selected percentage difference thresholds were $\tau = 5\%$ and $\tau = 10\%$. Both values were tested using a fixed number of patience rounds ($\mathcal{N} = 3$).

Analysis of the encoder weight relative difference trends indicated that the 10% threshold was consistently reached after approximately 5–6 rounds, whereas the 5% threshold was typically reached after 7–10 rounds.

It was observed that, although the activation of the freezing mechanism was delayed when using the more conservative $\tau = 5\%$, model performance remained unchanged while the time needed to complete the rounds was considerably reduced.

Nevertheless, employing the higher threshold ($\tau = 10\%$) could have yielded further gains in terms of training time, energy consumption,

¹ <https://github.com/ciroraggio/AugmentedDataLoader>

² <https://codecarbon.io/>

Table 2Results (mean $\pm$ standard deviation) obtained from 5 repetitions for each model and methodology.

Model	Metric	No adaptive freeze	Adaptive freeze $\mathcal{N} = 3, \tau = 5\%$
Simple U-Net [36]	Parameters [#]	1.727×10^7	7.857×10^6
	Time [h]	13.40 ± 0.76	11.15 ± 0.64
	Emissions [kgCO ₂ eq]	2.03 ± 0.09	1.69 ± 0.09
	Total energy [kWh]	5.32 ± 0.23	4.45 ± 0.23
Li et al. [37]	Parameters [#]	2.544×10^7	1.272×10^7
	Time [h]	10.11 ± 0.38	7.89 ± 0.22
	Emissions [kgCO ₂ eq]	1.51 ± 0.04	1.16 ± 0.03
	Total energy [kWh]	3.96 ± 0.10	3.05 ± 0.07
Spadea, Pileggi et al. [38]	Parameters [#]	4.791×10^7	3.065×10^7
	Time [h]	19.22 ± 0.75	17.42 ± 0.68
	Emissions [kgCO ₂ eq]	2.97 ± 0.10	2.70 ± 0.10
	Total energy [kWh]	7.80 ± 0.26	7.08 ± 0.27
Fu et al. [39]	Parameters [#]	3.492×10^7	2.729×10^7
	Time [h]	16.49 ± 0.57	13.81 ± 0.64
	Emissions [kgCO ₂ eq]	2.17 ± 0.07	1.80 ± 0.07
	Total energy [kWh]	5.69 ± 0.17	4.74 ± 0.20
Li et al. 2020 [40]	Parameters [#]	2.682×10^7	9.024×10^6
	Time [h]	14.10 ± 0.28	11.76 ± 0.59
	Emissions [kgCO ₂ eq]	1.85 ± 0.03	1.54 ± 0.07
	Total energy [kWh]	4.85 ± 0.09	4.04 ± 0.19

and emissions. However, this configuration resulted in performance degradation across different models, due to the anticipation of encoder freezing, thereby undermining the reliability of the least conservative approach.

Following these findings, using the same setup, the models were trained for 25 rounds across 5 repetitions, with the patience parameter ($\mathcal{N}$) set to 3 rounds to ensure that the threshold ($\tau = 5\%$) was consistently exceeded.

The final set of experiments validated the proposed methodology by comparing model performance in scenarios with and without adaptive encoder freezing.

4. Results

A detailed comparison of the average emissions, energy consumption, time performance and number of trainable parameters across five independent repetitions for each evaluated model, both with and without the proposed methodology, is presented in Table 2.

A consistent reduction in the number of trainable parameters was observed following the activation of the encoder freezing controlled using the proposed methodology. Therefore, a reduction in training time per local epoch was observed on all clients (see Fig. 2 and Table 3), leading to a global decrease in the total training time across all architectures, with improvements ranging from 9.4% (Spadea, Pileggi et al. [38]) to 22.0% (Li et al. [37]).

These temporal gains translated into proportional decreases in total energy consumption and CO₂eq emissions, with reductions between 9.1% and 23.2%. The Li et al. [37] model showed the largest benefits, while the Simple U-Net [36], Fu et al. [39], and Li et al. 2020 [40] achieved consistent improvements around 16%–17%.

Nevertheless, the box-plot comparison of the MAE results obtained on the unseen Centre E, for each model, is presented in Fig. 3. Statistical significance of MAE differences between adaptive freezing and non-freezing configurations was assessed using a paired t-test after verifying normality with the Shapiro–Wilk test. The statistical analysis indicated that for three of the five architectures (Simple U-Net [36], Fu et al. [39], and Li et al. [40]) no statistically significant difference in MAE was found ($p > 0.05$). In contrast, Li et al. [37] and Spadea, Pileggi et al. [38] demonstrated statistically significant reductions ($p = 3.82 \times 10^{-5}$ and $p = 9.26 \times 10^{-4}$ respectively) in terms of MAE.

The PSNR and SSIM metrics are reported in Table 4. No systematic differences between the two training strategies were observed, with only minor stochastic fluctuations across the evaluated architectures.

Table 3Training time (in min) per local epoch (mean $\pm$ standard deviation) before and after adaptive encoder freezing for each centre and model.

Centre	Model	Training time per local epoch [min]	
		Pre-freeze	Post-freeze
Centre A	Simple U-Net [36]	21.07 ± 0.65	14.38 ± 0.43
	Li et al. [37]	15.59 ± 0.53	10.05 ± 0.31
	Spadea, Pileggi et al. [38]	31.35 ± 1.04	25.36 ± 0.80
	Fu et al. [39]	26.29 ± 2.02	18.81 ± 1.11
	Li et al. 2020 [40]	21.37 ± 0.47	15.14 ± 0.33
Centre B	Simple U-Net [36]	20.12 ± 1.60	13.52 ± 0.90
	Li et al. [37]	14.79 ± 1.07	9.46 ± 0.64
	Spadea, Pileggi et al. [38]	29.34 ± 1.68	23.71 ± 1.26
	Fu et al. [39]	25.60 ± 2.13	18.25 ± 1.56
	Li et al. 2020 [40]	19.85 ± 0.49	14.02 ± 0.34
Centre C	Simple U-Net [36]	28.38 ± 1.87	19.31 ± 1.03
	Li et al. [37]	20.15 ± 0.44	13.32 ± 0.25
	Spadea, Pileggi et al. [38]	39.77 ± 0.66	32.41 ± 0.49
	Fu et al. [39]	35.80 ± 1.75	25.78 ± 1.21
	Li et al. 2020 [40]	29.29 ± 1.67	20.89 ± 1.20
Centre D	Simple U-Net [36]	30.54 ± 1.40	20.76 ± 0.82
	Li et al. [37]	22.03 ± 0.58	14.49 ± 0.36
	Spadea, Pileggi et al. [38]	44.03 ± 0.79	35.77 ± 0.59
	Fu et al. [39]	38.50 ± 1.90	27.71 ± 1.35
	Li et al. 2020 [40]	31.72 ± 1.41	22.57 ± 1.06

5. Discussion

This study introduces an adaptive encoder layer freezing strategy within federated encoder–decoder architectures for MRI-to-CT translation, following a Green AI by-design approach.

The proposed strategy was deliberately confined to the encoder layers. As demonstrated and discussed in the original study [21], although federated models tend to generalise to unseen data, they exhibit reduced specificity at the client level. Extending the freezing strategy to the decoder would exacerbate this limitation, further restricting the capacity of the global model to accommodate local data heterogeneity. By focusing on the encoder, the federated model can instead exploit inter-client consensus to enhance stability and efficiency, while preserving decoder flexibility for client-specific adaptation. This design choice is further supported by the transfer learning techniques, where encoder-level representations are known to capture domain-transferable features, while decoder layers tend to remain more task- and data-specific [47].

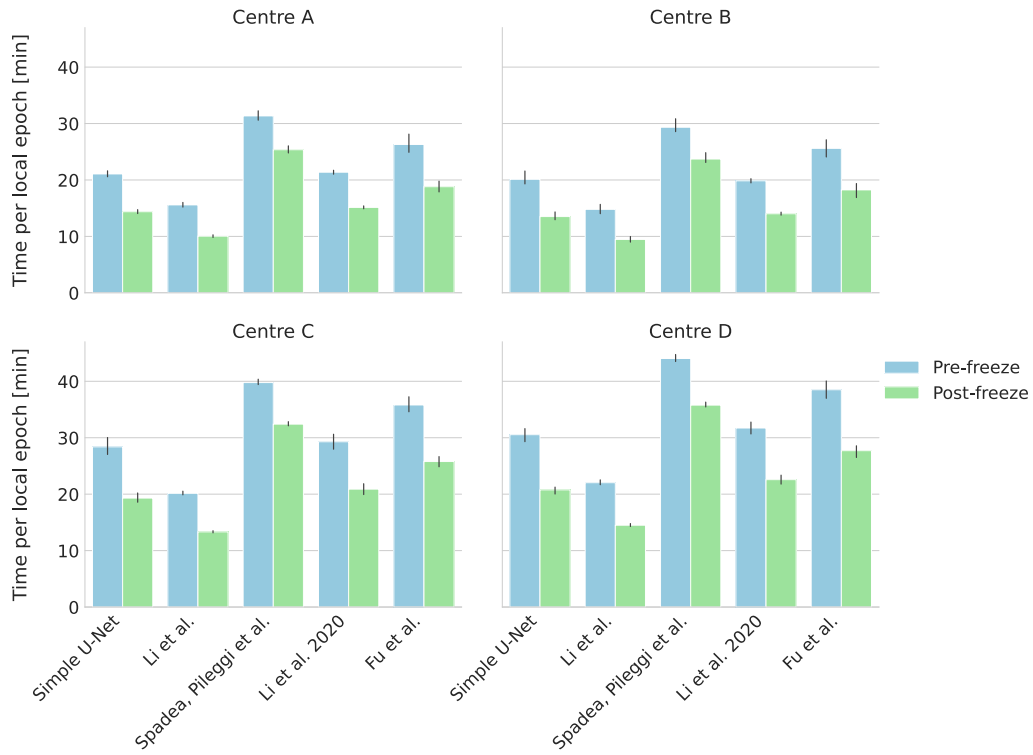


Fig. 2. Training time per local epoch for each participating client (Centres A–D), comparing the pre-freeze and post-freeze phases across the evaluated models.

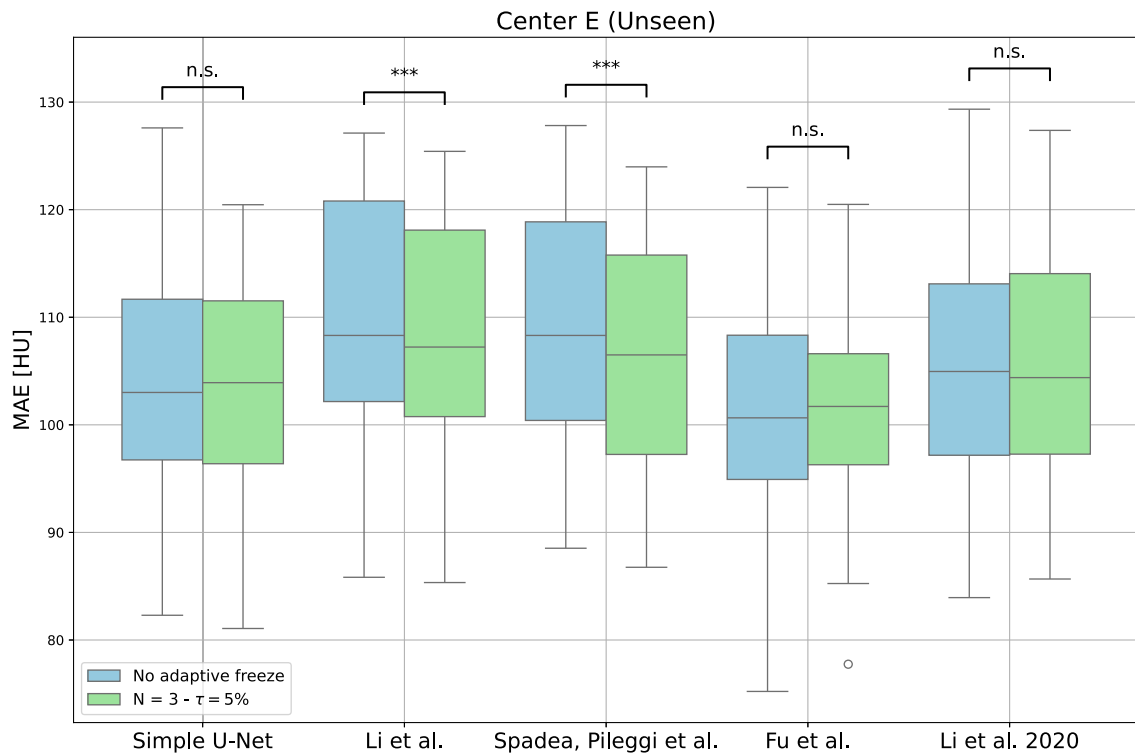


Fig. 3. Performance comparison of the models trained with and without the proposed patience-based adaptive encoder freezing. A p -value < 0.001 is indicated by ***.

Table 4

Comparison of PSNR and SSIM results, in terms of median and interquartile range for each model and methodology.

Model	Metric	No adaptive freeze	Adaptive freeze $\mathcal{N} = 3, \tau = 5\%$
Simple U-Net [36]	PSNR	26.52 (25.49–27.36)	26.48 (25.47–27.43)
	SSIM	0.89 (0.86–0.89)	0.89 (0.86–0.89)
Li et al. [37]	PSNR	26.33 (25.32–27.18)	26.42 (25.39–27.20)
	SSIM	0.89 (0.86–0.89)	0.880 (0.86–0.89)
Spadea, Pileggi et al. [38]	PSNR	26.49 (25.38–27.37)	26.43 (25.40–27.36)
	SSIM	0.89 (0.86–0.89)	0.89 (0.86–0.89)
Fu et al. [39]	PSNR	26.62 (25.53–27.45)	26.59 (25.56–27.48)
	SSIM	0.89 (0.87–0.89)	0.89 (0.86–0.89)
Li et al. 2020 [40]	PSNR	26.41 (25.44–27.30)	26.39 (25.40–27.30)
	SSIM	0.89 (0.86–0.89)	0.89 (0.86–0.89)

To demonstrate the effectiveness of our proposed strategy, we evaluated the method on multiple encoder–decoder architectures.

In consideration of the results presented in Table 2, on the Simple U-Net architecture by Ronneberger et al. [36], the adaptive freezing configuration (with parameters $\mathcal{N} = 3$ and $\tau = 5\%$) resulted in a global reduction in training time of 16.8% (from 13.40 ± 0.76 h to 11.15 ± 0.64 h). This temporal efficiency translated directly into proportional reductions in energy consumption and environmental impact, with total energy decreasing from 5.32 ± 0.23 kWh to 4.45 ± 0.23 kWh (16.4% reduction) and kgCO₂eq emissions decreasing from 2.03 ± 0.09 kg to 1.69 ± 0.09 kg (16.7% reduction).

Improvements were observed with the Li et al. [37] architecture, where the adaptive freezing mechanism achieved a remarkable 22.0% reduction in training time (from 10.11 ± 0.38 h to 7.89 ± 0.22 h). Correspondingly, this architecture exhibited the most substantial environmental benefits, with kgCO₂eq emissions reduced by 23.2% (from 1.51 ± 0.04 kg to 1.16 ± 0.03 kg) and energy consumption decreased by 23.0% (from 3.96 ± 0.10 kWh to 3.05 ± 0.07 kWh).

The positive trend was confirmed for the Fu et al. [39] architecture, which exhibited substantial efficiency improvements, with training time reduced by 16.2% (from 16.49 ± 0.57 h to 13.81 ± 0.64 h). This architecture also demonstrated notable environmental benefits, with kgCO₂eq emissions decreasing by 17.1% (from 2.17 ± 0.07 kg to 1.80 ± 0.07 kg) and energy consumption reduced by 16.7% (from 5.69 ± 0.17 kWh to 4.74 ± 0.20 kWh).

Similarly, the Li et al. 2020 [40] architecture showed consistent improvements with a 16.6% reduction in training time (from 14.10 ± 0.28 h to 11.76 ± 0.59 h), as well as corresponding decreases in emissions (16.8%, from 1.85 ± 0.03 kg to 1.54 ± 0.07 kg) and energy consumption (16.7%, from 4.85 ± 0.09 kWh to 4.04 ± 0.19 kWh).

Notably, the Spadea, Pileggi et al. [38] architecture, despite requiring the longest absolute training time, demonstrated consistent improvements with a 9.4% reduction in training duration (from 19.22 ± 0.75 h to 17.42 ± 0.68 h) across 25 rounds, with a corresponding reductions in emissions and energy consumption of $\approx 9\%$.

Furthermore, Fig. 2 and Table 3 report and compare the training time per local epoch before and after the application of the proposed methods across the four clients.

As a direct consequence of the reduction in the number of the trainable layers due to the encoder freeze, in Centre A, the Simple U-Net average time decreased from 21.07 ± 0.65 min to 14.38 ± 0.43 min per local epoch, corresponding to a 31.8% reduction, while Li et al. [37] reduced from 15.59 ± 0.53 min to 10.05 ± 0.31 min (35.5% reduction). Similar reduction trends were observed in Centres B, C, and D, with the largest absolute reductions observed in Spadea, Pileggi et al. on Centre D (from 44.03 ± 0.79 min to 35.77 ± 0.59 min) and Fu et al. on Centre D (from 38.50 ± 1.90 min to 27.71 ± 1.35 min).

These per-epoch reductions scaled across the full federated training process, leading to proportional decreases in total energy consumption and CO₂ emissions observed in Table 2, highlighting that adaptive

freezing both accelerates training and improves the environmental efficiency of the FL workflow, without adversely affecting model performance across heterogeneous client datasets. From an equity perspective, these reductions provide a concrete mechanism through which the proposed strategy has the potential to lower practical barriers to participation in FL initiatives. In real-world FL deployments, each centre is required to perform local training via its own computational infrastructure. Therefore, reducing the number of trainable layers (after encoder freezing) directly decreases the amount of local backpropagation and optimiser updates, with effects on GPU time, energy consumption, and associated operational costs. This is relevant for smaller hospitals or resource-constrained clinical centres, where access to high-end GPUs, dedicated computing clusters, or sustained electricity budgets may be limited. By lowering the computational demand at the client side, adaptive encoder freezing can make participation in multi-centre FL more feasible for institutions that might otherwise be excluded from collaborative AI development.

The statistical analysis of the models' performance – in terms of MAE – on the unseen Centre E dataset (Fig. 3) revealed that the proposed adaptive freezing mechanism preserved generalisation performance and occasionally marginally enhanced it.

For three of the five evaluated architectures (Simple U-Net [36], Fu et al. [39], and Li et al. [40]), no statistically significant difference in MAE was found ($p > 0.05$, t-test) on 23 unseen patients, indicating that computational efficiency gains were achieved without compromising model performance. The conduction of statistical testing across single clients was precluded by the limited number of test cases available per site.

Notably, the results for the Li et al. [37] and Spadea, Pileggi et al. [38] architectures, demonstrated statistically significant reductions in terms of MAE ($p < 0.001$, t-test). However, the absolute MAE reduction (≈ 2 HU) remains well below radiologically and clinically meaningful thresholds for image contrast and radiotherapy, where deviations greater than 20–50 HU are typically required to discriminate biological tissues and affect dose distributions [48,49]. Hence, these results indicated that the proposed adaptive freezing mechanism preserved model fidelity, with potential minor regularisation effects, but without producing relevant performance changes.

Indeed, no performance improvement or deterioration was observed in terms of PSNR and SSIM – metrics that are less sensitive than MAE – with only minor stochastic fluctuations resulting from the independent repetitions of the experiments. Comparable or identical findings across the different architectures and methodologies was also reported in the FedSynthCT-Brain study [21], particularly with regard to SSIM, further emphasising that MAE remains the pivotal metric for performance evaluation in this type of task.

To further investigate the impact of the adaptive freezing mechanism on the generated sCTs, additional qualitative comparisons are provided in the supplementary material (see Section S.1, Figure S1–S10). Specifically, for each evaluated model, two representative cases are presented: (i) the case with the smallest difference and (ii) the case

with the largest difference in MAE between the sCTs generated by the models trained with and without adaptive freezing, across the test set.

The visual comparisons confirmed that no systematic degradations are associated with the proposed methodology, and the observed differences were attributable to the inherent stochastic nature of the training process and the independent experiments.

The standard deviations reported for each metric across five repetitions indicate stable and reproducible results, with the adaptive freezing methodology consistently reducing variability in training time and resource consumption. The emissions reductions ranged from 9.1% to 23.2%, with an average reduction of $\approx 19.9\%$ across all models. These improvements are particularly important given the increasing emphasis on sustainable machine learning practices and the growing concern regarding the carbon footprint of DL model training [50]. Furthermore, this efficiency gain is especially valuable in clinical settings where computational resources may be limited and cost considerations are essential. The variability in efficiency gains stems from the fact that the computational benefit of freezing a network component depends not only on the number of trainable parameters that are frozen, but also on how the computational workload (FLOPs) is distributed throughout the architecture [51,52]. Consequently, reducing the number of trainable parameters does not necessarily result in proportional reductions in training time, energy consumption, or CO₂ emissions.

This behaviour is reflected in Table 2. For instance, Li et al. (2020) [40] exhibited the largest reduction in trainable parameters, yet achieved a training-time reduction of 16.6%. In contrast, Li et al. (2019) [37] achieved the most substantial efficiency gains despite a smaller parameter reduction. Similarly, the architecture proposed by Spadea, Pileggi et al. [38] demonstrated the lowest efficiency gains, which may be attributed to its deep multi-convolutional design, where a substantial portion of the computational workload remains active after encoder freezing. These findings demonstrated the efficacy of the patience-based adaptive freezing approach in a federated environment, with consistent improvements observed across various architectures, ranging from fundamental designs to more complex variants. However, the magnitude of the gains is still dependent upon the distribution of the computational workload within each architecture.

Notably, the parameters $\mathcal{N} = 3$ and $\tau = 5\%$ were selected empirically through the observational study presented in Section 3.5. This configuration achieved a favourable trade-off between computational efficiency and image-translation performance across the five evaluated federated encoder-decoder architectures. As is common for training-dynamics hyperparameters in DL, these values were determined through empirical validation rather than analytical derivation. While the proposed configuration proved robust in the considered experimental setting, further validation on additional architectures and image-to-image translation tasks will help assess its broader applicability.

6. Conclusion

In this study we proposed and empirically validated a patience-based adaptive encoder-freezing strategy to lay the foundation for the operationalisation of Green AI principles in FL for medical image translation.

The method was evaluated in the context of federated MRI-to-CT translation using multiple encoder-decoder architectures, with the aim of making federated strategies more accessible to a wider range of users in the medical field. Nevertheless, some limitations should be acknowledged. The evaluation was conducted on a limited set of encoder-decoder architectures and within the FedSynthCT-Brain MRI-to-CT translation setting. Further validation is therefore required to assess the generalisability of the proposed strategy to other architectures, imaging modalities, anatomical districts, and medical-imaging tasks.

The proposed method should also be investigated under more challenging non-IID conditions, as stronger differences across clients may affect encoder-weight stability and the optimal freezing time. While reducing local computational burden may lower participation barriers for resource-constrained centres, premature or globally synchronised encoder freezing could, in principle, reduce adaptability for under-represented institutions whose local data distributions are not sufficiently reflected in the aggregated model before freezing [53,54]. In the present work, this risk was mitigated by freezing only the encoder while preserving decoder flexibility for client-specific adaptation, and by verifying that performance was preserved on the unseen Centre E dataset. Future work should therefore assess site-stratified performance, rare-pattern robustness, and fairness across clients in more heterogeneous FL scenarios.

Additional limitations concern the manual tuning required for the τ and patience parameters and the controlled hardware setting used in the experiments. The selection of these parameters may require preliminary investigation of architecture-specific weight dynamics, thus, automated or systematic strategies for parameter selection should be explored to improve the practicality and generalisability of the method. Moreover, although the homogeneous hardware setting enabled an unbiased assessment of the adaptive freezing mechanism, future studies should investigate heterogeneous client hardware configurations and quantify additional client-side metrics, including communication overhead.

CRedit authorship contribution statement

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

Ethics statement

Written informed consent was obtained from all participants. The study complied with the ethical standards of the 1964 Declaration of Helsinki and its later amendments for research involving human subjects. All patient data were fully anonymised prior to analysis to ensure privacy and confidentiality. No personally identifiable information was used in this study.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

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

Appendix A. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.suscom.2026.101433>.

Data availability

The data used for the study acquired from Centre A, B and C are private. The data from Centre D and Centre E were extracted from the public SynthRAD2023 Grand Challenge dataset and are available at <https://doi.org/10.5281/zenodo.7260705>.

References

- [1] V. Bolón-Canedo, L. Morán-Fernández, B. Cancela, A. Alonso-Betanzos, A review of green artificial intelligence: Towards a more sustainable future, *Neurocomputing* (2024) 128096, <http://dx.doi.org/10.1016/j.neucom.2024.128096>.
- [2] S. Luccioni, Y. Jernite, E. Strubell, Power hungry processing: Watts driving the cost of ai deployment? in: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24*, Association for Computing Machinery, New York, NY, USA, 2024, pp. 85–99, <http://dx.doi.org/10.1145/3630106.3658542>.
- [3] G. Varoquaux, S. Luccioni, M. Whittaker, Hype, sustainability, and the price of the bigger-is-better paradigm in ai, in: *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, FAccT '25*, Association for Computing Machinery, New York, NY, USA, 2025, pp. 61–75, <http://dx.doi.org/10.1145/3715275.3732006>.
- [4] R. Schwartz, J. Dodge, N.A. Smith, O. Etzioni, Green AI, *Commun. ACM* 63 (2020) 54–63, <http://dx.doi.org/10.1145/3381831>.
- [5] H. Liu, Y. Wang, W. Fan, X. Liu, Y. Li, S. Jain, Y. Liu, A. Jain, J. Tang, Trustworthy ai: A computational perspective, *ACM Trans. Intell. Syst. Technol.* 14 (2022) 1–59, <http://dx.doi.org/10.1145/3546872>.
- [6] A. Jobin, M. Ienca, E. Vayena, The global landscape of ai ethics guidelines, *Nat. Mach. Intell.* 1 (2019) 389–399, <http://dx.doi.org/10.1038/s42256-019-0088-2>.
- [7] R. Raman, D. Pattnaik, H.H. Lathabai, C. Kumar, K. Govindan, P. Nedungadi, Green and sustainable ai research: an integrated thematic and topic modeling analysis, *J. Big Data* 11 (2024) 55, <http://dx.doi.org/10.1186/s40537-024-00920-x>.
- [8] H. Siala, Y. Wang, Shifting artificial intelligence to be responsible in healthcare: A systematic review, *Soc. Sci. Med.* 296 (2022) 114782, <http://dx.doi.org/10.1016/j.socscimed.2022.114782>, URL: <https://www.sciencedirect.com/science/article/pii/S0277953622000855>.
- [9] C. Chen, T. Liao, X. Deng, Z. Wu, S. Huang, Z. Zheng, Advances in robust federated learning: A survey with heterogeneity considerations, *IEEE Trans. Big Data* 11 (2025) 1548–1567, <http://dx.doi.org/10.1109/TBDATA.2025.3527202>, URL: <https://doi.ieeecomputersociety.org/10.1109/TBDATA.2025.3527202>.
- [10] A. Heidari, D. Javaheri, S. Tournaj, N.J. Navimipour, M. Rezaei, M. Unal, A new lung cancer detection method based on the chest CT images using federated learning and blockchain systems, *Artif. Intell. Med.* 141 (2023) 102572, <http://dx.doi.org/10.1016/j.artmed.2023.102572>, URL: <https://www.sciencedirect.com/science/article/pii/S0933365723000866>.
- [11] M. Zenk, U. Baid, S. Pati, A. Linardos, B. Edwards, M. Sheller, P. Foley, A. Aristizabal, D. Zimmerer, A. Gruzdev, J. Martin, R.T. Shinohara, A. Reinke, F. Isensee, S. Parampottupadam, K. Parekh, R. Floca, H. Kassem, B. Baheti, S. Thakur, V. Chung, K. Kushibar, K. Lekadir, M. Jiang, Y. Yin, H. Yang, Q. Liu, C. Chen, Q. Dou, P.-A. Heng, X. Zhang, S. Zhang, M.I. Khan, M.A. Azeem, M. Jafaritadi, E. Alhoniemi, E. Kontio, S.A. Khan, L. Mächler, I. Ezhov, F. Kofler, S. Shit, J.C. Paetzold, T. Loehr, H. Wiestler, H. Peiris, K. Pawar, S. Zhong, Z. Chen, M. Hayat, G. Egan, M. Harandi, E. Isik Polat, G. Polat, A. Kocyigit, A. Temizel, A. Tuladhar, L. Tyagi, R. Souza, N.D. Forkert, P. Mouches, M. Wilms, V. Shambhat, A. Maurya, S.S. Danannavar, R. Kalla, V.K. Anand, G. Krishnamurthi, S. Nalawade, C. Ganesh, B. Wagner, D. Reddy, Y. Das, F.F. Yu, B. Fei, A.J. Madhuranthakam, J. Maldjian, G. Singh, J. Ren, W. Zhang, N. An, Q. Hu, Y. Zhang, Y. Zhou, V. Siomos, G. Tarroni, J. Passerat-Palmbach, A. Rawat, G. Zizzo, S.R. Kadhe, J.P. Epperlein, S. Braghin, Y. Wang, R. Kanagavelu, Q. Wei, Y. Yang, Y. Liu, K. Kotowski, S. Adamski, B. Machura, W. Malara, L. Zarudzki, J. Nalepa, Y. Shi, H. Gao, S. Avestimehr, Y. Yan, A.S. Akbar, E. Kondrateva, H. Yang, Z. Li, H.-Y. Wu, J. Roth, C. Saueressig, A. Milesi, Q.D. Nguyen, N.J. Gruenhausen, T.-M. Huang, J. Ma, H.S.H. Singh, N.-Y. Pan, D. Zhang, R.A. Zeineldin, M. Futrega, Y. Yuan, G.M. Conte, X. Feng, Q.D. Pham, Y. Xia, Z. Jiang, H.M. Luu, M. Dobko, A. Carré, B. Tuchinov, H. Mohy ud Din, S. Alam, A. Singh, N. Shah, W. Wang, C. Sako, M. Bilello, S. Ghodasara, S. Mohan, C. Davatzikos, E. Calabrese, J. Rudie, J. Villanueva-Meyer, S. Cha, C. Hess, J. Mongan, M. Ingallhalikar, M. Jadhav, U. Pandey, J. Saini, R.Y. Huang, K. Chang, M.-S. To, S. Bhardwaj, C. Chong, M. Agzarian, M. Kozubek, F. Lux, J. Michálek, P. Matula, M. Kerkovský, T. Koprivová, M. Dostál, V. Vybíhal, M.C. Pinho, J. Holcomb, M. Metz, R. Jain, M.D. Lee, Y.W. Lui, P. Tiwari, R. Verma, R. Bareja, I. Yadav, J. Chen, N. Kumar, Y. Gusev, K. Bhuvaneswar, A. Sayah, C. Bencheqroun, A. Belouali, S. Madhavan, R.R. Colen, A. Kotrotsou, P. Vollmuth, G. Brugnara, C.J. Preetha, F. Sahm, M. Bendszus, W. Wick, A. Mahajan, C. Balaña, J. Capellades, J. Puig, Y.S. Choi, S.-K. Lee, J.H. Chang, S.S. Ahn, H.F. Shaykh, A. Herrera-Trujillo, M. Trujillo, W. Escobar, A. Abello, J. Bernal, J. Gómez, P. LaMontagne, D.S. Marcus, M. Milchenko, A. Nazeri, B. Landman, K. Ramadass, K. Xu, S. Chotai, L.B. Chambless, A. Mistry, R.C. Thompson, A. Srinivasan, J.R. Bapuraj, A. Rao, N. Wang, O. Yoshiaki, T. Moritani, S. Turk, J. Lee, S. Prabhudesai, J. Garrett, M. Larson, R. Jeraj, H. Li, T. Weiss, M. Weller, A. Bink, B. Pouymayou, S. Sharma, T.-C. Tseng, S. Adabi, A. Xavier Falcão, S.B. Martins, B.C.A. Teixeira, F. Sprenger, D. Menotti, D.R. Lucio, S.P. Niclou, O. Keunen, A.-C. Hau, E. Pelaez, H. Franco-Maldonado, F. Loayza, S. Quevedo, R. McKinley, J. Slotboom, P. Radojewski, R. Meier, R. Wiest, J. Trenkler, J. Pichler, G. Necker, A. Haunschild, S. Meckel, P. Guevara, E. Torche, C. Mendoza, F. Vera, E. Ríos, E. López, S.A. Velastin, J. Choi, S. Baek, Y. Kim, H. Ismael, B. Allen, J.M. Buatti, P. Zampakis, V. Panagiotopoulos, P. Tsiganos, S. Alexiou, I. Haliassos, E.I. Zacharakis, K. Moustakas, C. Kalogeropoulou, D.M. Kardamakis, B. Luo, L.M. Poisson, N. Wen, M. Vallières, M.A.L. Loutfi, D. Fortin, M. Lepage, F. Morón, J. Mandel, G. Shukla, S. Liem, G.S. Alexandre, J. Lombardo, J.D. Palmer, A.E. Flanders, A.P. Dicker, G. Ogbale, D. Oyekunle, O. Odafe-Oyibo, B. Osobu, M. Shu'aibu Hikima, M. Soneye, F. Dako, A. Dorcas, D. Murcia, E. Fu, R. Haas, J.A. Thompson, D.R. Ormond, S. Currie, K. Fatania, R. Frood, A.L. Simpson, J.J. Peoples, R. Hu, D. Cutler, F.Y. Moraes, A. Tran, M. Hamghalam, M.A. Boss, J. Gimpel, D. Kattil Veettil, K. Schmidt, L. Cimino, C. Price, B. Bialecki, S. Marella, C. Apgar, A. Jakab, M.-A. Weber, E. Colak, J. Kleesiek, J.B. Freymann, J.S. Kirby, L. Maier-Hein, J. Albrecht, P. Mattson, A. Karargiris, P. Shah, B. Menze, K. Maier-Hein, S. Bakas, Towards fair decentralized benchmarking of healthcare AI algorithms with the federated tumor segmentation (FeTS) challenge, *Nat. Commun.* 16 (2025) <http://dx.doi.org/10.1038/s41467-025-60466-1>.
- [12] A. Rauniyar, D.H. Hagos, D. Jha, J.E. Håkegård, U. Bagci, D.B. Rawat, V. Vlassov, Federated learning for medical applications: A taxonomy, current trends, challenges, and future research directions, *IEEE Internet Things J.* (2023) <http://dx.doi.org/10.1109/JIOT.2023.3329061>.
- [13] D. Narayanan, M. Nagpal, J. McGuire, S. Schweitzer, D.D. Cremer, Fairness perceptions of artificial intelligence: A review and path forward, *Int. J. Human-Computer Interact.* 40 (2024) 4–23, <http://dx.doi.org/10.1080/10447318.2023.2210890>, arXiv:<https://doi.org/10.1080/10447318.2023.2210890>.
- [14] D. Thakur, A. Guzzo, G. Fortino, F. Piccialli, Green federated learning: A new era of green aware ai, *ACM Comput. Surv.* 57 (2025) <http://dx.doi.org/10.1145/3718363>, URL: <https://doi.org/10.1145/3718363>.
- [15] C. Chen, H. Xu, W. Wang, B. Li, B. Li, L. Chen, G. Zhang, Communication-efficient federated learning with adaptive parameter freezing, in: *2021 IEEE 41st International Conference on Distributed Computing Systems, ICDCS, IEEE, 2021*, pp. 1–11, <http://dx.doi.org/10.1109/ICDCS51616.2021.00010>.
- [16] E. Malan, V. Peluso, A. Calimera, E. Macii, P. Montuschi, Automatic layer freezing for communication efficiency in cross-device federated learning, *IEEE Internet Things J.* 11 (2024) 6072–6083, <http://dx.doi.org/10.1109/JIOT.2023.3309691>.
- [17] Y. Wu, L. Li, C. Tian, T. Chang, C. Lin, C. Wang, C.-Z. Xu, Heterogeneity-aware memory efficient federated learning via progressive layer freezing, in: *2024 IEEE/ACM 32nd International Symposium on Quality of Service (IWQoS), IEEE, 2024*, pp. 1–10, <http://dx.doi.org/10.1109/IWQoS61813.2024.10682916>.
- [18] M.F. Spadea, M. Maspero, P. Zaffino, J. Seco, Deep learning based synthetic-ct generation in radiotherapy and pet: A review, *Med. Phys.* 48 (2021) 6537–6566, <http://dx.doi.org/10.1002/mp.15150>, URL: <https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.15150>, arXiv:<https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.15150>.
- [19] A. Thummerer, P. Zaffino, M. Spadea, A. Knopf, M. Maspero, Artificial intelligence to generate synthetic ct for adaptive particle therapy, in: *Imaging in Particle Therapy: Current Practice and Future Trends*, IOP Publishing Bristol, UK, 2024, pp. 8–1–8–16, <http://dx.doi.org/10.1088/978-0-7503-5117-1ch8>.
- [20] V. Rogowski, M.L. Terpstra, N. Wahl, F. Kamp, E. van der Bijl, A.J. Galapon, C. Kurz, B. Xin, Z. Sun, H. Min, G. Belous, J. Dowling, Y. Xia, S. Mei, F. Fan, A. Longuefosse, J.S. Gonzalez, M.D. Benito, A.G. Martin, F. Baldacci, V. Bousso, C. Hémon, J.-C. Nunes, J.-L. Dillenseger, Z. Zhang, J. Cai, H. Bing, T. Zuopeng, R. Brioso, D. Loiacono, G. Landry, A. Thummerer, M. Maspero, Generating synthetic computed tomography for radiotherapy: SynthRAD2025 challenge report, 2026, URL: <https://arxiv.org/abs/2605.13555>, arXiv:[2605.13555](https://arxiv.org/abs/2605.13555).
- [21] C.B. Raggio, M.K. Zabaleta, N. Skupien, O. Blanck, F. Cicone, G.L. Cascini, P. Zaffino, L. Migliorelli, M.F. Spadea, FedSynthCT-Brain: A federated learning framework for multi-institutional brain MRI-to-CT synthesis, *Comput. Biol. Med.* 192 (2025) 110160, <http://dx.doi.org/10.1016/j.combiomed.2025.110160>, URL: <https://www.sciencedirect.com/science/article/pii/S0010482525005116>.
- [22] L. Prechelt, Early Stopping — But When? Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 53–67, http://dx.doi.org/10.1007/978-3-642-35289-8_5.
- [23] J. Wang, G. Xie, Y. Huang, J. Lyu, F. Zheng, Y. Zheng, Y. Jin, Fedmedgan: Federated domain translation on unsupervised cross-modality brain image synthesis, *Neurocomputing* 546 (2023) 126282, <http://dx.doi.org/10.1016/j.neucom.2023.126282>, URL: <https://www.sciencedirect.com/science/article/pii/S09252321223004058>.
- [24] C.B. Raggio, P. Zaffino, M.F. Spadea, A privacy-preserving federated learning framework for generalizable cbct to synthetic ct translation in head and neck, *Front. Digit. Health Volume 8 - 2026* (2026) <http://dx.doi.org/10.3389/fdgh.2026.1812254>, URL: <https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgh.2026.1812254>.

- [25] B. McMahan, E. Moore, D. Ramage, S. Hampson, B.A. y. Arcas, Communication-efficient learning of deep networks from decentralized data, in: A. Singh, J. Zhu (Eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, in: *Proceedings of Machine Learning Research*, Vol. 54, PMLR, 2017, pp. 1273–1282, URL: <https://proceedings.mlr.press/v54/mcmahan17a.html>.
- [26] J. Mills, J. Hu, G. Min, Communication-efficient federated learning for wireless edge intelligence in iot, *IEEE Internet Things J.* 7 (2020) 5986–5994, <http://dx.doi.org/10.1109/JIOT.2019.2956615>.
- [27] M.-D. Nguyen, S.-M. Lee, Q.-V. Pham, D.T. Hoang, D.N. Nguyen, W.-J. Hwang, Hcfl: A high compression approach for communication-efficient federated learning in very large scale iot networks, *IEEE Trans. Mob. Comput.* 22 (2023) 6495–6507, <http://dx.doi.org/10.1109/TMC.2022.3190510>.
- [28] A.N. Onaizah, Y. Xia, A.J. Obaid, K. Hussain, Deep learning based brain tumour architecture for weight sharing optimization in federated learning, *Expert Syst.* 42 (2025) e13643, <http://dx.doi.org/10.1111/exsy.13643>, URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/exsy.13643>, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/exsy.13643.
- [29] T. Li, A.K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, V. Smith, Federated optimization in heterogeneous networks, in: I. Dhillon, D. Papailiopoulos, V. Sze (Eds.), *Proceedings of Machine Learning and Systems*, volume 2, 2020, pp. 429–450, URL: https://proceedings.mlsys.org/paper_files/paper/2020/file/1f5fe83998a09396be6477d9475ba0c-Paper.pdf.
- [30] S. Dayarathna, K.T. Islam, S. Uribe, G. Yang, M. Hayat, Z. Chen, Deep learning based synthesis of mri, ct and pet: Review and analysis, *Med. Image Anal.* 92 (2024) 103046, <http://dx.doi.org/10.1016/j.media.2023.103046>, URL: <https://www.sciencedirect.com/science/article/pii/S1361841523003067>.
- [31] E.M. Huijben, M.L. Terpstra, A.J. Galapon, S. Pai, A. Thummerer, P. Koopmans, M. Afonso, M. van Eijnatten, O. Gurney-Champion, Z. Chen, Y. Zhang, K. Zheng, C. Li, H. Pang, C. Ye, R. Wang, T. Song, F. Fan, J. Qiu, Y. Huang, J. Ha, J. Sung Park, A. Alain-Beaudoin, S. Bériault, P. Yu, H. Guo, Z. Huang, G. Li, X. Zhang, Y. Fan, H. Liu, B. Xin, A. Nicolson, L. Zhong, Z. Deng, G. Müller-Franzes, F. Khader, X. Li, Y. Zhang, C. Hémon, V. Boussoit, Z. Zhang, L. Wang, L. Bai, S. Wang, D. Mus, B. Kooiman, C.A. Sargeant, E.G. Henderson, S. Kondo, S. Kasai, R. Karimzadeh, B. Ibragimov, T. Helfer, J. Dafflon, Z. Chen, E. Wang, Z. Perko, M. Maspero, Generating synthetic computed tomography for radiotherapy: Synthrad2023 challenge report, *Med. Image Anal.* 97 (2024) 103276, <http://dx.doi.org/10.1016/j.media.2024.103276>, URL: <https://www.sciencedirect.com/science/article/pii/S1361841524002019>.
- [32] G. Goutam, S. Balasubramanian, D. Gera, R.R. Sarma, LayerOut: Freezing layers in deep neural networks, *SN Comput. Sci.* 1 (2020).
- [33] K. Pfeiffer, M. Rapp, R. Khalili, J. Henkel, Cocofl: Communication- and computation-aware federated learning via partial nn freezing and quantization, 2023, URL: <https://arxiv.org/abs/2203.05468>, arXiv:2203.05468.
- [34] E. Malan, V. Peluso, A. Calimera, E. Macii, Communication-efficient federated learning with gradual layer freezing, *IEEE Embed. Syst. Lett.* 15 (2023) 25–28, <http://dx.doi.org/10.1109/LES.2022.3190682>.
- [35] H. Sidahmed, Z. Xu, A. Garg, Y. Cao, M. Chen, Efficient and private federated learning with partially trainable networks, 2021, URL: <https://arxiv.org/abs/2110.03450>, arXiv:2110.03450.
- [36] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W.M. Wells, A.F. Frangi (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Springer International Publishing, Cham, 2015, pp. 234–241.
- [37] Y. Li, W. Li, P. He, J. Xiong, J. Xia, Y. Xie, CT synthesis from MRI images based on deep learning methods for MRI-only radiotherapy, in: 2019 International Conference on Medical Imaging Physics and Engineering, ICMIPe, 2019, pp. 1–6, <http://dx.doi.org/10.1109/ICMIPe47306.2019.9098190>.
- [38] M.F. Spadea, G. Pileggi, P. Zaffino, P. Salome, C. Catana, D. Izquierdo-Garcia, F. Amato, J. Seco, Deep convolution neural network (dcnn) multiplane approach to synthetic ct generation from mr images—application in brain proton therapy, *Int. J. Radiat. Oncology*Biophysics*Physics* 105 (2019) 495–503, <http://dx.doi.org/10.1016/j.ijrobp.2019.06.2535>, URL: <https://www.sciencedirect.com/science/article/pii/S0360301619334236>.
- [39] J. Fu, Y. Yang, K. Singhrao, D. Ruan, F.-I. Chu, D.A. Low, J.H. Lewis, Deep learning approaches using 2d and 3d convolutional neural networks for generating male pelvic synthetic computed tomography from magnetic resonance imaging, *Med. Phys.* 46 (2019) 3788–3798, <http://dx.doi.org/10.1002/mp.13672>, URL: <https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.13672>, arXiv:https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.13672.
- [40] Y. Li, W. Li, J. Xiong, J. Xia, Y. Xie, Comparison of supervised and unsupervised deep learning methods for medical image synthesis between computed tomography and magnetic resonance images, *BioMed Res. Int.* 2020 (2020) 5193707, <http://dx.doi.org/10.1155/2020/5193707>, URL: <https://onlinelibrary.wiley.com/doi/abs/10.1155/2020/5193707>, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1155/2020/5193707.
- [41] A. Thummerer, E. van der Bijl, A. Galapon Jr, J.J.C. Verhoeff, J.A. Langendijk, S. Both, C.N.A.T. van den Berg, M. Maspero, Synthrad2023 grand challenge dataset: Generating synthetic ct for radiotherapy, *Med. Phys.* 50 (2023) 4664–4674, <http://dx.doi.org/10.1002/mp.16529>, URL: <https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.16529>, arXiv:https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.16529.
- [42] S. Han, C. Hémon, B. Texier, Y. Kortli, A. Queffelec, R. de Crevoisier, J. Dowling, P. Greer, I. Bessières, A. Barateau, C. Lafond, J.-C. Nunes, Balancing data consistency and diversity: Preprocessing and online data augmentation for multi-center deep learning-based mr-to-ct synthesis, *Pattern Recognit. Lett.* 189 (2025) 56–63, <http://dx.doi.org/10.1016/j.patrec.2025.01.002>, URL: <https://www.sciencedirect.com/science/article/pii/S0167865525000029>.
- [43] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, 2019, URL: <https://arxiv.org/abs/1912.01703>, arXiv:1912.01703.
- [44] M.J. Cardoso, W. Li, R. Brown, N. Ma, E. Kerfoot, Y. Wang, B. Murrey, A. Myronenko, C. Zhao, D. Yang, V. Nath, Y. He, Z. Xu, A. Hatamizadeh, A. Myronenko, W. Zhu, Y. Liu, M. Zheng, Y. Tang, I. Yang, M. Zephyr, B. Hashemian, S. Alle, M.Z. Darestani, C. Budd, M. Modat, T. Vercauteren, G. Wang, Y. Li, Y. Hu, Y. Fu, B. Gorman, H. Johnson, B. Genereaux, B.S. Erdal, V. Gupta, A. Diaz-Pinto, A. Dourson, L. Maier-Hein, P.F. Jaeger, M. Baumgartner, J. Kalpathy-Cramer, M. Flores, J. Kirby, L.A.D. Cooper, H.R. Roth, D. Xu, D. Bericat, R. Floca, S.K. Zhou, H. Shuaib, K. Farahani, K.H. Maier-Hein, S. Aylward, P. Dogra, S. Ourselin, A. Feng, Monai: An open-source framework for deep learning in healthcare, 2022, URL: <https://arxiv.org/abs/2211.02701>, arXiv:2211.02701.
- [45] D.J. Beutel, T. Topal, A. Mathur, X. Qiu, J. Fernandez-Marques, Y. Gao, L. Sani, K.H. Li, T. Parcollet, P.P.B. de Gusmão, N.D. Lane, Flower: A friendly federated learning research framework, 2022, URL: <https://arxiv.org/abs/2007.14390>, arXiv:2007.14390.
- [46] L. Bouza, A. Bugeau, L. Lannelongue, How to estimate carbon footprint when training deep learning models? a guide and review, *Environ. Res. Commun.* 5 (2023) 115014, <http://dx.doi.org/10.1088/2515-7620/acf81b>.
- [47] J. Yosinski, J. Clune, Y. Bengio, H. Lipson, How transferable are features in deep neural networks? *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, MIT Press, Cambridge, MA, USA, 2014, pp. 3320–3328.
- [48] Q. Liu, J. Liang, C.W. Stanhope, D. Yan, The effect of density variation on photon dose calculation and its impact on intensity modulated radiotherapy and stereotactic body radiotherapy, *Med. Phys.* 43 (2016) 5717.
- [49] A.T. Davis, S. Muscat, A.L. Palmer, D. Buckle, J. Earley, M.G.J. Williams, A. Nisbet, Radiation dosimetry changes in radiotherapy treatment plans for adult patients arising from the selection of the CT image reconstruction kernel, *BJR Open* 1 (2019) 20190023.
- [50] P. Delanoë, D. Tchuente, G. Colin, Method and evaluations of the effective gain of artificial intelligence models for reducing co2 emissions, *J. Environ. Manag.* 331 (2023) 117261, <http://dx.doi.org/10.1016/j.jenvman.2023.117261>, URL: <https://www.sciencedirect.com/science/article/pii/S030147972300049X>.
- [51] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 770–778, <http://dx.doi.org/10.1109/CVPR.2016.90>.
- [52] A. Canziani, A. Paszke, E. Culurciello, An analysis of deep neural network models for practical applications, 2017, URL: <https://arxiv.org/abs/1605.07678>, arXiv:1605.07678.
- [53] X. Ma, J. Zhang, S. Guo, W. Xu, Layer-wise model aggregation for personalized federated learning, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 10082–10091, <http://dx.doi.org/10.1109/CVPR52688.2022.00985>.
- [54] D. Wu, L. Wong, B. Varghese, Fedfreeze: A dual-phase layer freezing framework for federated learning, *Future Gener. Comput. Syst.* 177 (2026) 108250, <http://dx.doi.org/10.1016/j.future.2025.108250>, URL: <https://www.sciencedirect.com/science/article/pii/S0167739X25005448>.

6.2.1 Supplementary Material

Supplementary Material for

Energy-Efficient Federated Learning via Adaptive Encoder Freezing for MRI-to-CT Conversion: A Green AI-Guided Research

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

^a *Institute of Biomedical Engineering, Karlsruhe Institute of Technology, Fritz-Haber-Weg 1, Karlsruhe, 76131, Baden-Württemberg, Germany*

^b *Department of Political Science, Università Degli Studi Di Teramo, Via Renato Balzarini 1, Teramo, 64100, Abruzzo, Italy*

^c *Department of Radiation Oncology, University Medical Center Schleswig-Holstein, Arnold-Heller-Strasse 3, Kiel, 24105, Schleswig-Holstein, Germany*

^d *Department of Experimental and Clinical Medicine, Magna Graecia University, Viale Europa, Catanzaro, 88100, Calabria, Italy*

*Corresponding author.

E-mail address: ciro.raggio@kit.edu

S.1 - Visual Comparison of Generated sCTs

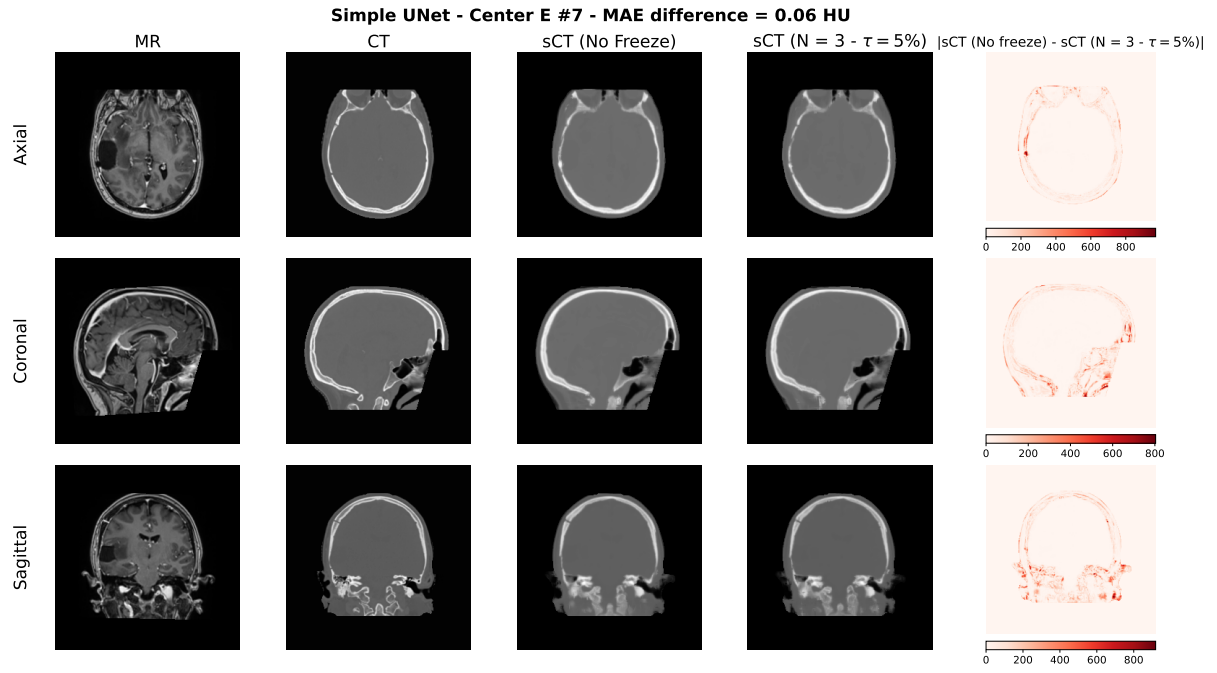
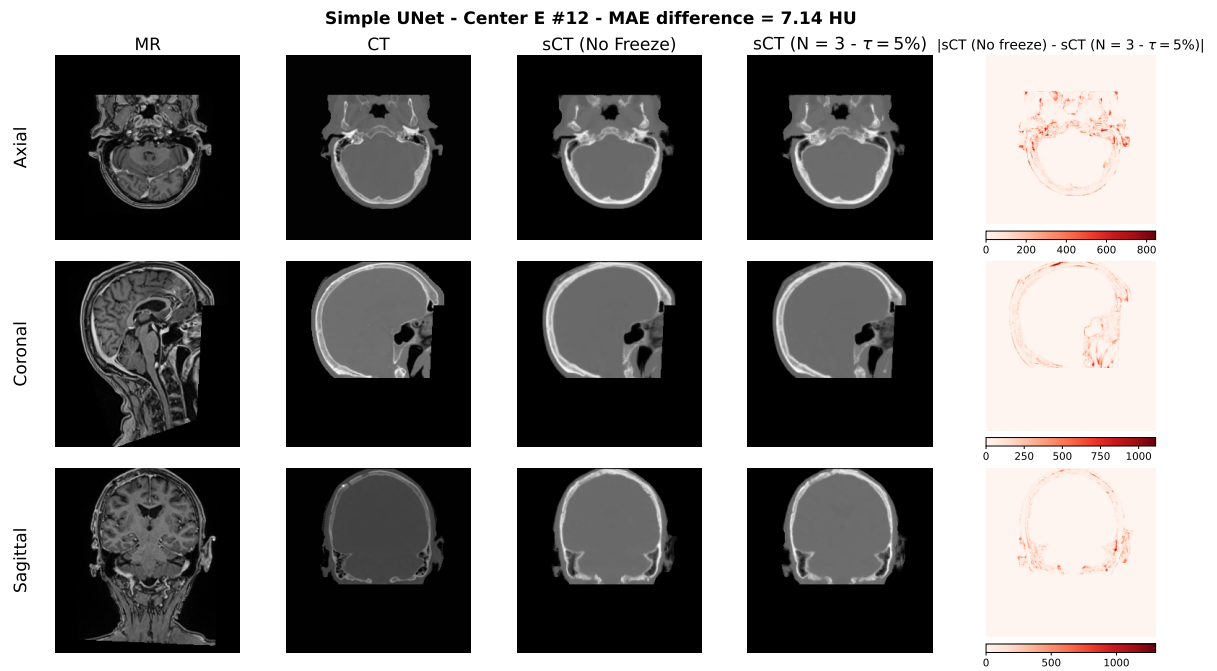
To qualitatively assess the potential impact of the adaptive freezing strategy on the generated synthetic CTs (sCTs), we conducted a visual comparison between the sCTs obtained with and without the proposed methodology. For each model employed in the study, we selected two representative test cases based on the following criteria:

- The case with the **minimum difference in terms of MAE**, thus the lowest difference between the sCTs MAE generated with and without adaptive freezing.
- The case with the **maximum difference in terms of MAE**, thus the highest difference between the sCTs MAE generated with and without adaptive freezing.

Each figure presents the following for the axial, coronal and sagittal plane of the central slice:

1. Input MR image;
2. Ground-truth CT image;
3. sCT generated without adaptive freezing;
4. sCT generated with the proposed adaptive freezing, using $\mathcal{N} = 3$ and $\tau = 5\%$;

5. Absolute difference map between the two sCTs.



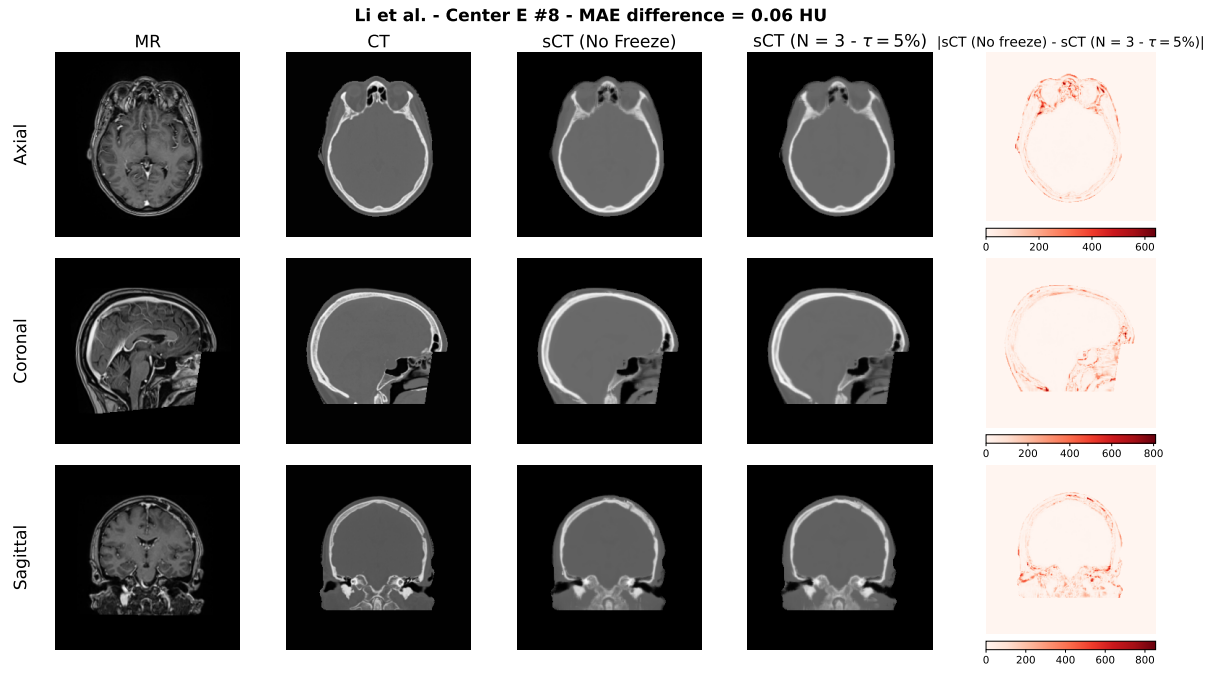


Figure S3: Li et al. architecture [2] – Minimum MAE difference case.

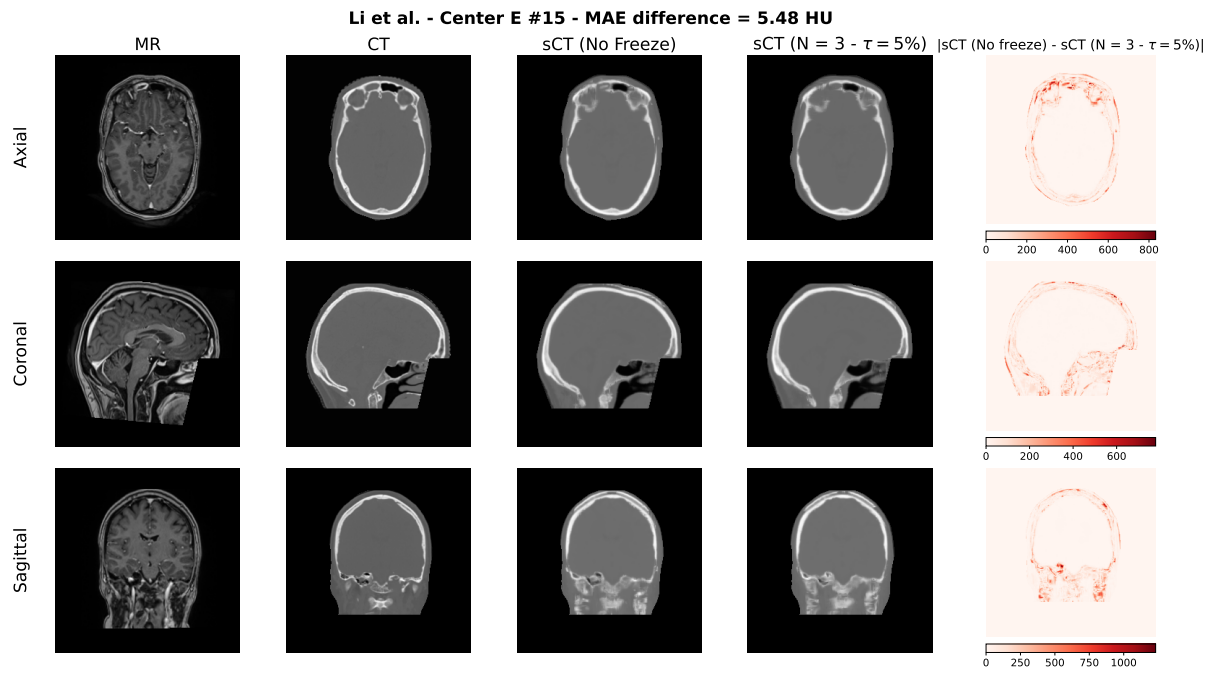


Figure S4: Li et al. architecture [2] – Maximum MAE difference case.

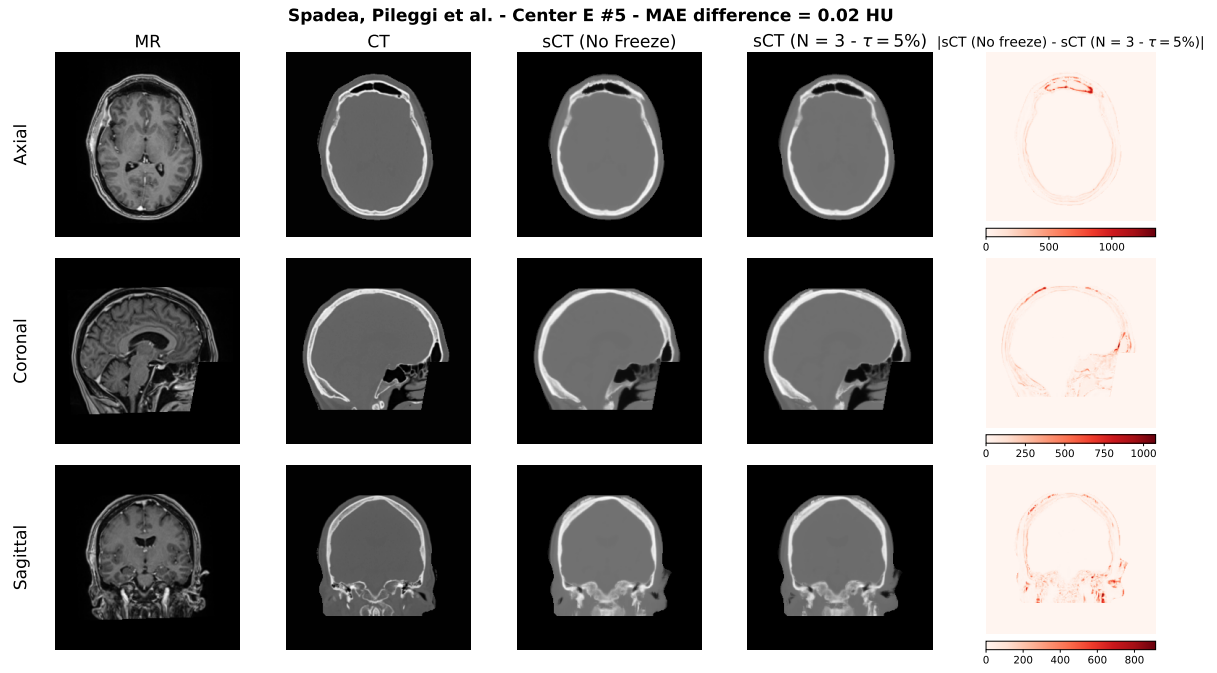


Figure S5: Spadea, Pileggi et al. architecture [3] – Minimum MAE difference case.

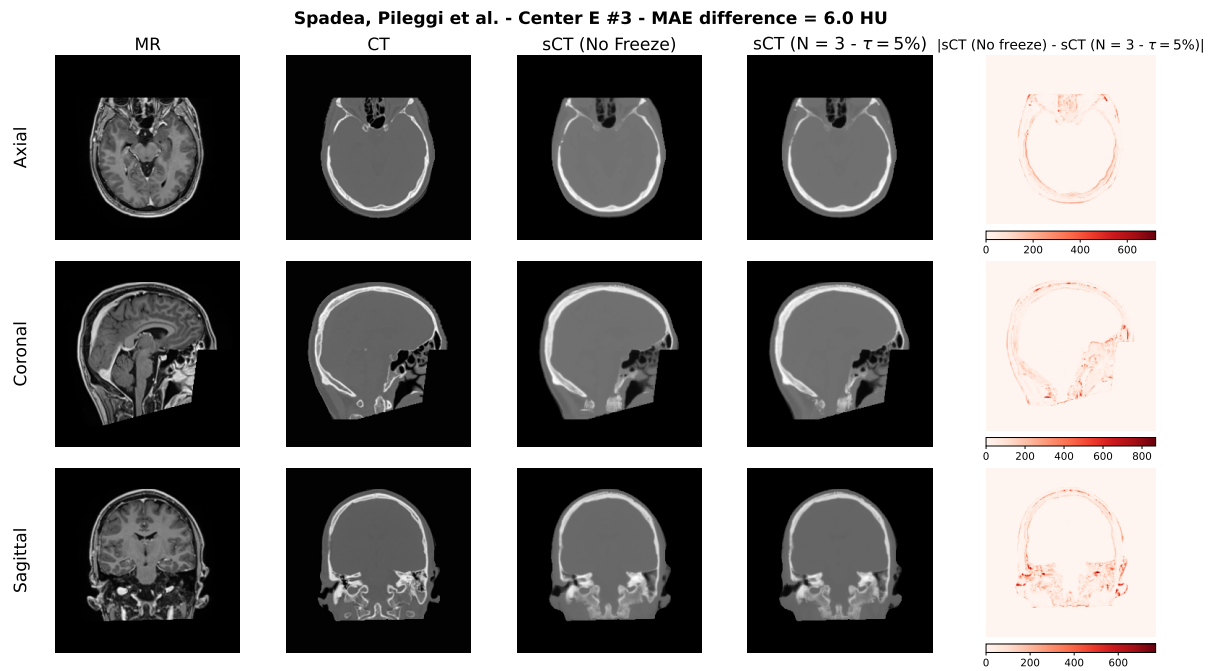


Figure S6: Spadea, Pileggi et al. architecture [3] – Maximum MAE difference case.

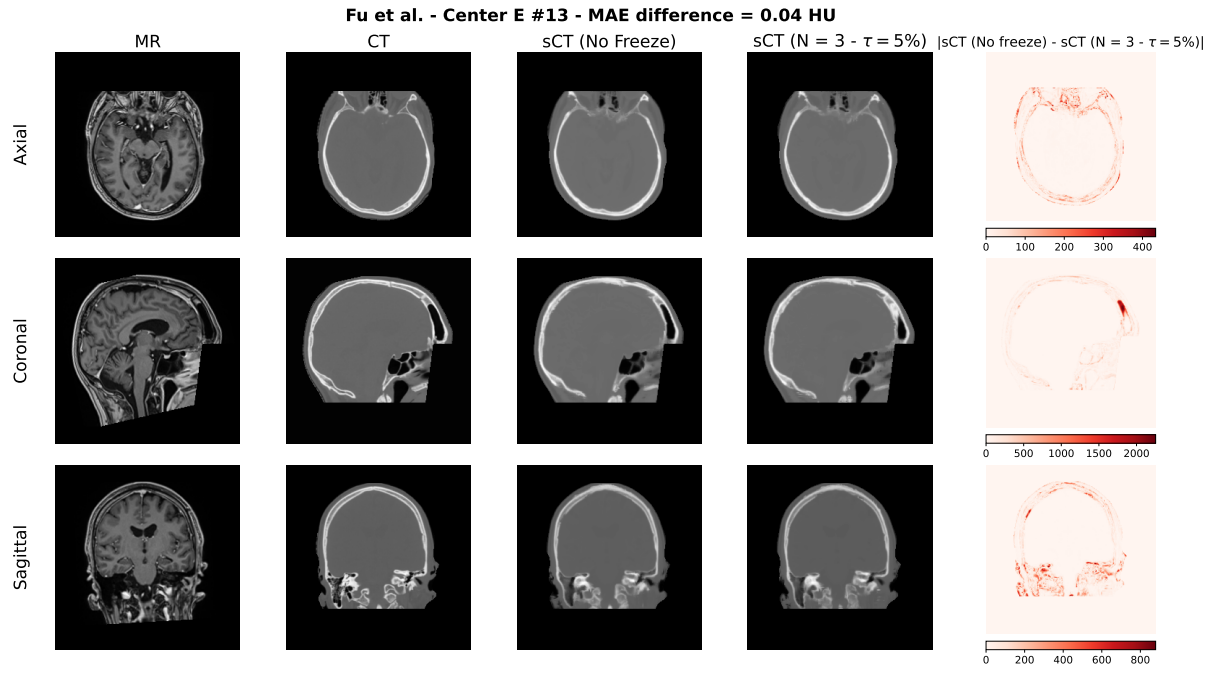


Figure S7: Fu et al. architecture [4] – Minimum MAE difference case.

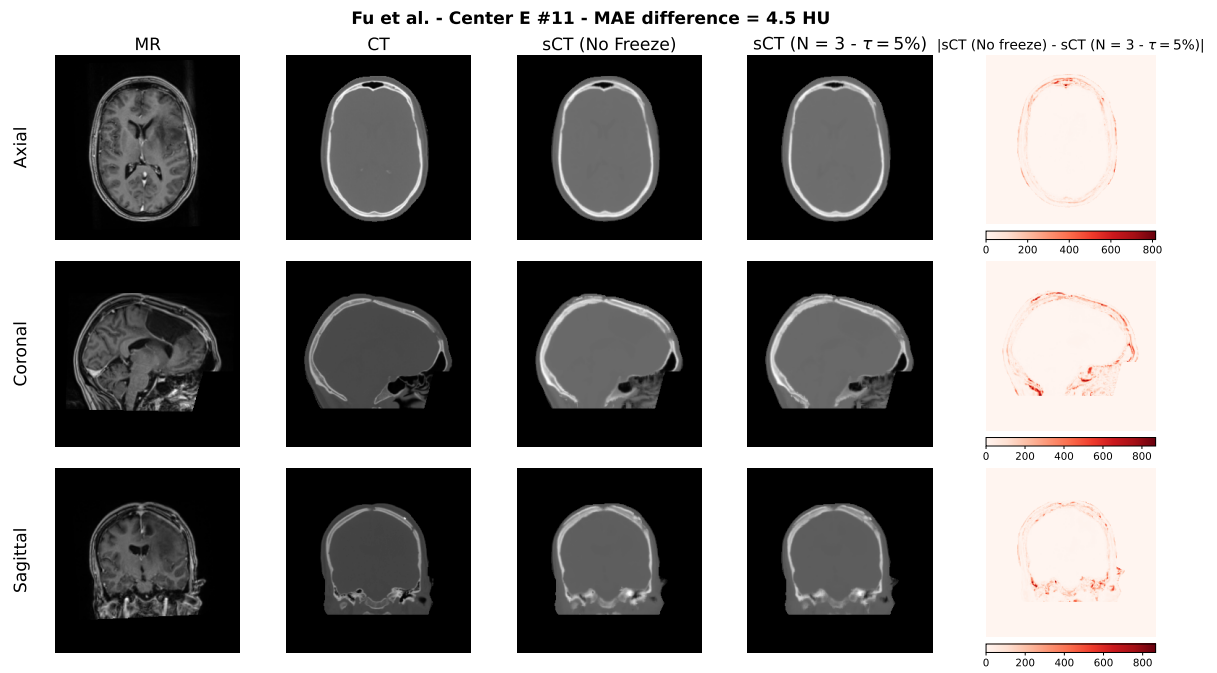


Figure S8: Fu et al. architecture [4] – Maximum MAE difference case.

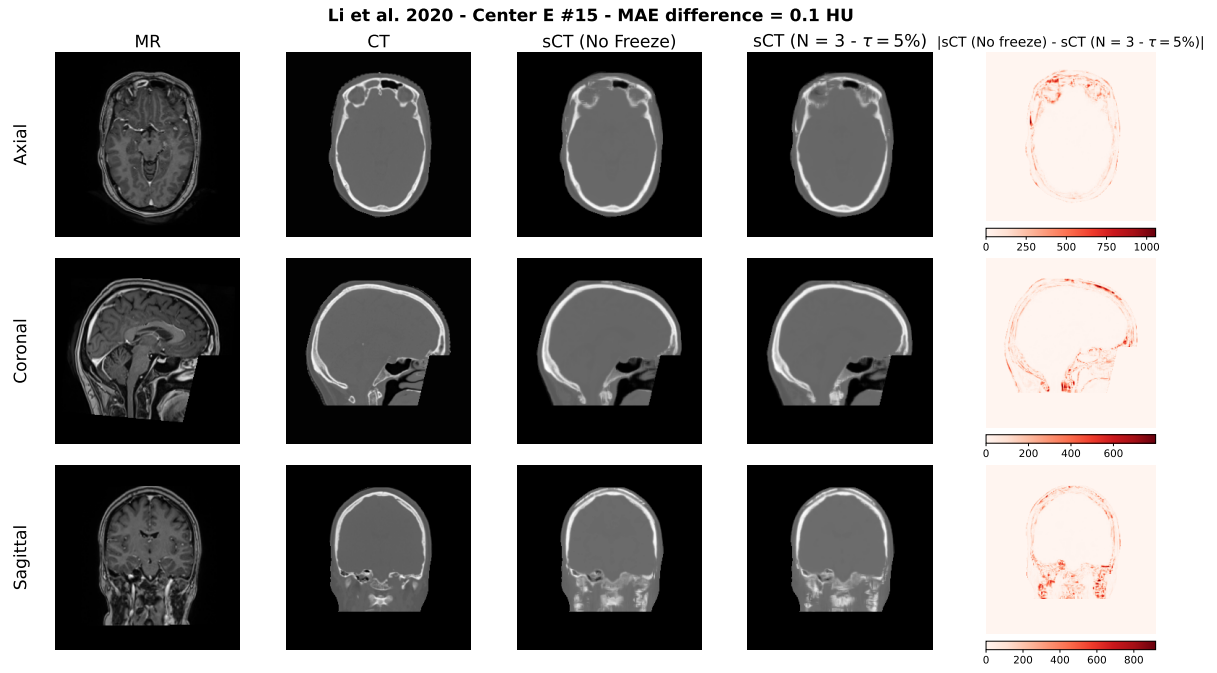


Figure S9: Li et al. 2020 architecture [5] – Minimum MAE difference case.

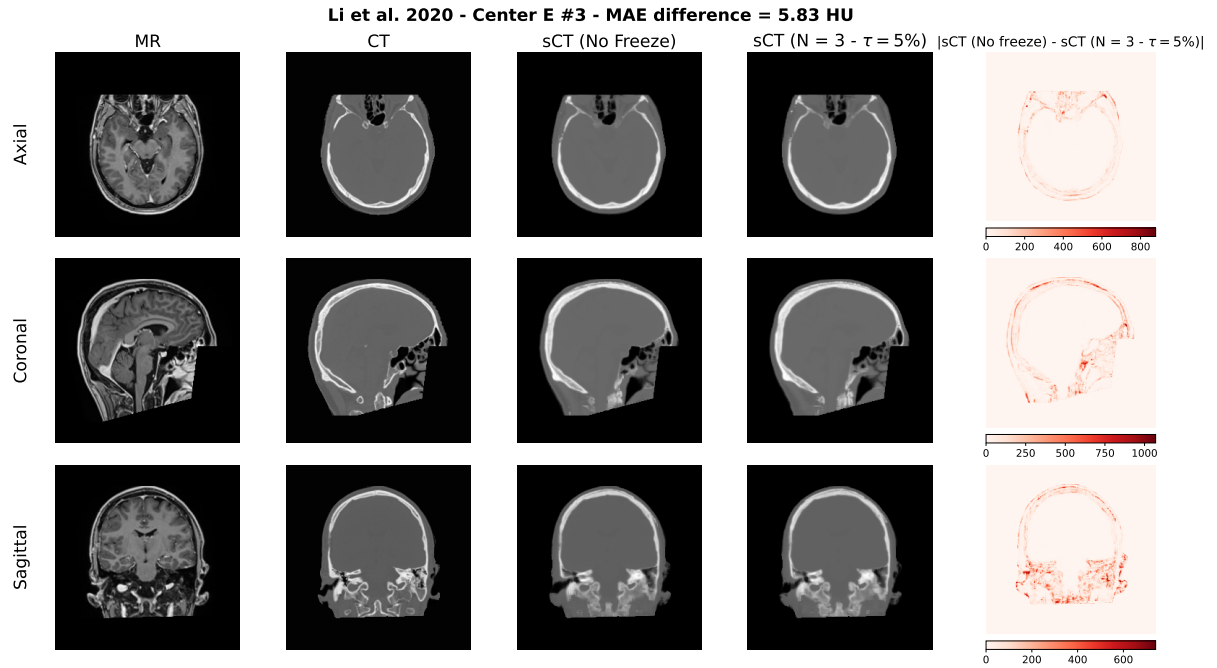


Figure S10: Li et al. 2020 architecture [5] – Minimum MAE difference case.

Across all models and both selected cases, no relevant and systematic structural differences were observed. The differences highlighted in the absolute difference maps and the variation in MAE are attributable to the stochastic nature of the training process (e.g., batch shuffling and optimization path variability) rather than the direct effect of the adaptive freezing mechanism.

These findings are aligned with the results observed in terms of MAE, PSNR and SSIM presented in this study.

References

- [1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.
- [2] Yafen Li, Wen Li, Pin He, Jing Xiong, Jun Xia, and Yaoqin Xie. CT synthesis from MRI images based on deep learning methods for MRI-only radiotherapy. In *2019 International Conference on Medical Imaging Physics and Engineering (ICMIPE)*, pages 1–6, 2019.
- [3] Maria Francesca Spadea, Giampaolo Pileggi, Paolo Zaffino, Patrick Salome, Ciprian Catana, David Izquierdo-Garcia, Francesco Amato, and Joao Seco. Deep convolution neural network (dcnn) multiplane approach to synthetic ct generation from mr images—application in brain proton therapy. *International Journal of Radiation Oncology*Biophysics*, 105(3):495–503, 2019.
- [4] Jie Fu, Yingli Yang, Kamal Singhrao, Dan Ruan, Fang-I Chu, Daniel A. Low, and John H. Lewis. Deep learning approaches using 2d and 3d convolutional neural networks for generating male pelvic synthetic computed tomography from magnetic resonance imaging. *Medical Physics*, 46(9):3788–3798, 2019.
- [5] Yafen Li, Wen Li, Jing Xiong, Jun Xia, and Yaoqin Xie. Comparison of supervised and unsupervised deep learning methods for medical image synthesis between computed tomography and magnetic resonance images. *BioMed Research International*, 2020(1):5193707, 2020.

Security Analysis of a Federated Learning Framework for Medical Image-to-Image Translation

7.1 Introduction

Federated Learning (FL) has emerged as a prominent paradigm for privacy-preserving collaborative training of Deep Learning (DL) models across multiple institutions. Nevertheless, the proposed adoption in clinical settings has often relied on the assumption that the exchange of model updates, rather than raw patient data, inherently guarantees security and privacy preservation [19]. However, this assumption has proven to be inadequate when considering realistic adversarial threat scenarios [122, 123]. Indeed, federated frameworks remain vulnerable to a range of attacks. These include gradient inversion, which enables private training data to be reconstructed from shared model updates [124], and membership inference, in which an adversary determines whether a specific sample was included in the training set [125]. Data poisoning represents a further vulnerability, enabling a compromised client to manipulate the global model by injecting adversarial inputs during local training [126]. Consequently, the absence of explicit security evaluation in federated frameworks represents a methodological limitation, particularly in medical imaging applications, where compromised model outputs may propagate undetected errors directly into clinical workflows.

Existing federated frameworks, including FedSynthCT (see Chapter 4–5), have demonstrated the feasibility of achieving effective cross-institutional generalisation without the direct exchange of patient data. However, previous studies [127, 128] have largely restricted security evaluations in federated medical imaging to classification tasks and GAN-based augmentation pipelines, leaving the vulnerability of federated image synthesis frameworks essentially unexplored. This gap is particularly significant because, unlike classification tasks, I2I translation models directly generate synthetic medical images that are subsequently

integrated into downstream clinical procedures, such as image segmentation and dose calculation, where subtle adversarial perturbations may remain undetected by conventional evaluation metrics.

The study presented in this Chapter investigated the security of FedSynthCT through a multi-threat evaluation [129]. The analysis was conducted within the experimental setting of FedSynthCT-Brain [99] (introduced in Chapter 4), which provided a reference environment. Three representative attack vectors were implemented and assessed: (i) Deep Leakage from Gradients (DLG) [124], a gradient inversion attack aimed at reconstructing private MRI data of clients from shared model updates; (ii) the Federated Membership Inference Attack (FedMIA) [125], which exploits aggregated client updates to infer patient-level membership information; and (iii) data poisoning, where a malicious client degrades global model performance by corrupting its local training data. For each attack category, corresponding defence strategies were also evaluated. Specifically, the robustness of the UNet architecture proposed in FedSynthCT-Brain and federated training configuration against DLG was characterised; Secure Aggregation (SecAgg) [130] was investigated as a cryptographic countermeasure against membership inference; and Byzantine-robust coordinate-wise median aggregation (FedMedian) [131] was examined as a defence against poisoning-induced degradation of model integrity.

The results demonstrated that the original framework exhibited an inherent robustness to gradient inversion attacks, primarily attributable to architectural and training-related factors, while simultaneously revealing a substantial vulnerability to membership inference in the absence of cryptographic protection mechanisms. SecAgg reduced attack efficacy without compromising synthesis quality. In contrast, data poisoning under standard aggregation produced severe degradation in model performance at high noise levels, whereas FedMedian considerably mitigated these effects, albeit with residual limitations under low-intensity noise injection. Overall, these findings demonstrate that federated I2I translation frameworks require explicit and multi-layered security evaluation, and that established defence strategies can provide effective and complementary protection against distinct classes of adversarial threats.

The content of this chapter has been adapted from a manuscript previously published in the peer-reviewed journal “Journal of Medical Systems”. Portions of the text are reproduced verbatim from the published article [129]. The author of this thesis led the conceptualisation of the study, the design of the attacks and the analytical evaluation. Furthermore, the author led the drafting and revision of the original manuscript in close collaboration with the co-authors.

7.2 Manuscript



Security Analysis of a Federated Learning Framework for Medical Image-to-Image Translation

Ciro Benito Raggio¹ · Lina Bucher² · Oliver Blanck³ · Francesco Cicone⁴ · Paolo Zaffino⁴ · Maria Francesca Spadea¹

Received: 24 April 2026 / Accepted: 27 June 2026
© The Author(s) 2026

Abstract

Federated Learning (FL) emerged as a privacy-preserving paradigm for collaborative training of deep learning models across institutions without sharing patient data. This approach has been applied to complex tasks such as medical image-to-image (I2I) translation, including MRI-to-synthetic CT (sCT) generation. However, existing federated I2I frameworks often assume privacy preservation as an inherent property of FL rather than a requirement to be explicitly validated, leaving their robustness to representative adversarial threat scenarios largely unexplored. In this study, we evaluated the vulnerability of a federated MRI-to-sCT translation framework (FedSynthCT-Brain) to three representative attack classes: Deep Leakage from Gradients (DLG), Federated Membership Inference Attack (FedMIA), and data poisoning. The efficacy of corresponding defense mechanisms, such as Secure Aggregation (SecAgg) and Byzantine-robust median aggregation (FedMedian), were assessed. DLG enabled only the recovery of coarse anatomical structures, with no clinically identifiable details ($SSIM \leq 0.16$, $PSNR \leq 11$ dB) across clients, suggesting limited vulnerability under the evaluated DLG setting. In contrast, FedMIA achieved high membership discrimination, with AUC scores between 0.92 and 0.99, revealing a critical privacy vulnerability. The introduction of SecAgg reduced AUC values to near-random levels (0.23–0.56) across all centers without impacting synthesis quality. Under high-noise poisoning, the standard federated averaging (FedAvg) aggregation rendered the federation inoperative, while FedMedian restored performance close to the no-poisoning baseline in most scenarios, with significant residual degradation in specific center configurations. At low noise levels, the advantage of FedMedian was less consistent, as low-level noise injection may be indistinguishable from natural heterogeneity across centers, potentially enabling stealthy degradation. These findings demonstrate that federated I2I translation frameworks are not inherently secure and require explicit, multi-layered evaluation. As FL is increasingly adopted in clinical workflows, our results underscore the necessity of integrating cryptographic, algorithmic, and infrastructural safeguards for secure deployment.

Keywords Federated learning · Image-to-image translation · Synthetic computed tomography · Security attacks

Ciro Benito Raggio and Lina Bucher contributed equally to this work.

✉ **Ciro Benito Raggio**
ciro.raggio@kit.edu
Lina Bucher
ia254@uni-heidelberg.de
Oliver Blanck
oliver.blanck@uksh.de
Francesco Cicone
cicone@unicz.it
Paolo Zaffino
p.zaffino@unicz.it
Maria Francesca Spadea
mf.spadea@kit.edu

- ¹ Institute of Biomedical Engineering, Karlsruhe Institute of Technology, Fritz-Haber-Weg 1, 76131 Karlsruhe, Baden-Württemberg, Germany
- ² Department of Physics and Astronomy, Heidelberg University, Im Neuenheimer Feld 226, 69120 Heidelberg, Baden-Württemberg, Germany
- ³ Department of Radiation Oncology, University Hospital Schleswig-Holstein, Feldstrasse 21, 24105 Kiel, Schleswig-Holstein, Germany
- ⁴ Department of Experimental and Clinical Medicine, Magna Graecia University, Viale Europa, 88100 Catanzaro, Calabria, Italy

Introduction

Deep Learning (DL) transformed medical imaging by enabling models to learn complex mappings between modalities directly from data [1]. Within this landscape, image-to-image (I2I) translation emerged as a prominent framework, encompassing tasks such as inter-sequence MRI translation and MRI- and Cone-Beam Computed Tomography (CBCT)-to-CT synthesis [2, 3]. Among these, synthetic CT (sCT) generation has attracted particular clinical interest in radiotherapy treatment planning, as it eliminates the need for dedicated CT acquisition, thus removing associated ionizing radiation dose and co-registration errors [3].

However, developing effective DL models requires large, diverse, multi-center datasets [4, 5], which institutions often cannot share due to regulatory constraints such as the GDPR [6, 7] and patient privacy requirements. This motivated the adoption of Federated Learning (FL), which enables collaborative training by exchanging only model updates rather than raw patient data [8]. Beyond its application in classification and segmentation tasks, FL has been successfully applied to federated MRI cross-modality translation [9–11]. More recently, it has been extended to sCT synthesis through FedSynthCT-Brain [12], a framework for brain MRI-to-sCT translation across four centers, and further extended to CBCT-to-sCT translation [13].

These approaches demonstrated effective cross-center generalization without requiring the exchange of patient data. However, none of these frameworks included an explicit security evaluation, treating privacy preservation as an inherent property of FL rather than a design requirement to be verified. Although FL mitigates direct data sharing, federated systems remain vulnerable to well-established adversarial threats, including privacy leakage from shared updates and integrity attacks that can compromise model behavior [14–16]. Prior work on FL security in medical imaging focused primarily on augmentation, classification and segmentation tasks [17], such as backdoor attacks in GAN-based data augmentation [18], poisoning for classification [19], and evasion attacks via adversarial perturbations in classification tasks [20].

To the best of the authors' knowledge, no study has evaluated security threats in the context of federated I2I translation. This work investigated representative adversarial threats and defense strategies, analyzing their trade-off between clinical utility, computational feasibility, and robustness.

Related Work

The security of FL systems encompasses two principal threat dimensions, including privacy attacks, which target the extraction of sensitive information from model updates, and integrity attacks, which aim to corrupt the global model through malicious updates. Both classes of attacks exploit the exchange of model updates in the FL protocol, which constitutes a primary attack surface in the absence of dedicated security mechanisms [15, 16].

Within privacy-related attacks, membership inference attacks (MIAs) exploit the tendency of overfitted models to behave differently on training versus unseen data. Nasr et al. [21] showed that FL models are not inherently more resistant to membership inference than centralized counterparts, and that client update vectors extend the attack surface. Zhu et al. proposed FedMIA [22], exploiting aggregated update information across clients and rounds. In addition, prior studies have investigated client participation inference from update dynamics, highlighting that even coarse-grained aggregation may still expose sensitive information about client involvement [23, 24].

In contrast to privacy-oriented attacks, integrity threats, such as data poisoning and backdoor attacks, represent a growing trend. Bagdasaryan et al. [25] demonstrated that a single compromised client can inject a targeted backdoor via model replacement within a single aggregation round, with subsequent work exploring increasingly stealthy strategies [26, 27]. In medical imaging, poisoning has been studied for classification [19] and federated GAN-based data augmentation [18], where the downstream impact is ultimately measured on classification performance. The implications for I2I translation are fundamentally different, as compromised synthesis models may introduce subtle intensity distortions directly propagated into clinical workflows (i.e., radiotherapy dose calculation), where such errors are difficult to detect using standard evaluation metrics.

Beyond MIAs and poisoning, gradient-based attacks represent another major privacy leakage vector. Zhu et al. demonstrated that private training data can be reconstructed from shared gradients via iterative optimization (Deep Leakage from Gradients), achieving pixel-wise accuracy on image data [28]. Subsequent work showed that reconstruction quality degrades with larger batch sizes [29], and batch normalization statistics substantially hinder standard attacks [30]. Furthermore, recent work extended inversion to diffusion-based reconstruction in medical imaging [31].

Several complementary defense strategies have been proposed [15, 16]. For instance, differential privacy (DP) [32] adds calibrated noise to client updates, providing formal privacy guarantees against gradient inversion and membership inference [21, 29], at the cost of a utility trade-off

particularly concerning in medical image synthesis. Secure aggregation (SecAgg) [33] prevents individual gradient exposure via cryptographic commitments, but offers no protection against poisoning. Byzantine-robust aggregation [34] mitigates integrity threats by down-weighting anomalous updates, but provides no privacy guarantees.

Despite extensive research on FL security, prior studies primarily focused on classification, segmentation, and data augmentation tasks. Consequently, established privacy and integrity attacks, as well as corresponding defense mechanisms, have not been evaluated in I2I translation settings, particularly in clinically relevant synthesis tasks such as sCT generation. To address this gap, we implemented three well-established attack vectors against FedSynthCT-Brain [12]: (i) deep leakage from gradients (DLG), (ii) federated membership inference, and (iii) data poisoning. Furthermore, SecAgg and FedMedian were implemented as representative defense strategies.

Materials and Methods

Datasets

The study included 102 patients across five centers (A–E). Centers A–C represented institutional datasets from the US, Italy, and Germany, while Centers D and E were extracted from the SynthRAD Grand Challenge 2023 [35]. Detailed dataset characteristics (i.e., scanner specifications, acquisition protocols) were reported in Table 1. Following the original FedSynthCT-Brain [12] implementation, each training center (Centers A, B, C, D) partitioned data into training, validation, and test sets to maximize training cases. Overall, this corresponded to a 70/10/20 split. Center E was used strictly for the external evaluation and generalization

assessment of the federated model and thus never participated in model training.

Notable heterogeneity was observed across sites regarding scanner vendors, field strengths, and voxel spacing. The MRI acquisitions included both 1.5T and 3T systems, while the CT scans were acquired with varying tube voltages (120–140 kVp) and heterogeneous reconstruction settings. The dimensions of the original images also varied considerably across centers, particularly for Centers C, D and E.

The pre-processing was performed at each client level, including rigid registration of MRI and CT volumes. N4 bias correction [36] and normalization to [0,1] were applied to MRI volumes. Subsequently, all MRI and CT volumes were resampled to a fixed dimension of $256 \times 256 \times 256$ using a combination of cropping, resizing, and constant padding. Non-anatomical regions (i.e., scanner couch and background) were masked, assigning -1000 HU to CT and 0 to MRI [12].

Federated Learning Framework

The FL workflow was implemented using the Flower framework [37], while model development and training were performed in PyTorch [38] and MONAI [39]. Following the original implementation, the model aggregation was performed via FedAvg, with local updates regularized using the FedProx [40] strategy, which introduces a proximal term in the local objective to limit the deviation from the global model. The proximal coefficient was set to $\mu = 3$ [12].

Furthermore, as delineated in the original implementation, a 2D UNet-based model [41] (with $\approx 2.68 \times 10^7$ trainable parameters) was employed as the foundation for all experiments, thereby facilitating the examination of the attacks on the original framework. The network consisted of four encoding levels with channel depths of 64, 128, 256, and 512, followed by a bottleneck block that doubled the

Table 1 Overview of dataset characteristics across the five participating centers. The table reports patient counts, image dimensions, scanner models, acquisition parameters, and voxel spacing for both MRI and CT modalities, highlighting inter-center heterogeneity

		Centers				
		A	B	C	D	E
MRI	Patients	15	14	21	29	23
	Original dimension	$256 \times 176 \times 256$	$256 \times 176 \times 248$	$226-357 \times 512 \times 512$	$167-213 \times 216-262 \times 250-277$	$167-262 \times 200-225 \times 225-248$
	Voxel size [mm^3]	$1 \times 1 \times 1$	$1 \times 1 \times 1$	$0.78 \times 0.78 \times 1$	$0.98-1.12 \times 0.98-1.12 \times 0.98-1.12$	$0.98 \times 0.98 \times 0.98$
	Scanner	MAGNETOM Trio	Biographm MR	Vantage Titan	MAGNETOM Avanto_fit, Skyra, Vida_fit, Prisma_fit	MAGNETOM Aera, Avanto_fit
	Field strength [T]	3	3	1.5	1.5–3	1.5–3
CT	Voxel size [mm^3]	$0.49-0.67 \times 0.49-0.67 \times 2.5$	$0.98 \times 0.98 \times 3.27$	$0.78 \times 0.78 \times 1$	$0.69-0.78 \times 0.69-0.79 \times 1-3$	$0.59-1.27 \times 0.59-1.27 \times 1-2$
	Scanner	LightSpeed QX/i	Discovery ST	Brilliance Sensation Open	Brilliance Big Bore	SOMATOM Definition AS
	Tube voltage [kVp]	140	120	120	120	120

channel count to 1024 prior to upsampling. Downsampling was performed via max pooling, and upsampling employed transposed convolutions with a kernel size of 4 and a stride of 2. Batch normalization was applied after each convolutional layer. LeakyReLU activations were used in the encoder and standard ReLU activations in the decoder. Skip connections then concatenated the encoder feature maps to the corresponding decoder levels [41]. Data augmentation was applied during training. Specifically, random spatial transformations were used, including left-right flipping along the spatial axis, random rotations, and random translations, implemented using MONAI.

Local training was performed using the Random Multi-2D approach for 1 epoch per round [12]. Optimization was carried out using Adam with a fixed learning rate of 1×10^{-4} and a batch size of 8 slices.

The federated training was conducted for 20 communication rounds. Client participation followed a full participation strategy, where all available clients (Centers A–D) contributed in each round.

Model selection and evaluation were performed by selecting the final global model obtained at the last communication round.

Experimental Scenarios

Three categories of attacks were simulated within the federated framework, as illustrated in Fig. 1: (a) gradient inversion, (b) membership inference, and (c) data poisoning. Where applicable, corresponding defense mechanisms were evaluated within the same experimental setting.

Gradient Inversion

The DLG attack [28] was implemented with the aim of reconstructing local inputs from shared gradients. Therefore, in this scenario, the sensitive data to reconstruct corresponded to the input MRIs. For each attack, a 2D dummy MRI slice was initialized as a learnable tensor within the anatomical mask. The reconstruction was performed by

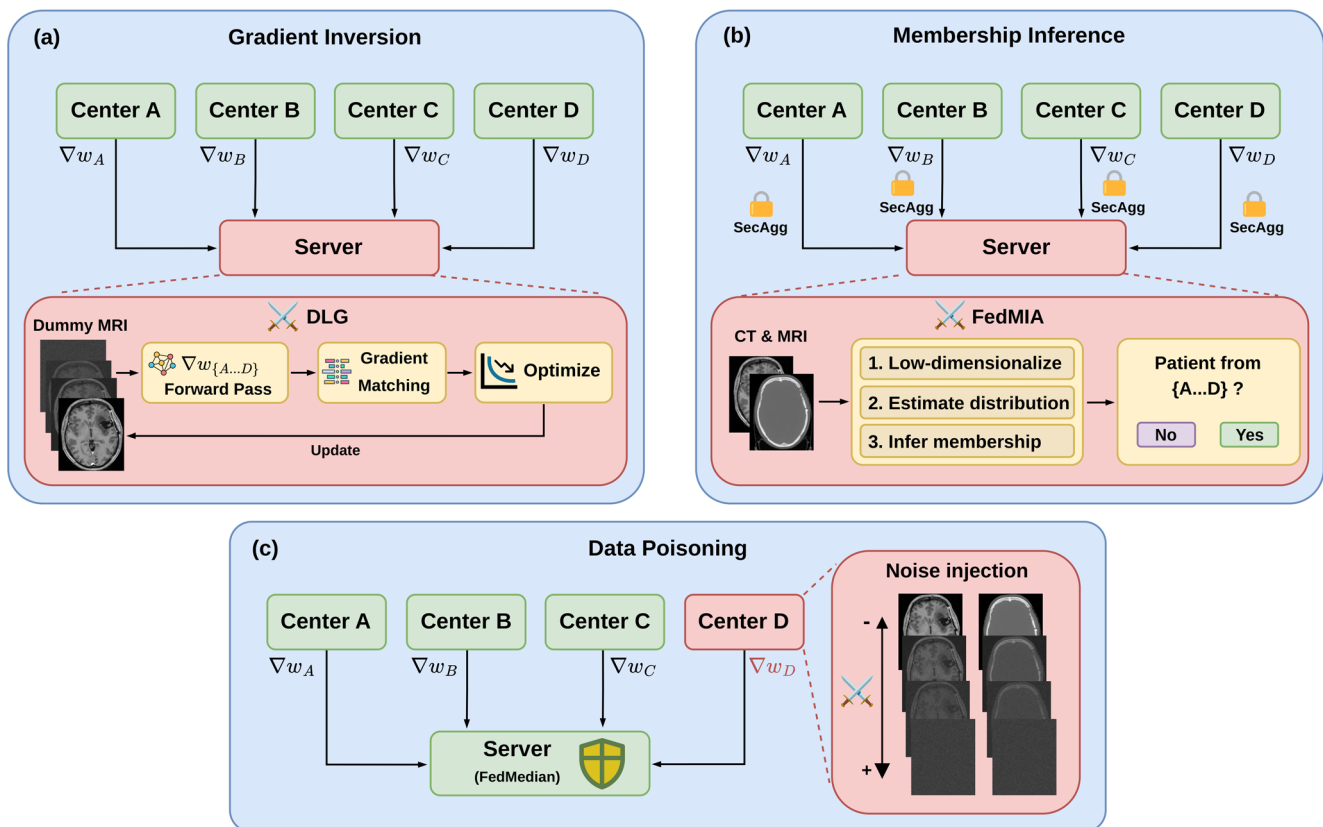


Fig. 1 Overview of the three attack scenarios and corresponding defenses evaluated in the proposed federated sCT synthesis framework. **(a) Gradient Inversion:** a honest-but-curious server attempts to reconstruct private MRI training data from the gradients $\nabla w_{\{A...D\}}$ via the DLG attack, which iteratively optimizes a dummy input through gradient matching. **(b) Membership Inference:** the FedMIA attack exploits client gradients to determine whether a given CT–MRI pair

was included in the training set of any participating center; SecAgg is adopted as a countermeasure, encrypting each client’s gradient prior to server-side aggregation. **(c) Data Poisoning:** a malicious client corrupts the federation by injecting additive noise into its local training data, thereby altering the contributed gradient; FedMedian aggregation is employed at the server as a robust defense against such poisoned updates

minimizing the gradient matching loss, defined as the squared difference between dummy and real gradients, computed exclusively within the anatomical mask to suppress background reconstruction.

The Adam and the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) optimizers [28] were evaluated separately. For Adam, a learning rate of $l_r = 0.1$ was used with a StepLR scheduler reducing l_r by a factor of $\lambda = 0.1$ every 10,000 iterations. For L-BFGS, a line search step size of 0.1 was adopted. Both optimizers were run for 30,000 iterations.

In this scenario, no explicit defense mechanism was required, and the analysis was conducted under the baseline [12] federated setting, while limiting the batch size to 1 to minimize DLG optimization complexity.

Membership Inference

For this scenario, the FedMIA [22] strategy was implemented. The attack targeted the server, which had access to the global model checkpoints at each round and to the client model updates uploaded prior to aggregation. The objective was to determine, for a given data sample, whether it was part of the training set of a specific target client. According to the FedMIA implementation, the attack followed a three-step procedure:

1. For each communication round r , the client update was defined as:

$$\Delta w_k^{(r)} = w_k^{(r)} - w^{(r)} \quad (1)$$

where $w_k^{(r)}$ represents the local model of client k and $w^{(r)}$ the global model. The cosine similarity between the gradient of the loss, computed on the global model with respect to each candidate sample (one MRI and CT batch), and the flattened $\Delta w_k^{(r)}$ was subsequently calculated, forming an $n_{\text{clients}} \times n_{\text{batches}}$ similarity matrix.

2. The non-membership distribution was estimated by fitting a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ to the similarity scores of non-target clients, where μ and σ represent the mean and standard deviation, respectively, and excluding samples with scores greater than $\mu + 3\sigma$ associated with high similarity and consequently potential membership.
3. The membership probability was computed as the Gaussian cumulative distribution function (CDF) evaluated at the target client score, with higher values indicating greater membership likelihood.

In order to establish the membership, in the simulated scenario the attacker had access to the original MRI and CT

data, providing an upper bound on attack and, consequently, defense effectiveness.

To mitigate this threat SecAgg [33] was applied. Under this protocol, individual client updates $\Delta w_k^{(r)}$ are masked through cryptographic pairwise secrets prior to transmission, ensuring that the server can only access the aggregated sum of updates without being able to reconstruct any individual contribution. This mechanism directly affects the assumptions underlying the FedMIA attack, which requires access to per-client update vectors $\Delta w_k^{(r)}$ to compute cosine similarity scores against candidate samples. Therefore, the application of SecAgg precludes the observation of these vectors, thereby rendering the calculation of similarity and, consequently, the estimation of membership probabilities, no longer viable. The experiments employed the built-in SecAgg+ [42] implementation provided by the Flower framework, which follows the additive masking scheme proposed by Bonawitz et al.

The FedMIA attack was subsequently re-evaluated under this aggregation setting to assess whether the attack pipeline remains applicable and to quantify the resulting privacy risk mitigation.

Data Poisoning

The impact of the data poisoning attack was assessed by corrupting the training data of a client during the federated training process. The primary objective of the attack was to degrade the performance of the global model on the remaining clients by injecting corrupted data into the federation. The poisoning strategy was implemented by introducing additive Gaussian noise to the MRI and CT volumes. For each patient, noise n was sampled from a Gaussian distribution with mean μ_x and standard deviation σ_x computed individually from the original CT or MRI volume x , and blended with it according to a noise level parameter $\gamma \in [0, 1]$, where $\gamma = 0$ preserves the original image and $\gamma = 1$ replaces it entirely with noise, according to:

$$x_{\text{poisoned}} = (1 - \gamma) \cdot x + \gamma \cdot n, \quad n \sim \mathcal{N}(\mu_x, \sigma_x) \quad (2)$$

The value of γ varied across three noise levels $\gamma \in \{0.1, 0.5, 1.0\}$. These correspond to 10%, 50% and 100% noise injection, respectively.

Subsequently, in order to quantify both vulnerability and robustness, the impact of data poisoning was evaluated under different aggregation strategies. As a baseline configuration, the standard FedAvg aggregation combined with FedProx [12] was first considered in the absence of adversarial attacks, establishing an upper bound on synthesis performance. The poisoning attack was then applied under the

FedAvg aggregation combined with FedProx to measure the degradation induced by a compromised client for each client in the federation (Centers A, B, C, D).

As a potential defense mechanism, a Byzantine-robust aggregation strategy based on coordinate-wise median (Fed-Median) [43] was evaluated as an alternative to FedAvg. In contrast to mean-based aggregation, FedMedian mitigates the influence of anomalous updates by selecting the median value across client parameters for each coordinate, thereby limiting the impact of outliers introduced by poisoned data. The integration with FedProx was maintained to ensure the closest possible alignment with the original framework.

Evaluation Metrics

Image synthesis quality and similarity were quantified using complementary metrics capturing both pixel-wise accuracy and perceptual fidelity [3, 12].

The Mean Absolute Error (MAE) measured pixel-wise prediction error between the synthesized image I_{Synth} and the ground-truth image I_{GT} in Hounsfield units (HU), and is defined as:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |I_{\text{Synth}}^{(i)} - I_{\text{GT}}^{(i)}|, \quad (3)$$

where N is the total number of voxels.

The Peak Signal-to-Noise Ratio (PSNR) evaluated reconstruction fidelity by comparing the maximum possible signal intensity I_{max} to the Mean Squared Error (MSE), defined as:

$$\text{PSNR} = 10 \log_{10} \left(\frac{I_{\text{max}}^2}{\text{MSE}} \right), \quad (4)$$

where

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N \left(I_{\text{syn}}^{(i)} - I_{\text{gt}}^{(i)} \right)^2. \quad (5)$$

The Structural Similarity Index (SSIM) assessed perceptual similarity by jointly comparing luminance, contrast, and structural information between image patches and was defined as:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \quad (6)$$

where μ_x and μ_y denoted the mean intensities, σ_x^2 and σ_y^2 the variances, and σ_{xy} the covariance of image patches x and y . Constants C_1 and C_2 were used to stabilize the division.

In the context of membership inference attacks, model privacy leakage was evaluated using the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). The ROC curve represents the trade-off between the True Positive Rate (TPR) and False Positive Rate (FPR), defined as:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}, \quad (7)$$

where TP, FP, TN, and FN denote true positives, false positives, true negatives, and false negatives, respectively. Consequently, the AUC summarized the ROC curve into a single scalar value, with higher values indicating stronger attack performance and thus greater privacy risk [22].

Results and Discussion

As described in Section “Gradient Inversion”, the DLG attack was performed using both the Adam and L-BFGS optimizers over 30, 000 iterations to ensure stable convergence. L-BFGS exhibited a slower and less stable behavior, with the gradient matching loss reaching a plateau after $\approx 10,000$ iterations. In contrast, Adam provided more stable optimization. However, as demonstrated in Fig. 2, the reconstruction improved within the initial 5, 000 iterations from a uniform initialization toward the anatomical shape of the original MRI, with no qualitative improvement beyond 15, 000 iterations. The reconstructed images failed to preserve identifiable anatomical details despite capturing coarse structural shapes, with SSIM of 0.16 ± 0.05 and PSNR of 11 ± 2 dB at optimizer convergence for the reference cases presented in Fig. 2, demonstrating the limited vulnerability of the framework against this attack. This can be attributed to multiple factors, including architectural and training factors, as well as the federated paradigm.

The UNet architecture employed in this study (proposed in [12]) incorporates batch normalization layers by design, which have been shown to substantially hinder gradient inversion attacks [30]. Furthermore, the federated training context implies that gradients are computed over multi-patient batches rather than single samples, a condition known to degrade reconstruction quality in gradient inversion settings [29]. The implemented DLG configuration represented an upper bound assessment of gradient inversion risk. The adoption of a reduced batch size resulted in an increased rate of information leakage per sample, thereby providing a more favorable scenario for the attacker in comparison to realistic FL implementations. In practical

Fig. 2 DLG-based reconstruction across federated centers. Reconstructions evolved from a uniform initialization to coarse anatomical shapes within the first 5,000 iterations, with no further qualitative improvement beyond 15,000 iterations

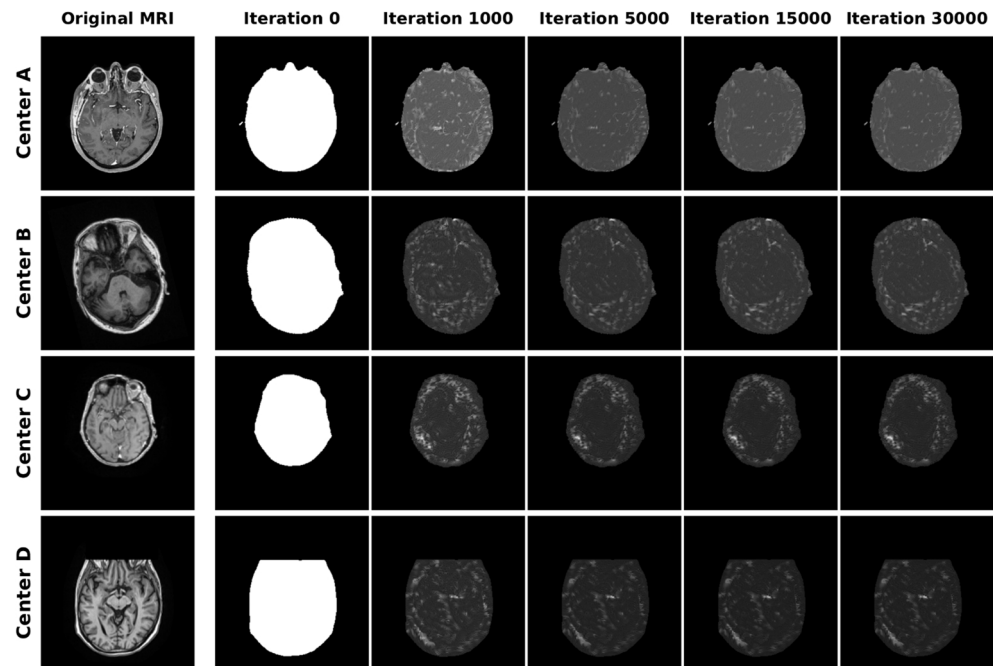
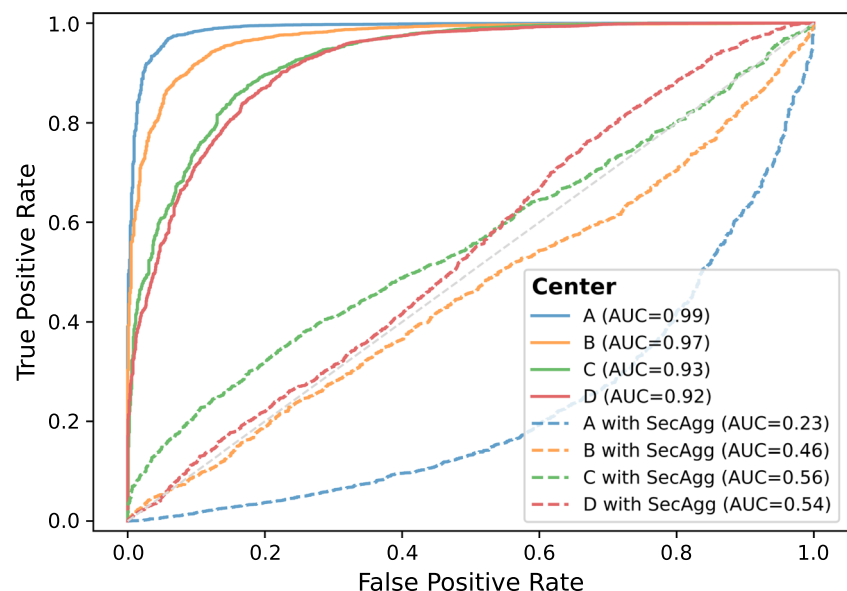


Fig. 3 ROC curves for membership inference attack (FedMIA) across clients, before and after the implementation of SecAgg, evaluated on MRI and CT batches at the final training round. In the original framework, the attack achieved high discrimination at all centers, indicating severe vulnerability to membership inference. The application of SecAgg resulted in a substantial decrease in AUC values, thereby indicating the efficacy of SecAgg in mitigating the attack



scenarios where gradients are aggregated across multiple samples per batch, the reduced granularity of updates is expected to further compromise the DLG's reconstruction performance.

The FedMIA attack was evaluated across all four training centers (A-D). As shown in Fig. 3, using the original aggregation strategy (FedAvg+FedProx) [12], the attack achieved accurate discrimination for all centers, with AUC values between 0.92 and 0.99 on the received model and client updates of the final training round. These results confirm that, within the original framework [12], the server could reliably identify whether a specific patient sample might

have been used in the training of a particular client, thereby highlighting a substantial privacy vulnerability.

The implementation of SecAgg substantially reduced the attack efficacy, as evidenced by the AUC values, which decreased to 0.23, 0.46, 0.56, and 0.54 for Centers A, B, C, and D, respectively (see Fig. 3). Notably, this reduction held even under a best-case attacker setting, where access to the original MRI and CT data provided an upper bound on attack effectiveness. Furthermore, the impact of SecAgg on model performance, as presented in Table 2, was found to be negligible. The MAE, SSIM, and PSNR values exhibited consistency across all centers before and after the implementation of SecAgg, with variations occurring within the

Table 2 Model performance across centers before and after applying Secure Aggregation (SecAgg) as a defense mechanism against Federated Membership Inference Attacks (FedMIA). Results are reported as mean and standard deviation

Center	SecAgg	MAE (HU)	SSIM	PSNR (dB)
A	X	92.9 ± 2.8	0.96 ± 0.01	33.4 ± 0.6
B	✓	87.0 ± 2.1	0.96 ± 0.01	33.7 ± 0.5
	X	122.9 ± 6.3	0.84 ± 0.02	21.2 ± 1.2
C	✓	121.2 ± 6.6	0.84 ± 0.02	21.3 ± 1.3
	X	96.0 ± 6.3	0.81 ± 0.03	21.6 ± 1.4
D	✓	94.3 ± 9.5	0.81 ± 0.03	21.6 ± 1.4
	X	68.5 ± 12.0	0.92 ± 0.01	26.5 ± 1.0
E	✓	68.9 ± 10.1	0.92 ± 0.01	26.5 ± 1.0
	X	102.4 ± 12.5	0.88 ± 0.03	26.1 ± 1.7
	✓	100.6 ± 10.6	0.88 ± 0.03	26.1 ± 1.7

reported results of the original framework baseline [12]. Overall, these results indicate that SecAgg provides robust privacy protection against FedMIA without compromising synthesis performance.

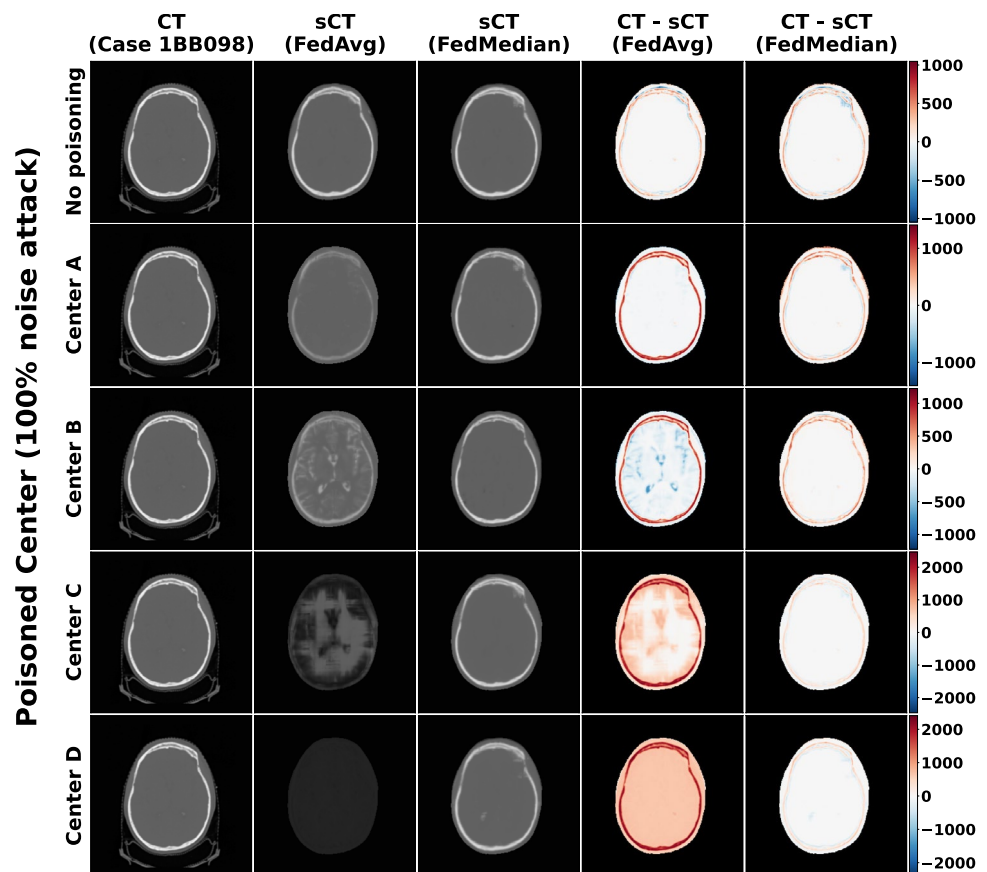
As detailed in Section “Data Poisoning”, the data poisoning attack was simulated by corrupting the training data of one client at a time (Centers A–D) at three different noise levels ($\gamma \in \{0.1, 0.5, 1.0\}$). The impact was then evaluated on all clients and on the external Center E. The results obtained under all scenarios are reported in Table 3 in terms of MAE, considered the most sensitive similarity metric in this context [12].

Under FedAvg, the application of high-noise poisoning ($\gamma = 1.0 - 100\%$) led to severe degradation of the model across all non-poisoned centers, thereby impeding the effective operability of the federation. Notably, the most severe

Table 3 Quantitative evaluation of the impact of data poisoning attacks on federated sCT synthesis. MAE [HU] is reported for each center under three noise levels ($\gamma \in \{0.1, 0.5, 1.0\}$) – indicating 10%, 50% and 100% of noise injected respectively – comparing FedAvg (top) and FedMedian (bottom) aggregation strategies. The no-poisoning baseline is reported in the last row. Bold values indicate the better-performing aggregation strategy for each center–noise

FedAvg		Evaluated Centers – MAE [HU]				
Poisoned Center	Noise	A	B	C	D	E (External)
A	100%	–	208.2 ± 22.2	215.4 ± 21.8	210.6 ± 38.6	192.3 ± 16.9
	50%	–	140.4 ± 13.3	106.5 ± 16.6	92.7 ± 17.7	121.6 ± 12.8
	10%	–	124.2 ± 7.1	94.3 ± 13.3	70.5 ± 9.1	110.9 ± 11.1
B	100%	183.3 ± 11.1	–	236.0 ± 24.1	248.5 ± 25.7	222.8 ± 14.5
	50%	92.8 ± 5.5	–	99.9 ± 18.8	84.0 ± 16.5	117.3 ± 12.4
	10%	92.1 ± 1.8	–	93.9 ± 14.6	72.9 ± 12.3	103.5 ± 9.9
C	100%	504.4 ± 28.3	841.6 ± 38.1	–	682.8 ± 57.3	623.7 ± 50.8
	50%	138.8 ± 15.0	165.0 ± 17.4	–	143.7 ± 26.7	180.4 ± 18.1
	10%	95.3 ± 2.4	131.5 ± 9.6	–	79.1 ± 12.2	119.0 ± 11.5
D	100%	764.9 ± 36.9	790.0 ± 36.8	740.3 ± 20.7	–	787.0 ± 26.0
	50%	186.3 ± 19.9	228.7 ± 25.7	181.9 ± 24.2	–	207.0 ± 19.3
	10%	94.9 ± 1.7	134.5 ± 8.0	98.8 ± 11.7	–	120.5 ± 11.1
No poisoning	–	87.0 ± 1.7	118.8 ± 7.9	95.0 ± 9.1	69.4 ± 8.3	107.1 ± 11.2
FedMedian		Evaluated Centers – MAE [HU]				
Poisoned Center	Noise	A	B	C	D	E (External)
A	100%	–	124.1 ± 12.6	104.1 ± 9.4	91.5 ± 12.2	115.2 ± 11.1
	50%	–	117.1 ± 9.7	97.3 ± 16.9	80.7 ± 12.4	115.5 ± 12.1
	10%	–	115.3 ± 7.5	97.1 ± 11.4	75.6 ± 10.5	110.6 ± 11.5
B	100%	92.9 ± 3.6	–	105.2 ± 12.2	82.2 ± 10.0	111.4 ± 10.0
	50%	87.6 ± 0.8	–	95.4 ± 12.7	72.0 ± 8.2	114.7 ± 11.7
	10%	88.3 ± 2.6	–	96.3 ± 11.9	74.7 ± 9.0	108.2 ± 10.5
C	100%	88.7 ± 9.8	118.4 ± 14.6	–	99.0 ± 12.0	124.3 ± 11.0
	50%	83.9 ± 1.5	109.8 ± 10.5	–	83.8 ± 14.6	119.3 ± 12.8
	10%	85.3 ± 0.2	110.4 ± 6.9	–	75.1 ± 9.9	108.9 ± 11.3
D	100%	85.0 ± 4.1	122.5 ± 14.1	126.6 ± 21.0	–	130.6 ± 10.4
	50%	83.1 ± 0.9	109.6 ± 8.6	96.3 ± 13.3	–	113.6 ± 12.0
	10%	86.9 ± 1.8	113.3 ± 10.1	94.7 ± 16.3	–	108.5 ± 10.2
No poisoning	–	89.2 ± 4.9	108.6 ± 5.3	98.1 ± 10.1	74.1 ± 9.5	108.6 ± 11.2

Fig. 4 Visual comparison of sCT for a representative patient from external Center E under a 100% noise data poisoning attack ($\gamma = 1.0$), corresponding to the worst-case scenario. The first row represents the no-poisoning baseline, while the other rows correspond to a different poisoned training center (Centers A-D). Columns show, from left to right, the reference CT, the sCT generated with FedAvg and FedMedian aggregation, and the corresponding pixel-wise difference maps (CT - sCT). FedMedian consistently restored image quality across all poisoning configurations, while FedAvg produced severely corrupted and invalid outputs



degradation was observed when Center C was poisoned, where FedAvg yielded MAE values of 504 HU, 841 HU, and 682 HU for Centers A, B, and D, respectively. These outcomes effectively render the model technically and clinically inoperative (see Fig. 4). A similar impact was observed when Center D was poisoned at the same noise level, with MAE values exceeding 740 HU across all non-poisoned centers. This greater impact can be attributed to the larger amount of data provided by Center C (21 patients) and Center D (29 patients) [12]. At intermediate noise levels ($\gamma = 0.5 - 50\%$), performance degradation remained critical, with MAE reaching high values (i.e., 229 HU) when Center D was poisoned. In conditions of low noise ($\gamma = 0.1 - 10\%$) poisoning, the operability of the federation was preserved with MAE values tending to the no-poisoning baseline. However, more substantial degradations were observed in specific center combinations, thereby confirming the impact of data poisoning also at low noise levels. The FedMedian aggregation strategy consistently mitigated or severely limited the attack under all poisoning scenarios at $\gamma = 1.0$ and $\gamma = 0.5$, often reducing MAE to the no-poisoning baseline across all center pairs (see Table 3). For instance, under a 100% noise injection on Center D, FedMedian achieved a substantial MAE reduction ($> 80\%$) across all centers compared with FedAvg. However,

a residual degradation persisted on Center C, where the MAE (≈ 127 HU) remained about 30 HU higher than the no-poisoning baseline, suggesting a potentially non-negligible impact in clinical settings. At 10% noise injection, the advantage of FedMedian over FedAvg was less consistent, and in several configurations FedAvg achieved equivalent MAE. This finding suggests that a 10% noise injection may not be sufficiently distinguishable from the natural heterogeneity of the data distribution across centers, which can originate from differences in scanner resolution or intrinsic acquisition noise. Notably, this observation aligns with a poisoning attack strategy, under which malicious clients inject deliberately low levels of noise to evade detection while still cumulatively degrading the global model performance. In a clinical context, particularly with regard to radiotherapy dose calculation workflows, these subtle and systematic distortions may potentially propagate directly into the sCT-based dose computation pipeline. Such distortions may not be captured by routine quality assurance checks that primarily assess gross anatomical plausibility rather than voxel-wise HU accuracy. In contrast to tasks where model corruption results in immediately observable categorical errors, sCT synthesis models may instead manifest spatially coherent but quantitatively biased HU estimates. This observation highlights a notable limitation of

aggregation strategies that are oriented towards robustness. While such strategies may prove effective in mitigating high-intensity attacks, they can remain partially vulnerable to low-amplitude, stealthy poisoning strategies.

Notably, the external Center E followed the same trend as the clients, demonstrating that the impact of poisoning extended beyond the federated training clients and that FedMedian effectively limited it in the external evaluation setting as well. Nevertheless, under FedMedian aggregation, the external Center E remained substantially affected under the 100% and 50% scenario, with MAE values ranging from 111 HU to 130 HU depending on the poisoned client, representing a consistent degradation with respect to the no-poisoning baseline of 108 HU. Indeed, statistically significant degradation (p -value < 0.05) was observed in most of these cases, with the exception of Center B poisoned at 100% (p -value > 0.05). Furthermore, a qualitative evaluation of the impact of data poisoning is presented in Fig. 4, showing the central-axial sCT slice of a representative patient from the external Center E under the worst-case attack scenario ($\gamma = 1.0$, i.e., 100% noise injection). This enables a direct comparison across poisoned training centers (Centers A–D) and the no-poisoning baseline. In detail, while poisoning Center A or Center B resulted in a degraded output, poisoning Centers C or D led to complete loss of structural information under FedAvg. In contrast, FedMedian demonstrated a consistent recovery of sCT in all poisoning scenarios, although some artifacts were observed (see FedMedian sCT - Poisoned Center D in Fig. 4).

However, the Differential Privacy defense mechanism was not investigated in this study. Although DP provides formal probabilistic privacy guarantees [32], its noise-injection mechanism introduces a performance-privacy trade-off [44] that poses a particular challenge for sCT synthesis, where voxel-level HU accuracy directly conditions downstream dosimetric calculations. Consequently, SecAgg was identified as the optimal solution for defending against the FedMIA threat model at the communication level. Nevertheless, additional investigation of DP for federated sCT synthesis may be a viable direction for future research.

Conclusion

This study evaluated the vulnerability of a federated MRI-to-sCT translation framework to three representative adversarial attack classes –gradient inversion, membership inference, and data poisoning– and assessed the efficacy of corresponding defense mechanisms. The DLG attack recovered only coarse anatomical structure without clinically identifiable detail, suggesting that the combination of architectural choices and federated training characteristics,

together with the complexity of the I2I task, may constitute a substantial barrier to pixel-accurate gradient inversion.

Nevertheless, the gradient inversion was performed in 2D, at the same granularity as the information exposed during training and aggregation. A comprehensive 3D formulation of gradient inversion under 3D (or patch-3D) training settings could represent a noteworthy direction for future investigations of privacy leakage in volumetric federated medical imaging workflows.

The FedMIA attack demonstrated near-perfect membership discrimination in the baseline framework, confirming that federated I2I translation models are not inherently secure. However, the application of SecAgg reduced AUC values to random levels across all centers without any substantial impact on synthesis quality. Data poisoning under original settings (FedAvg and FedProx) [12] caused severe synthesis degradation at high noise levels, while FedMedian consistently restored model performance close to the no-poisoning baseline, at the cost of some artifacts observed in the sCTs of the external Center E. Collectively, these findings demonstrate that federated I2I translation frameworks, as well as other FL applications, require explicit security evaluation. The application of well-established strategies (such as SecAgg and FedMedian) represent viable, complementary defenses against privacy and integrity threats. However, results also demonstrated that low levels of noise (i.e., 10%) injected by a malicious client could be difficult to detect and mitigate. Although Byzantine-robust aggregation methods, such as Krum [45], represent a promising candidate for future investigation, their practical implementation necessitates a minimum number of participating clients that exceeds the federation size examined in this study. Future work should therefore explore detection mechanisms compatible with small-scale federations, such as cosine similarity-based gradient screening [46]. Furthermore, although no measurable computational or communication overhead was observed for SecAgg at the federation scale considered in this study, the cost of cryptographic secure aggregation protocols is known to scale with the number of participating clients [33, 42]. Therefore, future studies involving larger federations should explicitly profile these overheads.

In light of these findings, the adoption of encryption protocols and additional infrastructure-level security measures (i.e., strong authentication protocols) appears to be a necessary complement to algorithmic defenses. While FL is regarded as a privacy-preserving and secure solution, its resilience is contingent upon the absence of dedicated countermeasures. As evidenced by the FedMIA results, the absence of SecAgg in a baseline federated framework renders FL critically susceptible to membership inference attacks under the right conditions. This highlights the necessity of addressing security in FL at multiple levels, including

algorithmic, cryptographic, and infrastructural aspects. Further studies should directly incorporate a security analysis when evaluating federated frameworks for medical I2I translation. When also combined with clinical studies, this would allow for a comprehensive assessment of the clinical implications of both attacks and defenses. Notably, this study was conducted with a limited number of clients. While this configuration is consistent with prior FL studies in the field of I2I translation, its application to large-scale clinical deployments is not fully reflected. Therefore, the observed security behavior in this work should be interpreted in the context of a small-scale federation. Future work should investigate larger federations with varying client numbers and controlled heterogeneity levels to better characterize the stability of attack and defense mechanisms under large-scale deployment conditions.

Acknowledgements The authors would like to thank Prof. Dr. J. Seco (Heidelberg University, German Cancer Research Center) for his support of this work and Dr. C. Catana (Massachusetts General Hospital) for supporting data sharing. The authors also acknowledge M. K. Zabaleta and N. Skupien for their contributions to the development of the original framework on which the present study builds.

Author Contributions Conceptualization: CBR, LB; Data curation: CBR, OB, FC, PZ; Formal analysis: CBR, LB, MFS; Investigation: CBR, LB; Methodology: CBR, LB; Project administration: CBR; Software: CBR, LB; Validation: CBR, LB; Visualization: CBR, LB, MFS; Writing - original draft: CBR, LB; Writing - review & editing: CBR, LB, OB, FC, PZ, MFS.

Funding Open Access funding enabled and organized by Projekt DEAL. This research received no external funding.

Data Availability Restrictions apply to the availability of the data supporting the findings of this study from Centers A, B and C which were used under licence for this study and are therefore not publicly available. The data from Center D and Center E were extracted from the public SynthRAD2023 Grand Challenge dataset and are available at <https://doi.org/10.5281/zenodo.7260705>.

Declarations

Competing Interests The authors declare no competing interests.

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not

included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Wang, J., Wang, S., and Zhang, Y., Deep learning on medical image analysis. *CAAI Trans. Intell. Technol.* 10(1):1–35, 2025. <https://doi.org/10.1049/cit2.12356>
2. Dayarathna, S., Islam, K.T., Uribe, S., Yang, G., Hayat, M., and Chen, Z., Deep learning based synthesis of mri, ct and pet: Review and analysis. *Med. Image Anal.* 92:103046, 2024. <https://doi.org/10.1016/j.media.2023.103046>
3. Spadea, M.F., Maspero, M., Zaffino, P., and Seco, J., Deep learning based synthetic-ct generation in radiotherapy and pet: A review. *Med. Phys.* 48(11):6537–6566, 2021. <https://doi.org/10.1002/mp.15150>
4. Texier, B., Hémon, C., Lekieffre, P., Collot, E., Tahri, S., Chourak, H., Dowling, J., Greer, P., Bessieres, I., Acosta, O., Boue-Rafle, A., Guevelou, J.L., de Crevoisier, R., Lafond, C., Castelli, J., Barateau, A., Nunes, J.-C.: Computed tomography synthesis from magnetic resonance imaging using cycle Generative Adversarial Networks with multicenter learning. *Phys. Imaging Radiat. Oncol.* 28:100511, 2023. <https://doi.org/10.1016/j.phro.2023.100511>
5. Suleman, M.U., Mursaleen, M., Khalil, U., Saboor, A., Bilal, M., Khan, S.A., Subhani, M.A., Hussnain, M.A., Tabassum, S.N., and Tahir, M., Assessing the generalizability of artificial intelligence in radiology: a systematic review of performance across different clinical settings. *Ann. Med. Surg.* 87(12):8803–8811, 2025. <https://doi.org/10.1097/ms9.0000000000004166>
6. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance), 2016. <http://data.europa.eu/eli/reg/2016/679/oj>
7. Wouters, B., Shaw, D., Sun, C., Ippel, L., Soest, J., Berg, B., Musmann, O., Koster, A., Kallen, C., Oppen, C., Dekker, A., Dumontier, M., and Townend, D., Putting the GDPR into practice: Difficulties and uncertainties experienced in the conduct of big data health research. *Eur. Data Protection Law Rev.* 7(2):206–216, 2021.
8. Mahmood, H., Alamgir, Z., Javed, S.T., Karim, S., and Awais, M., Federated Generative Models in Medical Imaging: Current Advances, Challenges, and Future Directions. *IEEE Access.* 14:5197–5217, 2026. <https://doi.org/10.1109/ACCESS.2026.3650810>
9. Wang, J., Xie, G., Huang, Y., Lyu, J., Zheng, F., Zheng, Y., and Jin, Y., Fedmed-gan: Federated domain translation on unsupervised cross-modality brain image synthesis. *Neurocomputing.* 546:126282, 2023. <https://doi.org/10.1016/j.neucom.2023.126282>
10. Dalmaz, O., Mirza, M.U., Elmas, G., Ozbey, M., Dar, S.U.H., Ceyani, E., Oguz, K.K., Avestimehr, S., and Çukur, T., One model to unite them all: Personalized federated learning of multi-contrast MRI synthesis. *Med. Image Anal.* 94:103121, 2024. <https://doi.org/10.1016/j.media.2024.103121>
11. Fiszer, J., Ciupek, D., Malawski, M., and Pieciak, T., Validation of ten federated learning strategies for multi-contrast image-to-image MRI data synthesis from heterogeneous sources. *bioRxiv.* 2025. <https://doi.org/10.1101/2025.02.09.637305>

12. Raggio, C.B., Zabaleta, M.K., Skupien, N., Blanck, O., Cicone, F., Cascini, G.L., Zaffino, P., Migliorelli, L., and Spadea, M.F., FedSynthCT-Brain: A federated learning framework for multi-institutional brain MRI-to-CT synthesis. *Comput. Biol. Med.* 192:110160, 2025. <https://doi.org/10.1016/j.combiomed.2025.110160>
13. Raggio, C.B., Zaffino, P., and Spadea, M.F., A privacy-preserving federated learning framework for generalizable cbct to synthetic ct translation in head and neck. *Frontiers in Digital Health*. Volume 8 - 2026, 2026. <https://doi.org/10.3389/fdgh.2026.1812254>
14. Ooijen, P.M.A., Darzi, E., and Dekker, A., In: De Cecco, C.N., Assen, M., and Leiner, T. (eds.) *Data Storage, Cloud Usage and Artificial Intelligence Pipeline*, pp. 45–55. Springer; Cham, 2022. https://doi.org/10.1007/978-3-030-92087-6_5
15. Hu, K., Gong, S., Zhang, Q., Seng, C., Xia, M., and Jiang, S., An overview of implementing security and privacy in federated learning. *Artif. Intell. Rev.* 57(8), (2024). <https://doi.org/10.1007/s10462-024-10846-8>
16. Feng, Y., Guo, Y., Hou, Y., Wu, Y., Lao, M., Yu, T., and Liu, G., A survey of security threats in federated learning. *Complex Intell. Syst.* 11(2), 2025. <https://doi.org/10.1007/s40747-024-01664-0>
17. Dong, J., Chen, J., Xie, X., Lai, J., and Chen, H., Survey on adversarial attack and defense for medical image analysis: Methods and challenges. *ACM Comput. Surv.* 57(3), 2024. <https://doi.org/10.1145/3702638>
18. Jin, R., and Li, X., Backdoor attack and defense in federated generative adversarial network-based medical image synthesis. *Med. Image Anal.* 90:102965, 2023. <https://doi.org/10.1016/j.media.2023.102965>
19. Kumar, K.N., Mohan, C.K., Cenkeramaddi, L.R., and Awasthi, N., Minimal data poisoning attack in federated learning for medical image classification: An attacker perspective. *Artif. Intell. Med.* 159:103024, 2025. <https://doi.org/10.1016/j.artmed.2024.103024>
20. Darzi, E., Dubost, F., Sijtsma, N.M., and Ooijen, P.M.A., Exploring adversarial attacks in federated learning for medical imaging. *IEEE Trans. Inf. Inf.* 20(12):13591–13599, 2024. <https://doi.org/10.1109/TII.2024.3423457>
21. Nasr, M., Shokri, R., and Houmansadr, A., Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In: *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 739–753, 2019. <https://doi.org/10.1109/SP.2019.00065>
22. Zhu, G., Li, D., Gu, H., Yao, Y., Fan, L., and Han, Y., FedMIA: An Effective Membership Inference Attack Exploiting "All for One" Principle in Federated Learning. In: *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20643–20653. IEEE Computer Society, Los Alamitos, CA, USA, 2025. <https://doi.org/10.1109/CVPR52734.2025.01922>
23. Hu, H., Salicic, Z., Sun, L., Dobbie, G., Yu, P.S., and Zhang, X., Membership Inference Attacks on Machine Learning: A Survey. *ACM Comput. Surv.* 54(11s):1–37, 2022. <https://doi.org/10.1145/3523273>
24. Li, H., Li, C., Wang, J., Yang, A., Ma, Z., Zhang, Z., and Hua, D., Review on security of federated learning and its application in healthcare. *Future Gener. Comput. Syst.* 144:271–290, 2023. <https://doi.org/10.1016/j.future.2023.02.021>
25. Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., and Shmatikov, V., How to backdoor federated learning. In: Chiappa, S., and Calandra, R. (eds.) *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research*, vol. 108, pp. 2938–2948. PMLR, 2020. <https://proceedings.mlr.press/v108/bagdasaryan20a.html>
26. Xie, C., Huang, K., Chen, P.-Y., and Li, B., Dba: Distributed backdoor attacks against federated learning. In: *8th International Conference on Learning Representations, ICLR 2020*, 2020. <https://openreview.net/forum?id=rkgyS0VFvr>
27. Zhang, H., Jia, J., Chen, J., Lin, L., and Wu, D., A3fl: adversarially adaptive backdoor attacks to federated learning. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23*. Curran Associates Inc., Red Hook, NY; USA, 2023
28. Zhu, L., Liu, Z., and Han, S., Deep leakage from gradients. Curran Associates Inc., Red Hook, NY; USA, 2019.
29. Huang, Y., Gupta, S., Song, Z., Li, K., and Arora, S., Evaluating gradient inversion attacks and defenses in federated learning. In: *Proceedings of the 35th International Conference on Neural Information Processing Systems. NIPS '21*. Curran Associates Inc., Red Hook, NY; USA, 2021.
30. Hatamizadeh, A., Yin, H., Molchanov, P., Myronenko, A., Li, W., Dogra, P., Feng, A., Flores, M.G., Kautz, J., and Xu, D., et al., Do gradient inversion attacks make federated learning unsafe? *IEEE Trans. Medical Imaging*. 42(7):2044–2056, 2023.
31. Wang, Z., Gan, D., Fang, W., Zhu, Y., and Liu, K., Gradinvdiff: Stealing medical privacy in federated learning via diffusion-based gradient inversion. In: Gee, J.C., Alexander, D.C., Hong, J., Iglesias, J.E., Sudre, C.H., Venkataraman, A., Golland, P., Kim, J.H., and Park, J. (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pp. 262–272. Springer, Cham, 2026.
32. Wei, K., Li, J., Ding, M., Ma, C., Yang, H.H., Farokhi, F., Jin, S., Quek, T.Q.S., and Vincent Poor, H., Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Trans. Inf. Forensics Secur.* 15:3454–3469, 2020. <https://doi.org/10.1109/TIFS.2020.2988575>
33. Bonawitz, K.A., Ivanov, V., Kreuter, B., Marcedone, A., McMahian, H.B., Patel, S., Ramage, D., Segal, A., and Seth, K., Practical secure aggregation for federated learning on user-held data. CoRR. [arxiv:abs/1611.04482](https://arxiv.org/abs/1611.04482), 2016.
34. Gong, X., Chen, Y., Wang, Q., and Kong, W., Backdoor attacks and defenses in federated learning: State-of-the-art, taxonomy, and future directions. *IEEE Wirel. Commun.* 30(2):114–121, 2023. <https://doi.org/10.1109/MWC.017.2100714>
35. Thummerer, A., Bijl, E., Galapon Jr, A., Verhoeff, J.J.C., Langendijk, J.A., Both, S., Berg, C.N.A.T., and Maspero, M., Synthrad2023 grand challenge dataset: Generating synthetic ct for radiotherapy. *Med. Phys.* 50(7):4664–4674, 2023. <https://doi.org/10.1002/mp.16529>
36. Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., and Gee, J.C., N4itk: Improved n3 bias correction. *IEEE Trans. Med. Imaging*. 29(6):1310–1320, 2010. <https://doi.org/10.1109/TMI.2010.2046908>
37. Beutel, D.J., Topal, T., Mathur, A., Qiu, X., Fernandez-Marques, J., Gao, Y., Sani, L., Li, K.H., Parcollet, T., Gusmão, P.P.B., and Lane, N.D., Flower: a friendly federated learning research framework. [arxiv:2007.14390](https://arxiv.org/abs/2007.14390) 2022.
38. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S., PyTorch: an imperative style, high-performance deep learning library. [arxiv:1912.01703](https://arxiv.org/abs/1912.01703) 2019
39. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., Nath, V., He, Y., Xu, Z., Hatamizadeh, A., Myronenko, A., Zhu, W., Liu, Y., Zheng, M., Tang, Y., Yang, I., Zephyr, M., Hashemian, B., Alle, S., Darestani, M.Z., Budd, C., Modat, M., Vercauteren, T., Wang, G., Li, Y., Hu, Y., Fu, Y., Gorman, B., Johnson, H., Genereaux, B., Erdal, B.S., Gupta, V., Diaz-Pinto, A., Dourson, A., Maier-Hein, L., Jaeger, P.F., Baumgartner, M., Kalpathy-Cramer, J., Flores, M., Kirby, J., Cooper, L.A.D., Roth, H.R., Xu, D., Bericat, D., Floca, R., Zhou, S.K., Shuaib, H., Farahani, K., Maier-Hein, K.H., Aylward, S., Dogra, P., Ourselin, S., and Feng, A., MONAI:

- An open-source framework for deep learning in healthcare, 2022. [arxiv:2211.02701](https://arxiv.org/abs/2211.02701)
40. Sahu, A.K., Li, T., Sanjabi, M., Zaheer, M., Talwalkar, A., and Smith, V., On the convergence of federated optimization in heterogeneous networks. CoRR. [arXiv:1812.06127](https://arxiv.org/abs/1812.06127) (2018)
 41. Li, Y., Li, W., Xiong, J., Xia, J., and Xie, Y., Comparison of supervised and unsupervised deep learning methods for medical image synthesis between computed tomography and magnetic resonance images. *BioMed Res. Int.* 2020(1):5193707, 2020. <https://doi.org/10.1155/2020/5193707>
 42. Bell, J.H., Bonawitz, K.A., Gascón, A., Lepoint, T., and Raykova, M., Secure single-server aggregation with (poly)logarithmic overhead. In: *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security. CCS '20*, pp. 1253–1269. Association for Computing Machinery, New York, NY, USA, 2020. <https://doi.org/10.1145/3372297.3417885>.
 43. Yin, D., Chen, Y., Kannan, R., and Bartlett, P., Byzantine-robust distributed learning: Towards optimal statistical rates. In: *Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research*, vol. 80, pp. 5650–5659. PMLR, 2018. <https://proceedings.mlr.press/v80/yin18a.html>
 44. Darzi, E., Sijtsema, N.M., and Ooijen, P., Weight-space noise for privacy-robustness trade-offs in federated learning. *Neural Comput. Appl.* 37(24):19687–19705, 2025. <https://doi.org/10.1007/s00521-025-11420-1>
 45. Blanchard, P., El Mhamdi, E.M., Guerraoui, R., and Stainer, J., Machine learning with adversaries: byzantine tolerant gradient descent. In *Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17*, pp. 118–128. Curran Associates Inc., Red Hook, NY, USA, 2017.
 46. Samy, A.E., and Girdzijauskas, Š., Mitigating sybil attacks in federated learning. In: Meng, W., Yan, Z., Piuri, V. (eds.) *Information Security Practice and Experience*, pp. 36–51. Springer, Singapore, 2023. https://doi.org/10.1007/978-981-99-7032-2_3

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Robust Neurocomputing Model in Efficient Federated Brain Tumour Segmentation with Sparsification and Weights Clustering

8.1 Introduction

Brain tumour segmentation constitutes one of the most clinically relevant and methodologically demanding applications of Artificial Intelligence (AI) in neuroimaging. Accurate delineation of tumour subregions from multi-parametric Magnetic Resonance Imaging (MRI) is essential for diagnosis, surgical planning, radiotherapy target definition, and monitoring disease progression over time. Indeed, among imaging modalities, MRI remains the reference standard for brain tumour assessment due to its superior soft-tissue contrast and non-invasive nature [132–134]. In recent years, Deep Learning (DL) methods substantially advanced the state of the art in automated tumour segmentation, with U-Net-based architectures [65] and their volumetric extensions becoming dominant paradigms in benchmark challenges such as Brain Tumor Segmentation (BraTS) [135].

Despite these advances, the development of robust DL models for neuro-oncological segmentation remains constrained by limited access to large, diverse, and well-annotated datasets. Medical imaging data are typically distributed across institutions and are subject to strict regulatory frameworks, such as the General Data Protection Regulation (GDPR), which restrict centralised data aggregation [136]. Federated Learning (FL) has emerged as a promising paradigm to address this limitation by enabling collaborative model training without sharing raw patient data [92, 99, 100, 137]. In this decentralised setting, local clients perform on-site optimisation and share model updates with a coordinating server, which aggregates them into a global model. Although FL mitigates data governance barriers, its deployment in

3D medical imaging introduces serious technical challenges. Volumetric models, such as 3D U-Nets, contain a large number of parameters, and the repeated transmission of full-precision weight updates during iterative aggregation rounds results in significant communication overhead [91, 138]. This issue is particularly critical in realistic hospital networks, where bandwidth may be limited and computational infrastructures may be heterogeneous. Traditional aggregation strategies such as FedAvg [92] are not inherently communication-aware and may therefore hinder the scalability of large-scale federated neuroimaging studies.

To address this limitation, an increasing number of studies investigated gradient compression techniques, including sparsification and quantization [139–141]. However, aggressive compression can degrade convergence stability and segmentation accuracy, particularly in non-independent and identically distributed (non-IID) scenarios, which characterise multi-centre clinical datasets [142]. In this context, Onaizah et al. [143] introduced *FedWSOComp*, an integrated framework combining weight magnitude-based sparsification, deterministic k -means quantization, and entropy-based encoding. The method demonstrated substantial communication savings while preserving model fidelity in brain tumour segmentation.

Nevertheless, the original evaluation of FedWSOComp primarily focused on scalability with respect to the number of participating clients and adopted fixed compression parameters. A systematic investigation into the relationship between sparsification rates, quantization granularity, and data heterogeneity, particularly in 3D volumetric segmentation, has not yet been conducted. Understanding these interactions is essential, as segmentation tasks are highly sensitive to boundary inaccuracies, and even minor degradations in spatial precision may compromise clinical reliability.

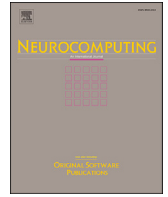
Building upon these considerations, the study presented in this Chapter provides a comprehensive analysis of FedWSOComp in the context of 3D brain tumour segmentation using the multi-institutional FeTS 2024 dataset [144]. Both IID and non-IID client distributions were explicitly simulated to reflect clinical FL scenarios. A 3D U-Net architecture was adopted in order to exploit volumetric contextual information, and 18 distinct parameters configurations were evaluated by varying sparsification retention and clustering granularity. Furthermore, a systematic ablation study distinguished the separate contributions of the clustering and sparsification components, and a dedicated communication analysis calculated the reduction in payload relative to the FedAvg baseline.

The experimental results demonstrated that aggressive compression severely degraded segmentation quality. In contrast, higher retention ratios combined with moderate-to-fine quantization achieved a stable balance between Dice Similarity Coefficient (DSC) and the 95th percentile of the Hausdorff distance (HD95) while substantially reducing communication cost. Importantly, the analysis revealed that the primary advantage of FedWSOComp arises from the synergistic interaction of its components rather than from sparsification or clustering in isolation. These findings provide quantitative guidance for the configuration of

communication-aware FL pipelines in clinical environments where bandwidth is constrained.

The content of this chapter has been adapted from a manuscript previously published in the peer-reviewed journal “Neurocomputing”. Portions of the text are reproduced verbatim from the published article [145]. The author of this thesis contributed to the methodological design, experimental implementation, and analytical evaluation of the extended FedWSOComp framework, including the systematic compression parameter study and the ablation analysis. As co-first author of the publication, the author actively participated in the conceptual development of the study, implementation of the 3D federated segmentation pipeline, statistical analysis of the results, and drafting and revising of the manuscript in close collaboration with the co-authors.

8.2 Manuscript



Towards robust neurocomputing model in efficient federated brain tumour segmentation with sparsification and weights clustering

Asaf Raza^{a,1} , Ciro Benito Raggio^{b,1}, Antonella Guzzo^{a,*} , Maria Francesca Spadea^b, Giancarlo Fortino^a

^a DIMES—Department of Informatics, Modeling, Electronics and Systems, University of Calabria, Via P. Bucci, Rende, 87036, Cosenza, Italy

^b Institute of Biomedical Engineering, Karlsruhe Institute of Technology, Fritz-Haber-Weg 1, Karlsruhe, 76131, Baden-Württemberg, Germany

HIGHLIGHTS

- Analyse the impact of FedWSOComp, an integrated strategy combining top-k sparsification, quantization, and entropy-based encoding.
- The performance of 18 different configurations (including IID and non-IID) was evaluated systematically.
- High retention (60%) with fine quantization (64 clusters) optimizes performance.

ARTICLE INFO

Communicated by N. Zeng

Keywords:

Brain tumour segmentation
Communication-efficient federated learning
Medical imaging
Quantization

ABSTRACT

Brain tumour segmentation is a key application of AI in neuroimaging. Recently, federated learning (FL) has emerged as a strategic and increasingly relevant paradigm in neural computing due to its ability to address key challenges in large-scale neural network training, such as data access, privacy, collaborative learning, and model robustness. However, its adoption is currently hindered by high communication costs and the heterogeneity of client data. In this study, we investigated an efficient FL framework for brain tumour segmentation based on communication-aware optimization. We evaluated FedWSOComp, which integrates sparsification, quantization, and entropy-based encoding, in combination with a 3D U-Net architecture under both homogeneous and heterogeneous data distributions. The multi-institutional FeTS 2024 dataset was employed and partitioned into independent and identically distributed (IID) and non-IID settings, with an independent test set of 67 patients. An overall of 18 configurations combined sparsification rates and quantization levels. Performance was measured using Dice Similarity Coefficient (DSC) and 95th percentile Hausdorff Distance (HD95). Experimental results demonstrated that aggressive compression caused severe degradation in segmentation quality, with HD95 exceeding 60 mm. In contrast, higher retention with finer quantization achieved the best balance between efficiency and accuracy, reaching a DSC of 0.98 ± 0.09 and HD95 of 10.40 ± 15.54 mm on the test set under non-IID conditions. The findings demonstrated that, when configured with moderate-to-fine quantization and high sparsification retention, FedWSOComp enabled accurate and communication-efficient federated brain tumour segmentation. This study provides quantitative evidence and practical guidance for the deployment of FL-based segmentation models in privacy-sensitive and bandwidth-constrained clinical settings.

1. Introduction

Brain tumour segmentation plays a vital role in identifying and delineating abnormal regions within brain imaging data, enabling accurate diagnosis and treatment planning and it is a major use case in AI for neuroimaging. Among the primary imaging modalities employed, computed

tomography (CT) and magnetic resonance imaging (MRI) are widely used to detect and segment tumours [1]. However, MRI is generally preferred due to its non-invasive nature [2], and it remains one of the most fundamental and essential tools in the diagnosis and segmentation of brain tumours [3]. In recent years, the tumour segmentation field

* Corresponding author.

Email address: antonella.guzzo@unical.it (A. Guzzo).

¹ These authors contributed equally to this work.

<https://doi.org/10.1016/j.neucom.2026.133142>

Received 15 December 2025; Received in revised form 9 February 2026; Accepted 20 February 2026

Available online 24 February 2026

0925-2312/© 2026 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

has made substantial progress through the application of deep learning (DL), which offers the potential to automate tumour segmentation with high accuracy [4]. The U-Net architecture, originally proposed by Ronneberger et al. [5], has become a standard in DL-based segmentation tasks. This architecture and its numerous variants continue to serve as the foundation for state-of-the-art models in benchmark challenges, such as the Brain Tumour Segmentation (BraTS) challenge, where it has consistently been employed by top-performing teams. Despite their success and the higher segmentation accuracy, DL models typically require large volumes of annotated data [6], and accessing extensive datasets distributed across multiple institutions remains a significant challenge due to stringent data protection regulations, such as the GDPR and HIPAA [7]. Federated Learning (FL) has been identified as a promising approach for neural computing because it enables the training of complex neural models on distributed and sensitive data without centralizing them. In a typical FL setup, participating clients train a local model on their private data and periodically share model weight updates with a central server. The server aggregates these updates and sends back the updated global model to all clients, enabling iterative improvement [8]. This decentralised approach mitigates data privacy concerns and fosters cross-institutional collaboration in medical imaging research [9–11]. Despite its well-recognized advantages, the application of FL to 3D medical images, such as MRI, poses significant challenges due to the high communication overhead, particularly when dealing with large DL models and resource-constrained clients [10,12]. Indeed, traditional FL aggregation strategies, such as federated averaging (FedAvg) [8], typically transmit full precision model updates during each communication round, resulting in a substantial bandwidth consumption [13]. To mitigate this, lossy compression techniques—including quantization and gradient sparsification—have been explored [14].

However, these methods often suffer from reduced accuracy and limited adaptability to the heterogeneity inherent in 3D volumetric imaging data in practical settings [15]. For instance, top- k gradient sparsification effectively reduces communication costs, it introduces approximation errors that impair convergence, especially at high sparsification levels [16]. Similarly, gradient quantization methods, such as QSGD [17] and TernGrad [18], have the potential to reduce communication overhead but often distort the gradient distributions and fail to optimise the bit efficiency in the quantization, leading to degraded model performance [19]. To address this limitation, Onaizah et al. introduced FedWSOComp [20], an advanced aggregation strategy that combined deterministic k -means clustering for sparse quantization, weight magnitude-based sparsification, and Huffman encoding. This approach significantly reduced communication costs while preserving model performance. In contrast to conventional compression methods, FedWSOComp placed emphasis on near-lossless weight recovery, an essential requirement for FL in clinical applications. However, FedWSOComp has not yet been analysed in terms of the individual effects of its core components, sparsification and clustering, particularly under conditions of data heterogeneity. Analysing the impact of sparsification and clustering in FedWSOComp is essential because these mechanisms directly affect communication efficiency, optimization dynamics, and model robustness. Sparsification reduces communication overhead but may introduce bias and noise in the weight updates, influencing convergence and the quality of learned neural representations. Clustering groups clients with similar data distributions or learning behaviors, mitigating non-independent and identically distributed (IID) effects and improving aggregation stability. Based on the above considerations, studying their combined impact is therefore crucial to understanding the trade-offs between efficiency, accuracy, and scalability in federated neurocomputing systems. In this study, we address this gap by investigating the contributions of these components within FedWSOComp in the context of 3D brain tumour segmentation. We conduct our experiments using the Federated Tumour Segmentation (FeTS) 2024 challenge dataset under both IID and non-IID settings. It is important to highlight that with respect to other FL domains, in 3D brain segmentation, it is crucial

to understand the individual effects of the FedWSOComp components, since brain MRI data is highly heterogeneous across institutions due to differences in scanners, protocols and patient populations. Furthermore, segmentation tasks are extremely sensitive to small errors that can alter anatomical boundaries or clinical interpretations. It is essential to understand whether communication savings from sparsification or robustness from clustering primarily drive performance, in order to ensure that models remain accurate, fair and reliable. This analysis not only strengthens scientific understanding, but is also a prerequisite for the trustworthy deployment of FL in clinical neuroimaging, where even minor performance trade-offs can have significant medical consequences. In summary, the key contributions of this study are as follows:

- We analysed the impact of quantization parameters in the FedWSOComp framework, including sparsification rate and clustering granularity, on model convergence and segmentation performance.
- By quantifying how varying compression levels affect convergence and segmentation quality in realistic heterogeneity scenarios, the study provides insights that extend beyond a single application and are broadly valuable to researchers optimizing federated neurocomputing systems.
- We validated the real-world applicability of FedWSOComp in FL scenarios by assessing the model's generalization ability on an independent test set comprising 67 patients from 3 separate institutions that were not involved in the federation.
- We evaluated the FedWSOComp framework on the recent FeTS 2024 dataset, focusing on 3D brain tumour segmentation under both IID and non-IID settings.

2. Related work

2.1. DL architectures and FL for medical image segmentation

In recent years, DL has become increasingly prominent in 3D medical image segmentation due to its ability to capture spatial continuity and mitigate inter-slice artefacts, which are common in 2D approaches. Nonetheless, building three-dimensional models typically requires substantial computational resources. This often led to the adoption of sub-volume (or patch-based) processing, which may compromise structural consistency and the representation of global features [21,22]. These limitations became even more critical in FL, where hardware constraints and communication costs further restricted the deployment of large models on client devices. To overcome such challenges, a wide range of architectural solutions has been proposed. DenseNet-based segmentation pipelines combined with ensemble learning, for instance, enhanced feature expressiveness and robustness across multimodal MRI data [23]. Similarly, lightweight convolutional modules and Dense U-Net variants offered a favourable balance between segmentation accuracy and inference speed [22]. Attention-integrated frameworks, such as SLCA-UNet, improved border delineation by focusing on salient contextual information [21]. Other hybrid encoder-decoder models sought to improve efficiency while retaining accuracy. Obayya et al. [24], for instance, proposed a nested U-Net with multi-scale skips and adapted loss functions, specifically designed to address class imbalance in low-grade glioma segmentation. Deep supervision and dual-path encoding further strengthened structural preservation at tumour boundaries [22], while attention-gated units facilitated feature fusion across tumour subregions [21].

Related advances in medical segmentation network design further highlighted the importance of global contextual modelling and receptive field enlargement. ELTS-Net extended the 3D U-Net paradigm for liver and tumour segmentation by incorporating dilated convolutions to expand the effective receptive field, together with dual attention mechanisms operating across spatial and channel dimensions. In addition, deep supervision was employed at multiple decoder levels to

constrain intermediate representations towards target regions and improve convergence [25]. Alternative paradigms for image segmentation have also been explored beyond fully supervised convolutional models. In this context, deep reinforcement learning has been applied to medical-related image segmentation by formulating region delineation as a sequential decision-making process guided by task-specific reward functions, demonstrating feasibility on gold immunochromatographic strip images segmentation [26].

All the above methods introduce important advances in improving accuracy of learning architectures but are orthogonal with respect to the communication cost. Parallel efforts focused on adapting 3D U-Nets to federated settings. Distributed training demonstrated the feasibility of privacy-preserving segmentation in FL [27], although the volumetric nature of updates significantly increased communication overhead. To alleviate this, decentralised aggregation protocols such as peer-to-peer FL improved scalability and resilience to system faults, while compression-aware strategies were introduced to manage high-dimensional MRI data [28].

2.2. Communication-efficient FL

Heterogeneity across clients further motivated the design of adaptive solutions. Zhang et al. [6] introduced client-personalised layers, enabling global consistency alongside local adaptability. Dai et al. [29] proposed FedATA, an attention-based transformer framework that weighted client contributions by relevance, thereby enhancing robustness to non-IID data. Privacy-preserving mechanisms also gained attention, with studies exploring encryption-based methods [30] and blockchain-inspired frameworks [31] for secure medical collaborations. Lightweight and efficient convolutional networks were further applied to brain cancer analysis in FL [32]. Complementary approaches included contrastive learning [33], standardised evaluation protocols [34], and the integration of transfer learning into federated pipelines, which proved particularly beneficial in cross-institutional scenarios [35,36]. Despite these advances, communication efficiency remained a central challenge. Compression methods such as sparsification and quantization were explored to reduce transmission costs. Top- k sparsification reduced the exchange of less relevant gradient updates [37,38], while weight quantization decreased bandwidth requirements by mapping parameters to lower-precision formats [39]. Adaptive schemes, including layer-wise quantization (FedLQ [37]) and multi-grained dequantization strategies (FedMQ + [38]), achieved improved trade-offs between accuracy and efficiency. Most existing techniques, however, treated sparsification, quantization, and encoding as separate processes, with limited emphasis on end-to-end reconstruction fidelity. Addressing this gap, FedWSOComp [20] proposed a compression-aware framework that integrated deterministic k -means quantization, soft Top- z sparsification [40], and Huffman encoding. This approach enabled high-fidelity reconstruction while achieving compression levels of up to $\times 1085$, resulting in 99% savings in both bandwidth and energy during communication. Although FedWSOComp demonstrated promising results in brain tumour segmentation, its evaluation primarily focused on scalability with increasing client numbers and fixed clustering and sparsification parameters.

Building on these findings, the present study investigated performance stability under varying heterogeneity, clustering, and sparsification conditions, thereby extending the applicability and usability of the original framework.

3. Materials and methods

3.1. Materials

This study employed the public and multi-institutional FeTS 2024 dataset [41]. This challenge extended the BraTS framework [42] in a federated setting, where images are located at local institutions. This condition is often observed in hospital networks, where privacy and

legal concerns pose challenges to the sharing of raw imaging data. Specifically, it integrates clinically acquired scans originating from the RSNA-ASNR-MICCAI BraTS 2021 challenge, the BraTS 2023 Glioma challenge, as well as additional independent datasets contributed by remote institutions participating in a real-world federated network [41]. As a result, the dataset encompasses substantial variability in acquisition protocols, scanners, patient populations, tumour appearance, and institutional data distributions.

The dataset includes multi-parametric MRI (mpMRI) images of the brain of 1251 different patients, each with a corresponding expert-labelled segmentation mask defining tumour subregions. Each scan consists of 4 modalities: i. T1-weighted, ii. T1-contrast enhanced (T1ce), iii. T2-weighted and iv. T2-FLAIR.

Although the dataset was collected from 33 real institutions, the metadata includes an artificial division into 23 partitions. In this study, the original partitions were used to create 4 major clients (or silos) with more data. This approach reduced the number of clients that needed to be trained while still enabling the simulation of non-uniform local datasets, which are characteristic of clinical FL implementations.

Using these 4 major clients, both homogeneous and heterogeneous partitioning arrangements have been defined. In the IID partitioning, 100 random samples per client were selected from each client. This ensured that each client received an equal number of scans and a roughly similar distribution of tumour characteristics. This scenario represented an optimistic and ideal case of FL in which clients work in balanced and similar data environments and served as a baseline for further analysis. For non-IID partitioning, the major clients were kept unbalanced. This strategy simulated a more challenging and more representative scenario of real-world clinical conditions, in which institutions differed significantly in both the quantity of available patient records and the quality of imaging. For each local client, the 10% of the total cases were used for the local validation set, 10% for the local testing set and remaining 80% for the local training set. Nevertheless, to ensure a clearer evaluation protocol, an independent testing set of 67 patients was used, differing from the original FedWSOComp study. This separation between training and testing data reflected a more realistic deployment setting and aimed to reduce potential evaluation bias. A comprehensive overview of the partitioning strategies employed is provided in Table 1. In addition, data distribution among clients was presented in Fig. 1.

Table 1

Client-level data allocation for IID and Non-IID configurations, with a separate independent test set on the server.

IID			
Client	# Patients	Percentage	Original partitions
Client 1	100	25%	1
Client 2	100	25%	18
Client 3	100	25%	4, 6, 13
Client 4	100	25%	2, 3, 5, 7, 8, 9, 10, 11, 12
Total	400	100.0%	
Non-IID			
Client	# Patients	Percentage	Original partitions
Client 1	511	43%	1
Client 2	382	32%	18
Client 3	151	13%	4, 6, 13, 21
Client 4	140	12%	2, 3, 5, 7, 8, 10, 11, 12, 14, 15, 17, 19, 22, 23
Total	1184	100.0%	
Independent test set			
	# Patients	Percentage	Original partitions
Server	67	100%	9, 16, 20

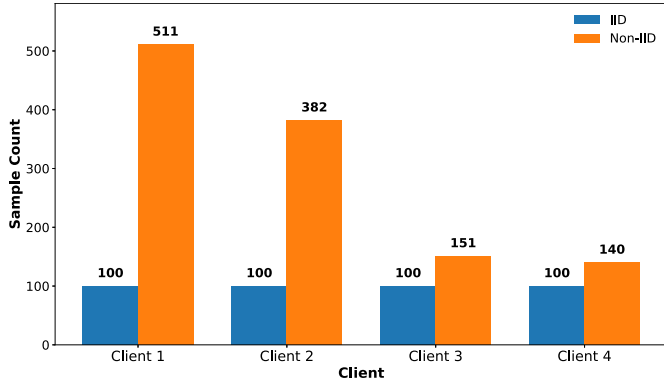


Fig. 1. The number of samples per client both in IID (blue) and non-IID (orange) settings. In an IID settings, clients received an equal number of patient cases. In a non-IID setting, the assignment of mixed data partitions to clients resulted in unbalanced local datasets, thereby simulating a real-world scenario.

3.2. Methods

3.2.1. Data preprocessing

The MONAI [43] framework was employed to implement the medical segmentation workflow, including the data preprocessing pipeline. To standardise spatial orientation across clients, all images were reoriented to the canonical RAS (Right-Anterior-Superior) coordinate system. Furthermore, images and labels were resampled to an isotropic voxel spacing of $1.0 \times 1.0 \times 1.0 \text{ mm}^3$, ensuring uniform spatial resolution across the dataset. Constant padding was applied to match the input dimensions to $256 \times 256 \times 160$ voxels. A min-max intensity normalization was applied for each modality, scaling values to the $[0, 1]$ range. During training, random data augmentation was applied in a deterministic manner at the client level, including flipping along the spatial axes, intensity scaling, and intensity shifting.

3.2.2. Deep learning architecture

While the original FedWSOComp study [20] employed a 2D U-Net [5], we adopted a 3D U-Net to better exploit volumetric information in MRI data [44]. As discussed in Section 2, this architecture has been widely used in biomedical image segmentation tasks due to its encoder-decoder structure and skip connections, which preserve spatial detail by combining high-resolution features from the contracting path with those from the expanding path. The implemented model consisted of five hierarchical levels. At each downsampling stage, the number of feature channels was doubled, starting from 16 up to 256. Each block applied a $3 \times 3 \times 3$ convolution, followed by batch normalization and ReLU activation. Downsampling was performed via max pooling with a stride of 2. In the decoding path, transposed convolutions were used to upsample the feature maps, which were concatenated with the corresponding encoder features via skip connections to preserve localisation accuracy. A dropout rate of 0.2 was applied at the bottleneck to mitigate overfitting. The input to the model was a four-channel 3D volume comprising the T1, T1ce, T2, and FLAIR MRI modalities. A single-channel output volume was obtained through sigmoid activation, representing voxel-wise probabilities of tumour presence. The model was trained using Dice loss, which directly measured the overlap between predicted tumour masks and ground truth annotations.

3.2.3. Aggregation strategy

The FedWSOComp [20] strategy was built upon the widely adopted FedAvg algorithm [8], introducing quantization-aware modifications to reduce communication overhead between clients and the central server. While standard FedAvg involved the transmission of full-precision model updates, FedWSOComp employed compression techniques to enable scalability across numerous clients, particularly

within bandwidth-constrained settings such as federated medical image segmentation [20]. The training process began with the initialisation of the global model weights $w^{(0)}$ by the server, which then broadcast them to all participating clients. Each client trained the model locally over a fixed number of epochs, producing an updated model $w_i^{(t)}$ and computing the local update as:

$$\Delta w_i^{(t)} = w_i^{(t)} - w^{(t)} \quad (1)$$

Rather than transmitting $\Delta w_i^{(t)}$ directly, clients compressed their updates into binary representations $C_i^{(t)}$ using the FedWSOComp encoding pipeline. Only the compressed updates were transmitted to the server. Upon receiving the encoded updates, the server executed the following operations:

- Huffman decoding to recover the quantised index sequence z_i ;
- Index-to-centroid mapping to reconstruct the quantised update:

$$z_i \rightarrow \hat{w}_i = \mu_{z_i} \quad (2)$$

- Aggregation of the reconstructed updates via a weighted FedAvg scheme.

The global model was subsequently updated as:

$$w^{(t+1)} = w^{(t)} + \sum_{i=1}^N \frac{n_i}{\sum_j n_j} \cdot \Delta \hat{w}_i \quad (3)$$

where n_i denoted the number of local samples held by client i , and $\Delta \hat{w}_i$ represented the decoded and reconstructed update provided by that client. Following aggregation, the updated global model $w^{(t+1)}$ was redistributed to all clients to initiate the next training round. This process was repeated for a predefined number of communication rounds. All compression and decompression steps were deterministic, ensuring round-trip consistency without the need for additional metadata. This design led to considerable communication savings, as also reported in Onaizah et al. [20]. As in the original FedWSOComp algorithm, compression was performed independently for each layer, treating model parameters as separate tensors. Sparsification was applied using a top- z scheme [45], in which each tensor was pruned by identifying the z -quantile of absolute magnitudes and clipping smaller values while preserving their sign. This approach reshaped the update distribution without introducing zero values and was shown to improve convergence, particularly for bias tensors when using low sparsification thresholds [20]. Following sparsification, each tensor was quantised via deterministic k -means clustering, with centroids computed per layer. Unlike stochastic or fixed-level quantization schemes, this method considered the statistical distribution of tensor values, thereby reducing update variance and preserving representativeness. Raw values were replaced by cluster indices, which were subsequently encoded using entropy-based methods such as Huffman coding. More frequent indices were assigned shorter bit-lengths, further enhancing compression efficiency [20]. An overview of the complete FedWSOComp pipeline employed in this study, which illustrates the full compression workflow, is provided in Fig. 2.

3.2.4. Compression parameter analysis

The FedWSOComp strategy relies on two principal compression parameters: the number of quantization clusters (C) and the sparsification rate (S_R). In the original study [20], the number of clusters was fixed at $C = 3$, and sparsification was applied conservatively. The authors reported improved convergence, particularly for bias tensors, when sparsification thresholds were set below the Top-0.2—thus 20%—level. In the present study, we extended this analysis by systematically exploring a broader range of compression configurations. Specifically, the number of clients was held constant while performing a grid search over the two parameters, as follows:

- Number of clusters $C \in \{3, 32, 64\}$

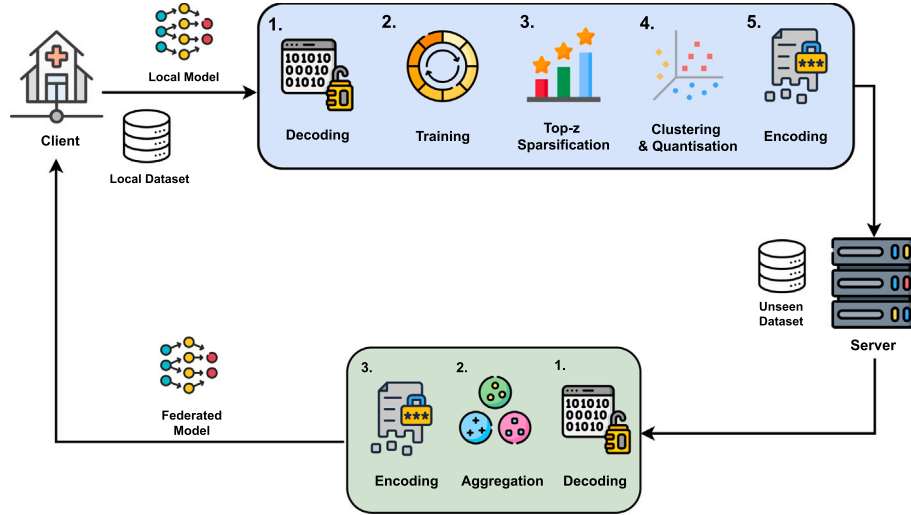


Fig. 2. Overview of the complete FedWSOComp pipeline adopted in this study.

- Soft Top- z sparsification rate $S_R \in \{20, 30, 60\}$

All possible combinations were evaluated in both IID and non-IID scenarios, yielding 18 distinct configurations. Each setting was tested using the complete FedWSOComp compression and aggregation pipeline, allowing for an isolated assessment of the relationship between compression intensity and segmentation performance under otherwise identical experimental conditions.

3.2.5. Ablation study

To disentangle the individual contributions of the different components of the FedWSOComp framework, we performed a structured ablation study under both IID and non-IID data distributions. The objective was to isolate the functional impact of each mechanism, by selectively disabling or activating specific components while keeping all other experimental conditions unchanged. Therefore, three additional experimental pipelines were implemented:

1. *FedAvg baseline.* The FedAvg strategy was used without any compression, clustering, or sparsification mechanisms. This configuration served as a reference upper-bound baseline in terms of communication fidelity and segmentation performance, allowing us to quantify the effect introduced by FedWSOComp.
2. *Clustering-only configuration (C-only).* In this setting, only the clustering-based quantization component of FedWSOComp was enabled. The number of quantization clusters was varied as $C \in \{3, 32, 64\}$, while the sparsification rate was fixed to $S_R = 0$, effectively disabling gradient sparsification.
3. *Sparsification-only configuration (S-only).* With this configuration, the clustering mechanism was disabled and only the Soft Top- z sparsification component was applied. The sparsification rate was varied as $S_R \in \{20\%, 30\%, 60\%\}$, allowing an isolated analysis of the impact of gradient pruning on model convergence and segmentation performance in the absence of quantization.

All ablation configurations were evaluated under both IID and non-IID client distributions, using the same number of clients, model architecture, training schedule, and data partitions adopted in the full FedWSOComp experiments. This controlled design ensured that any observed performance variations could be attributed exclusively to the inclusion or removal of specific FedWSOComp components, enabling a principled decomposition of their individual and combined contributions to communication efficiency and learning performance.

3.2.6. Communication efficiency evaluation

To enable a principled and reproducible analysis of communication efficiency across different the configurations, we defined a unified evaluation protocol for measuring transmission costs and compression effectiveness. The objective was to quantify the impact of each compression strategy on the communication payload exchanged between clients and server, independently of model performance metrics. For each experimental configuration presented in Section 3.2.5, the communication cost was computed as the total number of transmitted bits per communication round, accounting exclusively for model updates exchanged during the aggregation phase.

The baseline transmission cost was defined using the uncompressed FedAvg protocol, where full-precision model parameters are transmitted without any form of compression, quantization, or sparsification.

This value provided a fixed reference point for all subsequent comparisons.

Communication efficiency was quantified using two complementary metrics:

- the *compression ratio*, defined as the ratio between the FedAvg baseline payload and the compressed payload size, providing a multiplicative measure of communication reduction.
- the *payload reduction percentage*, defined as the relative decrease in transmitted bits with respect to the FedAvg baseline.

These metrics allowed for a consistent comparison across compression mechanisms, including the single components ablation study, and their combined application. All communication measurements were performed under identical experimental conditions, including model architecture, number of clients, training schedule, aggregation protocol, and data distribution, ensuring that differences in transmission cost arose solely from the adopted compression strategy.

3.2.7. Experimental DL protocol and metrics

The federation was implemented using the Flower framework [46], while the DL components were developed using PyTorch [47] and MONAI² [43]. Model optimization was performed using the Adam optimiser, with a learning rate set to 1×10^{-4} and a weight decay of 1×10^{-5} . A cosine annealing learning rate scheduler was employed, with the cycle

² All experiments were conducted on a Supermicro AS-4124GS-TNR 4U Rackmount GPU SuperServer equipped with an NVIDIA A100 80 GB GPU, 16 CPU cores, and 128 GB of RAM

length set to the number of local training epochs. Mixed-precision training was enabled via automatic loss scaling using PyTorch *GradScaler*. Due to the substantial memory requirements of 3D training, the batch size was set to 1. For all experiments, clients were trained for 25 rounds, each consisting of 5 local epochs.

Segmentation performance was primarily assessed using standard medical imaging segmentation metrics:

- the Dice Similarity Coefficient (DSC), which quantified the volumetric overlap between the predicted and ground truth tumour masks. The DSC, bounded between 0 and 1, was defined as:

$$DSC = \frac{2|A \cap B|}{|A| + |B|} \quad (4)$$

where A and B denoted the predicted and ground truth binary masks, respectively.

- the 95th percentile Hausdorff Distance (HD95) [48] was computed to evaluate spatial agreement between segmentations, offering complementary insight into anatomical accuracy. The HD95 metric corresponded to the 95th percentile of all pairwise distances, reducing the influence of outliers.

4. Results

4.1. Compression parameter analysis

As reported in Section 3.2.4, the FedWSOComp strategy was assessed under 18 different compression configurations, by varying both the quantization cluster size and the Soft Top-z sparsification retention ratio, in the context of IID and non-IID data distributions. Tables 2 and 3 summarise the performance of the best-performing models, defined as those achieving the highest validation performance during training.

These models were subsequently evaluated on an independent test dataset that remained unseen during training. Both the global server models and the local clients models were considered, allowing the assessment of cross-site generalization as well as client-specific adaptation.

For Client 1, under non-IID conditions with $C = 64$ and $S_R = 60\%$, the local model achieved a DSC of 0.99 ± 0.03 and an HD95 of 5.15 ± 4.17 mm, indicating high segmentation quality when compression is lower. The federated model, trained under the same parameters and evaluated on the independent test set, obtained a DSC of 0.98 ± 0.09 and an HD95 of 10.40 ± 15.54 mm, confirming that the applied sparsification ratio

retained sufficient information to preserve boundary fidelity in both global and local models. In IID conditions, DSC values peaked at $S_R = 30\%$ with $C = 32$ (0.85 ± 0.24), although further increasing S_R to 60% improved boundary accuracy, yielding an HD95 of 9.75 ± 7.38 mm at $C = 64$, with a slight trade-off in DSC (0.81 ± 0.31). This suggested that, in IID settings, volumetric overlap was maximised at $S_R = 30\%$, while boundary accuracy improved at higher retention. Interestingly, this trade-off disappeared under non-IID conditions, where high C values combined with $S_R = 60\%$ yielded the best overall performance across both metrics.

For Client 2, which displayed greater intra-dataset variability, the best non-IID outcome was reached at $C = 64$, $S_R = 60\%$, with DSC 0.79 ± 0.34 and HD95 10.19 ± 8.24 mm. Here, boundary accuracy surpassed that of the global model, while DSC remained comparable. In IID settings, results were more consistent across parameter combinations. The best outcome was again observed with $C = 64$, $S_R = 60\%$ (DSC 0.53 ± 0.44 , HD95 11.64 ± 8.37 mm), with a marginal DSC improvement at $C = 32$, $S_R = 60\%$ (DSC 0.56 ± 0.47 , HD95 12.87 ± 9.14 mm).

For Client 3, the non-IID setting with $C = 32$, $S_R = 60\%$ resulted in a local HD95 of 11.95 ± 9.24 mm. The corresponding global model under the same parameters achieved DSC 0.98 ± 0.12 and HD95 13.74 ± 15.71 mm, reflecting stable performance. In IID conditions, the lowest HD95 was observed at $C = 32$, $S_R = 60\%$ (11.95 ± 9.24 mm), whereas the highest DSC was obtained at $C = 64$, $S_R = 20\%$ (0.92 ± 0.13), though at the expense of substantially higher HD95.

For Client 4, segmentation accuracy reached its maximum, with non-IID DSC values of 1.00 ± 0.00 in several configurations. However, boundary precision was not always preserved, as illustrated by $C = 64$, $S_R = 60\%$ (non-IID), which achieved HD95 of 10.40 ± 15.54 mm despite near-perfect DSC. In IID conditions, DSC peaked at 0.99 ± 0.04 with $C = 32$, $S_R = 30\%$ and $C = 64$, $S_R = 30\%$, while boundary accuracy further improved at higher retention (e.g., $C = 64$, $S_R = 60\%$, HD95 14.18 ± 9.96 mm).

A comparison of the results obtained on the independent set (reported as “Server” in Tables 2 and 3) revealed a consistent pattern: moderate-to-fine quantization ($C = 32$ or $C = 64$) combined with higher retention ($S_R = 60\%$) yielded the most favourable balance between DSC and HD95. Specifically, under IID, $C = 32$ or 64 with $S_R = 60\%$ achieved DSC values of 0.98 ± 0.12 and 0.98 ± 0.13 , and HD95 of 13.74 ± 15.71 mm and 14.11 ± 15.74 mm, respectively. Under non-IID conditions, the same retention ratio with $C = 64$ produced optimal performance (DSC 0.98 ± 0.09 , HD95 10.40 ± 15.54 mm). Conversely, low retention ($S_R = 20\%$)

Table 2

Test results under IID data distribution for different combinations of C and S_R . Dice Similarity Coefficient (DSC) and 95th percentile Hausdorff Distance (HD95) are reported as mean $\pm$ standard deviation for the server and each client.

C	S_R (%)	Metric	Server	Client 1	Client 2	Client 3	Client 4
3	20	DSC	0.77 ± 0.21	0.24 ± 0.12	0.21 ± 0.17	0.69 ± 0.32	0.74 ± 0.24
		HD95	69.23 ± 13.28	81.00 ± 17.40	88.44 ± 7.26	69.28 ± 16.99	68.60 ± 11.51
3	30	DSC	0.93 ± 0.14	0.47 ± 0.25	0.40 ± 0.40	0.78 ± 0.26	0.94 ± 0.13
		HD95	57.70 ± 19.48	74.02 ± 19.25	85.46 ± 9.93	59.33 ± 24.69	57.16 ± 16.56
3	60	DSC	0.90 ± 0.14	0.63 ± 0.28	0.37 ± 0.34	0.71 ± 0.26	0.93 ± 0.09
		HD95	60.47 ± 18.88	75.41 ± 18.48	86.90 ± 10.44	58.08 ± 22.77	63.19 ± 15.99
32	20	DSC	0.95 ± 0.13	0.54 ± 0.25	0.46 ± 0.37	0.89 ± 0.18	0.94 ± 0.08
		HD95	61.23 ± 18.73	73.35 ± 19.42	83.39 ± 12.31	60.08 ± 21.94	61.71 ± 16.63
32	30	DSC	0.97 ± 0.12	0.85 ± 0.24	0.43 ± 0.40	0.78 ± 0.33	0.99 ± 0.04
		HD95	25.25 ± 24.93	26.95 ± 20.46	46.95 ± 35.99	24.04 ± 13.53	22.24 ± 18.19
32	60	DSC	0.98 ± 0.12	0.80 ± 0.32	0.56 ± 0.47	0.75 ± 0.40	0.98 ± 0.07
		HD95	13.74 ± 15.71	10.69 ± 6.95	12.87 ± 9.14	11.95 ± 9.24	14.20 ± 9.84
64	20	DSC	0.95 ± 0.12	0.56 ± 0.27	0.48 ± 0.37	0.92 ± 0.13	0.97 ± 0.05
		HD95	58.60 ± 20.52	72.47 ± 19.76	80.90 ± 14.01	58.31 ± 22.91	59.71 ± 18.02
64	30	DSC	0.97 ± 0.12	0.85 ± 0.24	0.45 ± 0.39	0.80 ± 0.31	0.99 ± 0.04
		HD95	26.59 ± 25.50	32.42 ± 19.03	48.32 ± 36.29	25.41 ± 16.46	27.74 ± 21.01
64	60	DSC	0.98 ± 0.13	0.81 ± 0.31	0.53 ± 0.44	0.75 ± 0.40	0.97 ± 0.08
		HD95	14.11 ± 15.74	9.75 ± 7.38	11.64 ± 8.37	12.23 ± 8.87	14.18 ± 9.96

Table 3

Test results under non-IID data distribution for different combinations of C and S_R . Dice Similarity Coefficient (DSC) and 95th percentile Hausdorff Distance (HD95) are reported as mean $\pm$ standard deviation for the server and each client.

C	S_R (%)	Metric	Server	Client 1	Client 2	Client 3	Client 4
3	20	DSC	0.92 $\pm$ 0.15	0.55 $\pm$ 0.28	0.42 $\pm$ 0.38	0.82 $\pm$ 0.23	0.92 $\pm$ 0.14
		HD95	61.35 $\pm$ 18.48	74.32 $\pm$ 18.77	82.08 $\pm$ 13.39	58.33 $\pm$ 22.56	61.08 $\pm$ 15.35
3	30	DSC	0.96 $\pm$ 0.11	0.76 $\pm$ 0.28	0.45 $\pm$ 0.38	0.83 $\pm$ 0.23	0.99 $\pm$ 0.03
		HD95	54.02 $\pm$ 22.78	69.89 $\pm$ 19.51	71.85 $\pm$ 25.79	41.03 $\pm$ 29.92	51.61 $\pm$ 22.15
3	60	DSC	0.98 $\pm$ 0.09	0.83 $\pm$ 0.23	0.48 $\pm$ 0.37	0.84 $\pm$ 0.22	1.00 $\pm$ 0.00
		HD95	50.21 $\pm$ 25.27	68.50 $\pm$ 16.99	69.82 $\pm$ 27.20	39.93 $\pm$ 27.89	47.92 $\pm$ 26.84
32	20	DSC	0.98 $\pm$ 0.12	0.83 $\pm$ 0.25	0.42 $\pm$ 0.40	0.81 $\pm$ 0.30	1.00 $\pm$ 0.00
		HD95	37.66 $\pm$ 29.13	50.16 $\pm$ 10.80	69.19 $\pm$ 19.19	35.92 $\pm$ 29.82	38.13 $\pm$ 26.86
32	30	DSC	0.98 $\pm$ 0.13	0.85 $\pm$ 0.25	0.40 $\pm$ 0.41	0.77 $\pm$ 0.36	0.99 $\pm$ 0.04
		HD95	27.02 $\pm$ 27.49	32.43 $\pm$ 21.42	49.31 $\pm$ 33.66	33.27 $\pm$ 27.47	28.79 $\pm$ 26.55
32	60	DSC	0.97 $\pm$ 0.14	0.70 $\pm$ 0.44	0.31 $\pm$ 0.45	0.70 $\pm$ 0.46	0.91 $\pm$ 0.26
		HD95	13.34 $\pm$ 15.23	87.39 $\pm$ 68.91	17.93 $\pm$ 12.65	15.21 $\pm$ 19.05	12.10 $\pm$ 7.62
64	20	DSC	0.98 $\pm$ 0.12	0.82 $\pm$ 0.25	0.48 $\pm$ 0.37	0.84 $\pm$ 0.25	1.00 $\pm$ 0.00
		HD95	44.74 $\pm$ 27.34	59.76 $\pm$ 9.61	74.73 $\pm$ 15.02	37.32 $\pm$ 29.96	45.79 $\pm$ 23.61
64	30	DSC	0.97 $\pm$ 0.14	0.90 $\pm$ 0.21	0.37 $\pm$ 0.42	0.74 $\pm$ 0.40	0.95 $\pm$ 0.11
		HD95	22.25 $\pm$ 24.93	29.58 $\pm$ 21.86	43.54 $\pm$ 34.15	28.92 $\pm$ 23.14	28.23 $\pm$ 27.08
64	60	DSC	0.98 $\pm$ 0.09	0.99 $\pm$ 0.03	0.79 $\pm$ 0.34	0.83 $\pm$ 0.34	0.98 $\pm$ 0.05
		HD95	10.40 $\pm$ 15.54	5.15 $\pm$ 4.17	10.19 $\pm$ 8.24	11.79 $\pm$ 10.45	17.56 $\pm$ 15.31

combined with coarse quantization ($C = 3$) proved least effective, with IID DSC as low as 0.77 ± 0.21 and HD95 69.23 ± 13.28 mm, and non-IID DSC 0.92 ± 0.15 with HD95 61.35 ± 18.48 mm.

Fig. 3 further illustrates these trends. In IID settings, $C = 32$ with $S_R = 60\%$ produced smooth, monotonic DSC improvements with minimal oscillations, indicative of stable convergence. In contrast, coarse quantization with low retention (e.g., $C = 3$, $S_R = 20\%$) resulted in unstable trajectories, with rapid degradation of DSC during training. In non-IID conditions, configurations with high retention ($S_R = 60\%$) and fine quantization ($C = 64$) retained DSC above 0.95 throughout, demonstrating robustness against heterogeneous updates. In addition, to complement the quantitative findings, a fixed-slice comparison was performed to highlight qualitative differences between configurations, as shown in Figs. 4 and 5 for IID and non-IID settings, respectively.

4.2. Ablation study

As described in Section 3.2.5, to assess the individual contribution of each component of the *FedWSOcomp* framework, we conducted an ablation study under both IID and non-IID data distributions. In the following Section, we report the results of each ablation configuration, including a systematic comparison with the full *FedWSOcomp* pipeline using the reference configuration $C = 32$ and $S_R = 60\%$.

The results reported in Tables 4 and 5 correspond to the best-performing communication rounds, thus the models achieving the highest validation performance during the federated training process. Performance is reported for both the server (independent test set) and the clients.

Under IID conditions (Table 4), the FedAvg baseline achieved strong segmentation performance at both server and client levels, with a noticeable degradation on Client 2, which—as also reported in Section 4.1—exhibited the highest intra-dataset variability.

When quantization was isolated ($C - only$), with coarse quantization ($C = 3$), segmentation fidelity deteriorated substantially across both metrics, indicating that the transmitted updates were excessively distorted to preserve tumour boundaries. In contrast, moderate-to-fine quantization ($C = 32$, $C = 64$) preserved both overlap and boundary accuracy at the server level, achieving an average server-side performance of 0.98 ± 0.12 and HD95 10.82 ± 15.72 mm for $C = 32$, and DSC 0.99 ± 0.02 with HD95 9.01 ± 11.62 mm for $C = 64$. In the latter configuration, HD95 improved consistently both at the server and client levels, indicating

enhanced boundary preservation. This behaviour can be attributed to the sensitivity of boundary-based metrics to inconsistencies in local gradient directions and anatomical representations across heterogeneous clients. While FedAvg performs a global parameter averaging that can smooth and distort fine-grained boundary features, clustering-based aggregation groups clients with similar data distributions, thereby preserving coherent geometric and anatomical structures.

When sparsification was evaluated independently ($S_R - only$), with low-to-moderate sparsification ($S_R = 20\%$, $S_R = 40\%$), segmentation quality was already substantially degraded, as reflected by moderate DSC values and high HD95, indicating pronounced boundary distortions. With stronger sparsification ($S_R = 60\%$), volumetric overlap remained relatively high, with an average server-side DSC of 0.94 ± 0.13 . However, boundary accuracy still deteriorated, with an average server-side HD95 of 41.05 ± 30.26 mm. These findings suggested that sparsification alone can preserve volumetric overlap while introducing significant boundary errors.

Under non-IID conditions (Table 5), the same overall trends were observed. FedAvg remained robust at both server and client levels. The $C - only$ configurations, showed that coarse clustering induced severe boundary degradation, with an average server-side HD95 of 68.69 ± 14.58 mm for $C = 3$, whereas finer configurations ($C = 32$, $C = 64$) preserved both overlap and boundary accuracy.

For the $S_R - only$ configurations, boundary degradation was consistently observed across sparsity ratios. At low sparsification ($S_R = 20\%$), although overlap remained relatively high, boundary accuracy was already severely compromised, indicating once again significant contour deviations.

The same behaviour was observed with moderate and higher sparsification ($S_R = 60\%$). Overlap remained high (i.e., server-side DSC of 0.98 ± 0.08), but boundary accuracy remained unsatisfactory compared to FedAvg and $C - only$ aggregation with $C \in \{32, 64\}$.

4.3. Communication efficiency evaluation

Communication efficiency results are reported in Table 6. The baseline transmission cost was constant across all modes (1.54×10^8 bits). The $C - only$ configurations achieved compression ratios ranging from $\times 3.13$ to $\times 3.77$, corresponding to payload reductions of 68.11% – 73.48% . The highest compression was obtained for the coarsest clustering setting ($C = 3$, $\times 3.77$), which, however, corresponds to the configuration

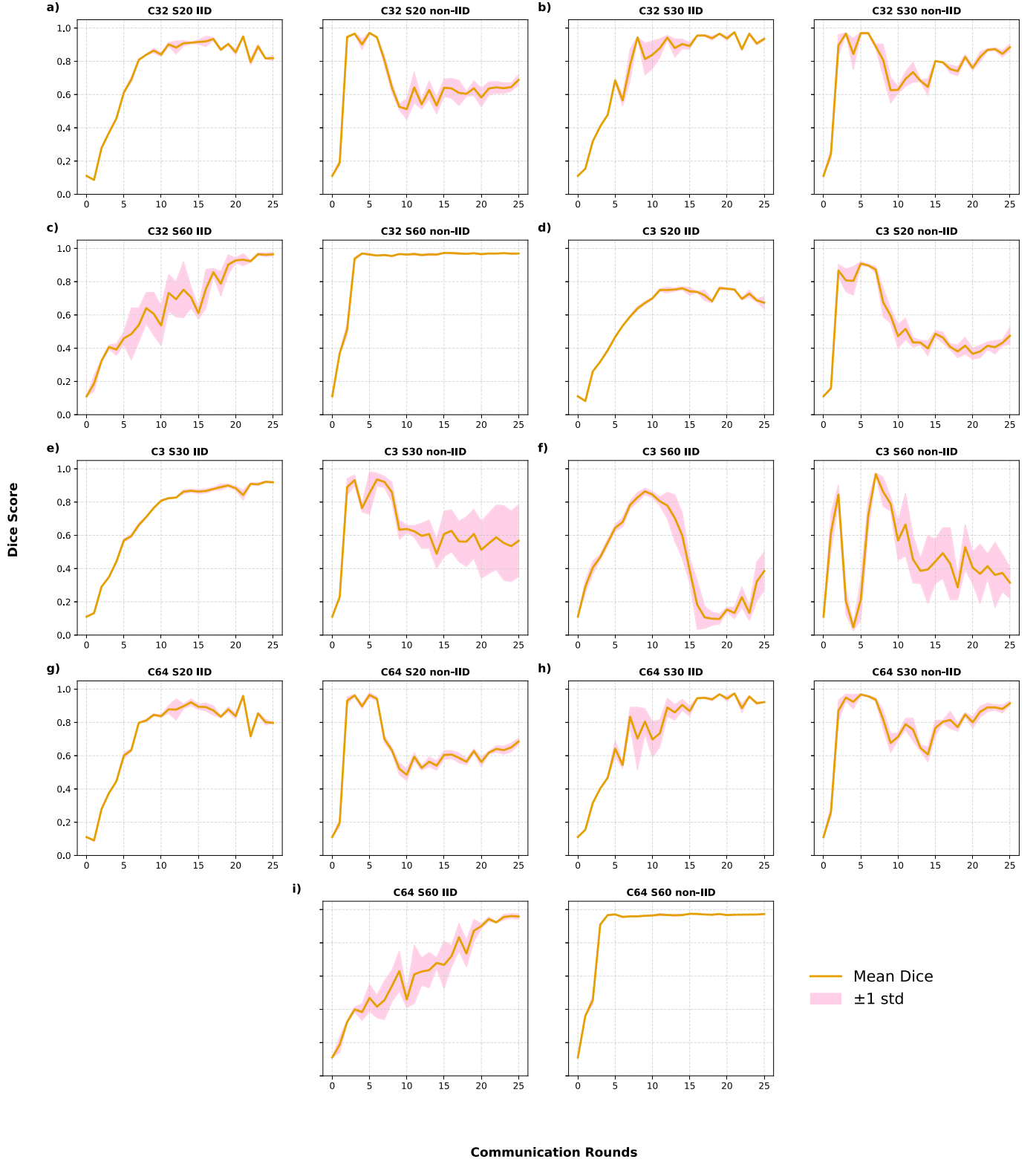


Fig. 3. Performance trends for varying S_R and C values under IID and non-IID settings.

yielding the poorest segmentation performance in Table 4, highlighting a clear trade-off between communication efficiency and model accuracy.

For the S_R -only configurations, communication gains were strongly dependent on the sparsity level. Specifically, S_R -only with $S_R = 20\%$ resulted in no effective reduction, matching the FedAvg baseline, while $S_R = 30\%$ produced only a negligible improvement ($\times 1.01$, 0.93%). Increasing sparsity to $S_R = 60\%$ led to a modest communication

reduction ($\times 1.24$, 19.48%), which remained substantially weaker than the gains achieved by C -only strategies.

In contrast, the combined configuration –reported as FedWSOcomp– using the $C = 32$ and $S_R = 60\%$ configuration achieved a good compression ratio of $\times 3.40$ with a 70.65% payload reduction, providing communication savings comparable to C -only while simultaneously preserving segmentation performance.

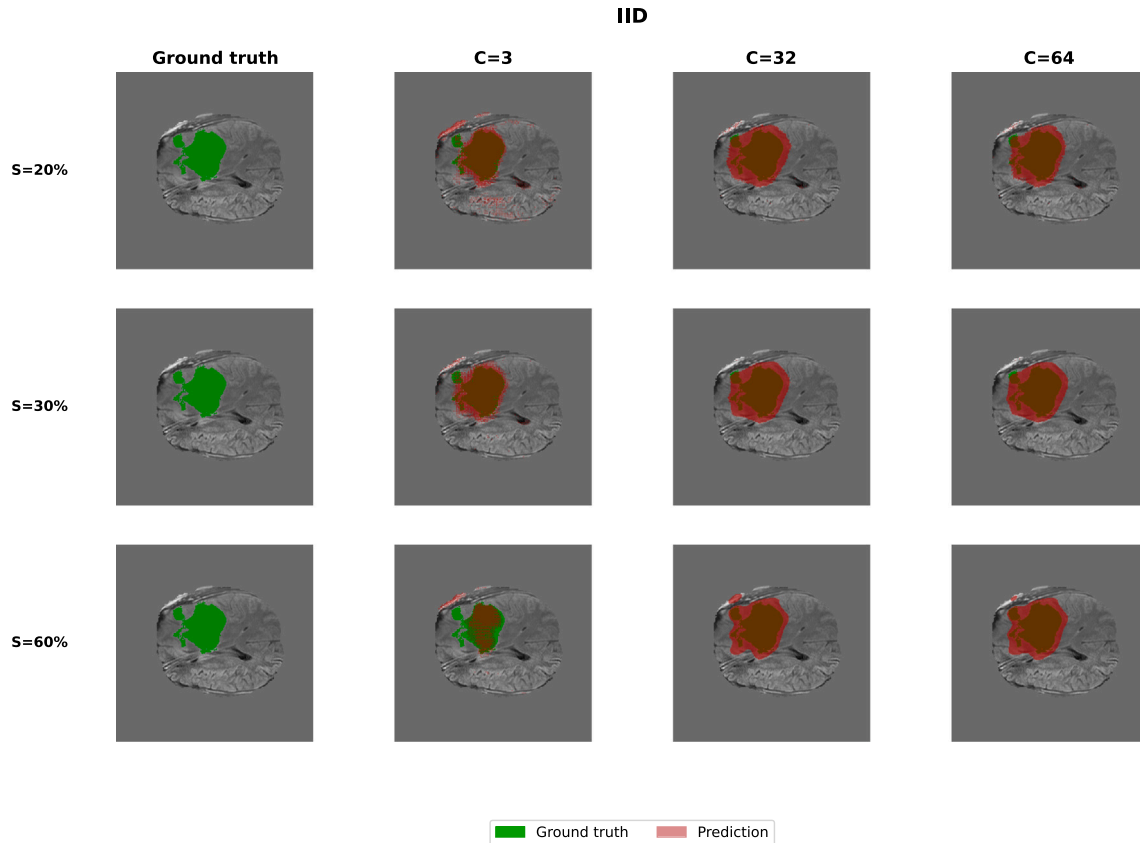


Fig. 4. Qualitative comparison of segmentation predictions across all compression configurations under IID conditions. Results are shown for different sparsification rates ($S_R = 20\%$, 30% , 60%) and quantization levels ($C = 3, 32, 64$), with the corresponding ground truth masks.

5. Discussion

The experimental evaluation of FedWSOComp under both IID (Table 2) and non-IID (Table 3) conditions provided insights into the relationship between sparsification, quantization, and federated aggregation in 3D brain tumour segmentation. At the global level, the results indicated that extreme compression ($C = 3$, $S_R = 20\%$) led to severe performance degradation, with the DSC falling below 0.80 and the HD95 exceeding 60 mm. These findings confirmed that aggressive sparsification combined with coarse quantization caused substantial information loss, limiting the global model's ability to capture tumour morphology. In contrast, configurations with higher retention ($S_R = 60\%$) and moderate-to-fine quantization ($C = 32$ or $C = 64$) consistently achieved the best trade-off between accuracy and communication efficiency. In non-IID settings, the combination of $C = 64$ and $S_R = 60\%$ reached better overall performance, maintaining high segmentation quality despite heterogeneous client distributions. This observation aligned with prior studies that emphasised the importance of retaining high-magnitude parameters to preserve representational capacity while reducing communication costs [16,40,49]. Conceptually, FedWSOComp shared similarities with the dynamic layer-wise sparsification strategy introduced by Zhang et al. [40], but it differed in its adoption of a fixed *soft* Top- z policy. This softer retention mechanism appeared effective in preserving fine anatomical structures, which are particularly relevant in volumetric tumour segmentation.

In addition to the quantitative analysis, the qualitative results presented in Figs. 4 and 5 provide further insight into the behaviour of FedWSOComp under different compression settings. In the IID case, coarse quantization with $C = 3$ and low retention ($S_R = 20\%$) produced poorly defined segmentations, including spurious regions extending beyond the brain volume. Increasing retention to $S_R = 30\%$ and $S_R = 60\%$ reduced these artefacts, but the segmentations remained poor. Using

finer quantization ($C = 32$ and $C = 64$), the predictions covered the ground truth more consistently, with only minor deviations along the outer lesion boundaries. However, the difference between $C = 32$ and $C = 64$ was negligible, indicating limited benefit from doubling the number of clusters. Under non-IID conditions, the larger number of samples available to each client led to more precise segmentations overall. As in the IID setting, $C = 3$ resulted in inaccurate predictions, particularly at $S_R = 20\%$ and $S_R = 30\%$. An improvement was noticed only using $S_R = 60\%$, although the model still failed to match the ground truth with high accuracy. In contrast, with $C = 32$ and $C = 64$, segmentation quality clearly improved as retention increased, yielding satisfactory results at $S_R = 30\%$ and excellent delineations at $S_R = 60\%$. Overall, these qualitative observations confirmed quantitative results. A coarse quantization ($C = 3$) consistently degraded segmentation quality, with the degradation being most pronounced when fewer training cases were available (IID scenario). By contrast, moderate-to-fine quantization ($C = 32$ or $C = 64$) combined with higher retention produced more accurate and reliable lesion delineations, particularly in heterogeneous non-IID scenarios.

Notably, substantial variability was observed across clients. For instance, Client 2, trained on a more heterogeneous dataset, exhibited reduced stability, whereas Client 4, with more homogeneous data, achieved near-perfect DSC. However, higher DSC values were sometimes associated with increased HD95, suggesting a trade-off between volumetric overlap and boundary precision. Such site-specific differences reinforced the need to assess FL models at both independent institutions and individual clients levels, in agreement with recent federated studies [10,11,34].

Under IID conditions, $C = 32$ provided the most favourable balance, achieving tighter variance bounds and higher DSC compared with $C = 3$ or $C = 64$. Although $C = 64$ yielded marginal robustness advantages

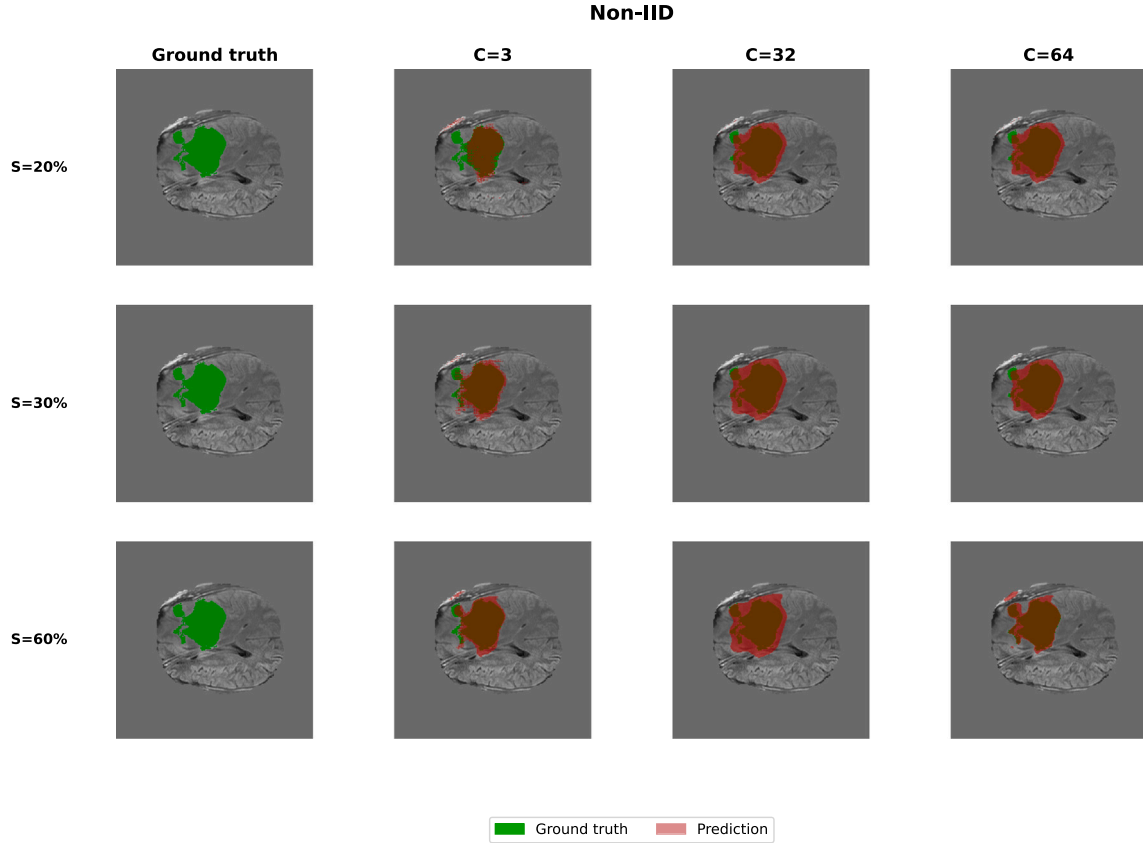


Fig. 5. Qualitative comparison of segmentation predictions across all compression configurations under non-IID conditions. Results are shown for different sparsification rates ($S_R = 20\%, 30\%, 60\%$) and quantization levels ($C = 3, 32, 64$), with the corresponding ground truth masks.

Table 4

Test results under IID data distribution for the four experiment modes: FedAvg, CH ($S_R = 0, C \in \{3, 32, 64\}$), SH ($C = 0, S_R = 60\%$), and SKH ($C = 32, S_R = 60\%$). Dice Similarity Coefficient (DSC) and 95th percentile Hausdorff Distance (HD95) are reported as mean $\pm$ standard deviation for the server and each client.

Pipeline	C	S_R (%)	Metric	Server	Client 1	Client 2	Client 3	Client 4
FedAvg	0	0	DSC	0.99 $\pm$ 0.04	0.90 $\pm$ 0.30	0.71 $\pm$ 0.39	0.98 $\pm$ 0.05	1.00 $\pm$ 0.00
			HD95	12.58 $\pm$ 18.87	12.32 $\pm$ 16.79	11.30 $\pm$ 9.18	11.39 $\pm$ 9.55	14.12 $\pm$ 13.54
<i>C - only</i>	3	0	DSC	0.59 $\pm$ 0.21	0.46 $\pm$ 0.26	0.40 $\pm$ 0.28	0.54 $\pm$ 0.24	0.68 $\pm$ 0.18
			HD95	73.28 $\pm$ 12.69	76.19 $\pm$ 15.31	76.47 $\pm$ 14.14	73.21 $\pm$ 13.12	64.77 $\pm$ 8.36
<i>C - only</i>	32	0	DSC	0.98 $\pm$ 0.12	0.88 $\pm$ 0.30	0.63 $\pm$ 0.45	0.79 $\pm$ 0.34	1.00 $\pm$ 0.00
			HD95	10.82 $\pm$ 15.72	13.48 $\pm$ 18.47	12.50 $\pm$ 9.42	10.79 $\pm$ 9.57	14.04 $\pm$ 12.51
C - only	64	0	DSC	0.99 $\pm$ 0.02	0.89 $\pm$ 0.30	0.78 $\pm$ 0.32	0.97 $\pm$ 0.07	1.00 $\pm$ 0.00
			HD95	9.01 $\pm$ 11.62	8.55 $\pm$ 7.97	9.07 $\pm$ 7.77	8.07 $\pm$ 7.30	13.37 $\pm$ 13.85
<i>S_R - only</i>	0	20	DSC	0.74 $\pm$ 0.23	0.26 $\pm$ 0.14	0.40 $\pm$ 0.32	0.87 $\pm$ 0.17	0.64 $\pm$ 0.16
			HD95	61.49 $\pm$ 18.29	74.59 $\pm$ 17.84	83.43 $\pm$ 12.43	61.53 $\pm$ 21.26	60.95 $\pm$ 16.42
<i>S_R - only</i>	0	30	DSC	0.74 $\pm$ 0.17	0.30 $\pm$ 0.14	0.40 $\pm$ 0.27	0.70 $\pm$ 0.28	0.65 $\pm$ 0.27
			HD95	68.89 $\pm$ 12.76	77.19 $\pm$ 17.07	85.70 $\pm$ 9.29	68.51 $\pm$ 16.51	69.02 $\pm$ 11.55
S_R - only	0	60	DSC	0.94 $\pm$ 0.13	0.72 $\pm$ 0.28	0.80 $\pm$ 0.31	1.00 $\pm$ 0.00	0.92 $\pm$ 0.08
			HD95	41.05 $\pm$ 30.26	30.61 $\pm$ 20.60	32.73 $\pm$ 31.09	27.14 $\pm$ 18.73	42.50 $\pm$ 28.49

under non-IID distributions, this came at the cost of increased communication. Similar behaviour was observed in high-retention scenarios ($S_R = 60\%$), consistent with the findings of Jia et al. [50], who showed that inappropriate parameter choices could amplify inter-client variance. In this work, deterministic k -means quantization combined with Huffman coding appeared to mitigate such instability by constraining convergence trajectories.

When compared with FedSparse [49], which primarily relied on sparse gradient updates, FedWSOComp integrated quantization directly into the sparsification pipeline and employed entropy coding to enable exact weight reconstruction. This design minimised communication

payloads while maintaining boundary integrity—an essential aspect in 3D segmentation, where even minor inaccuracies may compromise clinical reliability.

Interestingly, the ablation results revealed that HD95 degradation was more pronounced under IID conditions than in non-IID settings for the FedAvg baseline. This counterintuitive behaviour was likely due to the limited number of patients used to construct a perfectly balanced IID partition, which may reduce boundary generalization in individual experiments. In contrast, the inherent heterogeneity of the non-IID setting appears to provide more stable anatomical variability, supporting improved generalization in some cases.

Table 5

Test results under non-IID data distribution for the four experiment modes: FedAvg ($C = 0$, $S_R = 0$), CH ($S_R = 0$, $C \in \{3, 32, 64\}$), SH ($C = 0$, $S_R = 60\%$), and SKH ($C = 32$, $S_R = 60\%$). Dice Similarity Coefficient (DSC) and 95th percentile Hausdorff Distance (HD95) are reported as mean $\pm$ standard deviation for the server and each client.

Pipeline	C	S_R (%)	Metric	Server	Client 1	Client 2	Client 3	Client 4
FedAvg	0	0	DSC	0.99 $\pm$ 0.05	0.96 $\pm$ 0.17	0.83 $\pm$ 0.34	0.94 $\pm$ 0.16	0.97 $\pm$ 0.07
			HD95	8.99 $\pm$ 13.75	5.49 $\pm$ 3.88	7.84 $\pm$ 8.55	8.50 $\pm$ 7.97	13.67 $\pm$ 14.40
$C - only$	3	0	DSC	0.85 $\pm$ 0.18	0.74 $\pm$ 0.27	0.60 $\pm$ 0.34	0.79 $\pm$ 0.22	0.91 $\pm$ 0.12
			HD95	68.69 $\pm$ 14.58	71.69 $\pm$ 15.16	73.57 $\pm$ 14.53	67.25 $\pm$ 15.19	61.13 $\pm$ 11.09
$C - only$	32	0	DSC	0.98 $\pm$ 0.08	0.98 $\pm$ 0.14	0.88 $\pm$ 0.31	0.94 $\pm$ 0.24	0.97 $\pm$ 0.09
			HD95	9.91 $\pm$ 14.47	6.09 $\pm$ 5.29	7.13 $\pm$ 7.31	7.80 $\pm$ 7.21	15.39 $\pm$ 15.27
C – only	64	0	DSC	0.99 $\pm$ 0.05	0.97 $\pm$ 0.16	0.84 $\pm$ 0.33	0.94 $\pm$ 0.19	0.98 $\pm$ 0.08
			HD95	8.51 $\pm$ 13.03	5.51 $\pm$ 4.22	8.27 $\pm$ 9.27	8.33 $\pm$ 7.95	12.18 $\pm$ 15.27
$S_R - only$	0	20	DSC	0.89 $\pm$ 0.17	0.93 $\pm$ 0.17	0.86 $\pm$ 0.24	0.94 $\pm$ 0.10	0.89 $\pm$ 0.17
			HD95	57.43 $\pm$ 19.32	56.07 $\pm$ 20.65	62.66 $\pm$ 16.77	48.63 $\pm$ 22.21	54.40 $\pm$ 13.41
$S_R - only$	0	30	DSC	0.99 $\pm$ 0.05	0.96 $\pm$ 0.12	0.88 $\pm$ 0.23	0.97 $\pm$ 0.07	0.99 $\pm$ 0.04
			HD95	56.00 $\pm$ 22.62	56.23 $\pm$ 24.09	57.18 $\pm$ 21.47	44.58 $\pm$ 26.80	54.91 $\pm$ 18.17
$S_R - only$	0	60	DSC	0.98 $\pm$ 0.08	0.97 $\pm$ 0.17	0.93 $\pm$ 0.12	1.00 $\pm$ 0.00	0.97 $\pm$ 0.11
			HD95	23.04 $\pm$ 27.07	16.32 $\pm$ 24.15	10.92 $\pm$ 13.38	24.74 $\pm$ 37.29	36.40 $\pm$ 25.61

Table 6

Communication costs for different compression pipelines. For each method, the minimum transmitted payload observed across training rounds is reported.

Pipeline	C	S_R (%)	Baseline bits	Compressed bits	Compression ratio	Payload reduction (%)
FedAvg	0	0	1.54×10^8	1.54×10^8	$\times 1.00$	0.00
$C - only$	3	0	1.54×10^8	4.09×10^7	$\times 3.77$	73.48
$C - only$	32	0	1.54×10^8	4.73×10^7	$\times 3.26$	69.31
$C - only$	64	0	1.54×10^8	4.91×10^7	$\times 3.13$	68.11
$S - only$	0	20	1.54×10^8	1.54×10^8	$\times 1.00$	0.00
$S - only$	0	30	1.54×10^8	1.53×10^8	$\times 1.01$	0.93
$S - only$	0	60	1.54×10^8	1.24×10^8	$\times 1.24$	19.48
FedWSOComp	32	60	1.54×10^8	4.52×10^7	$\times 3.40$	70.65

The communication results reported in Table 6 further clarified that the full FedWSOComp pipeline represents a more practical operating point than clustering or sparsification alone. Although the $C - only$ configuration achieved substantial compression (up to $3.77\times$), the most aggressive configuration ($C = 3$) was associated with the weakest segmentation performance –presented in Tables 4 and 5– and the largest boundary errors.

Furthermore, the $S_R - only$ configurations have not yielded consistent communication savings across retention levels while simultaneously increasing the boundary errors represented by the HD95.

The findings of this study, when considered in conjunction with the ablation and communication analyses, indicate that the primary advantage of FedWSOComp arises from the synergistic combination of sufficiently fine-grained quantization with high retention, rather than from either compression mechanism in isolation. From a pragmatic standpoint, the configuration $C = 32$ with $S_R = 60\%$ was identified as the most balanced operating regime under IID conditions with constrained sample availability, while $C = 64$ with $S_R = 60\%$ was identified as the most effective configuration under non-IID settings. It was demonstrated that both configurations consistently achieved high DSC while maintaining low HD95, and simultaneously delivered high communication reductions without the instability associated with coarse quantization or low-retention sparsification.

Overall, these configurations delineate a stable and deployable trade-off between communication efficiency and segmentation reliability, thus supporting the assertion that FedWSOComp may be considered suitable for clinical FL applications with constrained computational and network infrastructures, without compromising the robustness of tumour delineation.

6. Conclusion

This work analysed FedWSOComp, an FL strategy for communication-efficient 3D medical image segmentation, by systematically varying its key components, parameters and analysing their effects on model performance and compression. The study showed how different sparsification levels and quantization granularities influenced the trade-off between segmentation accuracy and communication efficiency, providing new insights into the practical applicability of the method. The results confirmed that moderate-to-fine quantization combined with higher retention ratios preserved segmentation quality while still achieving substantial communication savings. Moreover, the analysis of client-specific outcomes highlighted the importance of considering heterogeneity in federated scenarios, as improvements varied according to local data distributions. A limitation of this study concerned the construction of the IID scenario. By allocating 100 patients to each client, the total number of patients available in this case was lower than in the non-IID setting, which explained the performance degradation observed in some IID experiments. Overall, this study deepened the understanding of FedWSOComp and its operational boundaries, offering guidance for its deployment in bandwidth-constrained clinical networks. Future investigations may explore adaptive parameter selection, integration with privacy-preserving techniques, and large-scale validation across multiple institutions and imaging modalities.

CRedit authorship contribution statement

Asaf Raza: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Ciro Benito Raggio:** Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Antonella Guzzo:** Writing – review & editing, Validation, Supervision, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Maria Francesca Spadea:** Writing – review & editing, Validation, Supervision, Methodology, Formal analysis, Conceptualization. **Giancarlo Fortino:** Supervision, Project administration, Funding acquisition.

Funding

This work was partially supported by the following projects: (1) POS RADIOAMICA project funded by the Italian Ministry of Health (CUP: H53C22000650006); (2) PNRR project FAIR-Future AI Research (PE00000013), Spoke 9, under the NRRP MUR program funded by the NextGenerationEU.

Declaration of competing interest

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

Data availability statement

The data used in this study were collected from the public Federated Tumor Segmentation (FeTS) Challenge 2024 dataset and are available at: <https://doi.org/10.5281/zenodo.10990500>.

References

- [1] T.A. Soomro, L. Zheng, A.J. Afifi, A. Ali, S. Soomro, M. Yin, J. Gao, Image segmentation for mr brain tumor detection using machine learning: a review, *IEEE Rev. Biomed. Eng.* 16 (2022) 70–90, <https://doi.org/10.1109/RBME.2022.3185292>
- [2] S. Karim, G. Tong, Y. Yu, A.A. Laghari, A.A. Khan, M. Ibrar, F. Mehmood, Developments in brain tumor segmentation using MRI: deep learning insights and future perspectives, *IEEE Access* 12 (2024) 26875–26896, <https://doi.org/10.1109/ACCESS.2024.3365048>
- [3] S. Alphonse, F. Mathew, K. Dhanush, V. Dinesh, Federated learning with integrated attention multiscale model for brain tumor segmentation, *Sci. Rep.* 15 (2025) 11889, <https://doi.org/10.1038/s41598-025-96416-6>
- [4] F.J. Dorfner, J.B. Patel, J. Kalpathy-Cramer, E.R. Gerstner, C.P. Bridge, A review of deep learning for brain tumor analysis in MRI, *NPJ Precis. Oncol.* 9 (2025) 2, <https://doi.org/10.1038/s41698-024-00789-2>
- [5] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, in: *Medical Image Computing and Computer-Assisted Intervention (MICCAI) of LNCS*, vol. 9351, Springer, 2015, pp. 234–241, https://doi.org/10.1007/978-3-319-24574-4_28
- [6] W. Zhang, W. Jin, S. Rho, F. Jiang, C.-F. Yang, A federated learning framework for brain tumor segmentation without sharing patient data, *Int. J. Imaging Syst. Technol.* 34 (2024) e23147, <https://doi.org/10.1002/ima.23147>
- [7] K.S. Adewole, E. Alozie, H. Olagunju, N. Faruk, R.Y. Aliyu, A.L. Imoize, A. Abdulkarim, Y.O. Imam-Fulani, S. Garba, B.A. Baba, et al., A systematic review and meta-data analysis of clinical data repositories in Africa and beyond: recent development, challenges, and future directions, *Discover Data* 2 (2024) 8, <https://doi.org/10.1007/s44248-024-00012-4>
- [8] H.B. McMahan, E. Moore, D. Ramage, S. Hampson, B.A. Y Arcas, Communication-efficient learning of deep networks from decentralized data, *arXiv preprint arXiv:1602.05629*, 2023.
- [9] C. Shranthika, P. Saedi, I.V. Bajić, Decentralized learning in healthcare: a review of emerging techniques, *IEEE Access* 11 (2023) 54188–54209, <https://doi.org/10.1109/ACCESS.2023.3281832>
- [10] C.B. Raggio, M.K. Zabaleta, N. Skupien, O. Blanck, F. Cicone, G.L. Cascini, P. Zaffino, L. Migliorelli, M.F. Spadea, FedSynthCT-Brain: a federated learning framework for multi-institutional brain MRI-to-CT synthesis, *Comput. Biol. Med.* 192 (2025) 110160, <https://doi.org/10.1016/j.combiomed.2025.110160>
- [11] C.B. Raggio, P. Zaffino, M.F. Spadea, A Privacy-Preserving federated learning framework for generalizable CBCT to synthetic CT translation in head and neck, *arXiv preprint arXiv:2506.08654* 2025
- [12] S.S. Sandhu, H.T. Gorji, P. Tavakolian, K. Tavakolian, A. Akhbardeh, Medical imaging applications of federated learning, *Diagnostics* 13 (2023) 3140, <https://doi.org/10.3390/diagnostics13193140>
- [13] C. Ren, H. Yu, H. Peng, X. Tang, B. Zhao, L. Yi, A.Z. Tan, Y. Gao, A. Li, X. Li, Z. Li, Q. Yang, Advances and open challenges in federated foundation models, *IEEE Commun. Surv. Tutorials* 28 (2026) 2087–2126, <https://doi.org/10.1109/COMST.2025.3552524>
- [14] S. Gallus, A. Mercer, P. Singh, D. Cho, Optimization, communication, and personalization in federated learning for massive networks, *Preprints* (2025), <https://doi.org/10.20944/preprints202507.1037.v1>
- [15] A. Yang, J. Liu, B. Yang, Z. Mo, K. Li, Balancing communication overhead and accuracy in compression integration: a survey, *J. Supercomput.* 81 (2025) 964, <https://doi.org/10.1007/s11227-025-07451-z>
- [16] A.N. Sahu, A. Dutta, A.M. Abdelmoniem, T. Banerjee, M. Canini, P. Kalnis, Rethinking gradient sparsification as total error minimization, in: *NeurIPS 2021 - Advances in Neural Information Processing Systems*, 2021, <https://arxiv.org/abs/2108.00951>
- [17] D. Alistarh, D. Grubic, J.Z. Li, R. Tomioka, M. Vojnovic, Qsgd: communication-efficient SGD via gradient quantization and encoding, in: *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, Curran Associates Inc., Red Hook, NY, USA, 2017, pp. 1707–1718.
- [18] W. Wen, C. Xu, F. Yan, C. Wu, Y. Wang, Y. Chen, H. Li, TernGrad: ternary gradients to reduce communication in distributed deep learning, in: *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, Curran Associates Inc., Red Hook, NY, USA, 2017, pp. 1508–1518.
- [19] L. Wang, W. Wu, J. Zhang, H. Liu, G. Bosilca, M. Herlihy, R. Fonseca, FFT-based gradient sparsification for the distributed training of deep neural networks, in: *Proceedings of the 29th International Symposium on High-Performance Parallel and Distributed Computing, HPDC '20*, Association for Computing Machinery, New York, NY, USA, 2020, pp. 113–124, <https://doi.org/10.1145/3369583.3392681>
- [20] A.N. Onaizah, Y. Xia, A.J. Obaid, K. Hussain, Deep learning based brain tumour architecture for weight sharing optimization in federated learning, *Expert Syst.* 42 (2025) e13643, <https://doi.org/10.1111/essy.13643>
- [21] P.S. Tejaswini, J. Thirveni, K.R. Venugopal, A novel SLCA-UNet architecture for automatic MRI brain tumor segmentation, *Biomed. Signal Process. Control.* 100 (2025) 107047, <https://doi.org/10.1016/j.bspc.2024.107047>
- [22] K.G. Revathi, M.C.P. Shirley, S. Sreethar, et al., Brain tumor segmentation using optimized depth wise separable convolutional neural network with dense U-Net, *Knowl.-Based Syst.* (2025) 113678, <https://doi.org/10.1016/j.knsys.2025.113678>
- [23] A. Shaikh, S. Amin, M.A. Zeb, A. Sulaiman, M.S. Al Reshan, H. Alshahrani, Enhanced brain tumor detection and segmentation using densely connected convolutional networks with stacking ensemble learning, *Comput. Biol. Med.* 186 (2025) 109703, <https://doi.org/10.1016/j.combiomed.2025.109703>
- [24] M. Obayya, A. Alshuhail, K. Mahmood, M.H. Alanazi, M. Alqahtani, N.O. Aljehane, H. Almansour, M.A. Al-Hagery, A novel u-net model for brain tumor segmentation from MRI images, *Alex. Eng. J.* 126 (2025) 220–230, <https://doi.org/10.1016/j.aej.2025.04.051>
- [25] X. Guo, Z. Wang, P. Wu, Y. Li, F.E. Alsaadi, N. Zeng, Elts-net: an enhanced liver tumor segmentation network with augmented receptive field and global contextual information, *Comput. Biol. Med.* 169 (2024) 107879, <https://doi.org/10.1016/j.combiomed.2023.107879>
- [26] N. Zeng, H. Li, Z. Wang, W. Liu, S. Liu, F.E. Alsaadi, X. Liu, Deep-reinforcement-learning-based images segmentation for quantitative analysis of gold immunochromatographic strip, *Neurocomputing* 425 (2021) 173–180, <https://doi.org/10.1016/j.neucom.2020.04.001>, <https://www.cienccedirect.com/science/article/pii/S09252323220035385>
- [27] M.E. Yahiaoui, M. Dourdour, R. Abdulghafor, S. Turaev, M. Gasm, A. Bennour, A. Aborujilah, M. Sarem al, Federated learning with privacy preserving for multi-institutional three-dimensional brain tumor segmentation, *Diagnostics* 14 (2024) 2891, <https://doi.org/10.3390/diagnostics14242891>
- [28] E. Albalawi, T.R. Mahesh, A. Thakur, V.V. Kumar, M. Gupta, S.B. Khan, A. Almusharraf, Integrated approach of federated learning with transfer learning for classification and diagnosis of brain tumor, *BMC Med. Imaging* 24 (2024) 110, <https://doi.org/10.1186/s12880-024-01261-0>
- [29] J. Dai, H. Wu, H. Liu, L. Yu, X. Hu, X. Liu, D. Geng, FedATA: adaptive attention aggregation for federated self-supervised medical image segmentation, *Neurocomputing* 613 (2025) 128691, <https://doi.org/10.1016/j.neucom.2024.128691>
- [30] R. Kumar, C.M. Bernard, A. Ullah, R.U. Khan, J. Kumar, D.K.B. Kulevome, R. Yunbo, S. Zeng, Privacy-preserving blockchain-based federated learning for brain tumor segmentation, *Comput. Biol. Med.* 177 (2024) 108646, <https://doi.org/10.1016/j.combiomed.2024.108646>
- [31] G. Appasami, N. Savarimuthu, Federated learning for secure medical MRI brain tumor image classification, *Eur. Phys. J. Spec. Top.* (2025) 1–15, <https://doi.org/10.1140/epjs/s11734-025-01516-z>
- [32] J.B. Awotunde, C.O. Abikoye, B. Brahma, E.T. Oladipupo, A. Bandyopadhyay, Federated learning augmented with convolutional neural networks for brain cancer classification, *Proc. Comput. Sci.* 258 (2025) 2617–2626, <https://doi.org/10.1016/j.procs.2025.04.523>
- [33] Z. Liu, F. Wu, Y. Wang, M. Yang, X. Pan, FedCL: federated contrastive learning for multi-center medical image classification, *Pattern Recognit.* 143 (2023) 109739, <https://doi.org/10.1016/j.patcog.2023.109739>
- [34] A. Kazlouski, I.M. Perez, F. Noor, M. Högerman, O. Ettala, T. Pahikkala, A. Airola, Towards practical federated learning and evaluation for medical prediction models, *Int. J. Med. Inform.* (2025) 106046, <https://doi.org/10.1016/j.ijmedinf.2025.106046>
- [35] P.N. Srinivasu, G.J. Lakshmi, S.C. Narahari, J. Shafi, J. Choi, M.F. Ijaz, Enhancing medical image classification via federated learning and pre-trained model, *Egypt. Inform. J.* 27 (2024) 100530, <https://doi.org/10.1016/j.eij.2024.100530>
- [36] Y.N. Tan, V.P. Tinh, P.D. Lam, N.H. Nam, T.A. Khoa, A transfer learning approach to breast cancer classification in a federated learning framework, *IEEE Access* 11 (2023) 27462–27476, <https://doi.org/10.1109/ACCESS.2023.3257562>
- [37] Y. Xu, J. Cheng, H. Xu, C. Guo, Y. Liao, Z. Yao, Towards layer-wise quantization for heterogeneous federated clients, *Comput. Netw.* 264 (2025) 111223, <https://doi.org/10.1016/j.comnet.2025.111223>
- [38] M. Cao, H. Wang, Y. Yuan, J. Lu, X. Cai, D. Yu, M. Zhao, FedMQ+: towards efficient heterogeneous federated learning with multi-grained quantization, *J. Syst. Archit.* (2025) 103460, <https://doi.org/10.1016/j.sysarc.2025.103460>
- [39] H. Hu, W. Du, Y. Li, Y. Wang, pFL-SBPM: a communication-efficient personalized federated learning framework for resource-limited edge clients, *Future Gener. Comput. Syst.* 171 (2025) 107849, <https://doi.org/10.1016/j.future.2025.107849>
- [40] H. Zhang, T. Wu, Z. Ma, F. Li, J. Liu, Dynamic layer-wise sparsification for distributed deep learning, *Futur. Gener. Comput. Syst.* 147 (2023) 1–15, <https://doi.org/10.1016/j.future.2023.04.022>
- [41] S. Bakas, A. Linardos, S. Pati, U. Baid, M. Sheller, B. Edwards, A. Karargyris, P. Mattson, B. Menze, M. Bilello, S. Mohan, J.B. Freymann, J.S. Kirby, C. Davatzikos, H. Fathallah-Shaykh, R. Wiest, A. Jakab, R.R. Colen, A. Kotrotsou, D. Marcus, M. Milchenko, A. Nazeri, M.-A. Weber, A. Mahajan, The federated tumor segmentation (FeTS) challenge 2024, *Nat. Commun.* (2024), <https://doi.org/10.5281/ZENODO.10990500>
- [42] S. Bakas, et al., The multimodal brain tumor image segmentation benchmark (BRATS), *IEEE Trans. Med. Imaging* 41 (2022) 2773–2781, <https://doi.org/10.1109/TMI.2022.3194052>
- [43] M.J. Cardoso, W. Li, R. Brown, N. Ma, E. Kerfoot, Y. Wang, B. Murrey, A. Myronenko, C. Zhao, D. Yang, V. Nath, Y. He, Z. Xu, A. Hatamizadeh, A.

- Myronenko, W. Zhu, Y. Liu, M. Zheng, Y. Tang, I. Yang, M. Zephyr, B. Hashemian, S. Alle, M.Z. Darestani, C. Budd, M. Modat, T. Vercauteren, G. Wang, Y. Li, Y. Hu, Y. Fu, B. Gorman, H. Johnson, B. Genereaux, B.S. Erdal, V. Gupta, A. Diaz-Pinto, A. Dourson, L. Maier-Hein, P.F. Jaeger, M. Baumgartner, J. Kalpathy-Cramer, M. Flores, J. Kirby, L.A.D. Cooper, H.R. Roth, D. Xu, D. Bericat, R. Floca, S.K. Zhou, H. Shuaib, K. Farahani, K.H. Maier-Hein, S. Aylward, P. Dogra, S. Ourselin, A. Feng, MONAI: an open-source framework for deep learning in healthcare, arXiv preprint [arXiv:2211.02701](https://arxiv.org/abs/2211.02701), 2022.
- [44] P. Agrawal, N. Katal, N. Hooda, Segmentation and classification of brain tumor using 3d-unet deep neural networks, *Int. J. Cogn. Comput. Eng.* 3 (2022) 199–210, <https://doi.org/10.1016/j.ijcce.2022.11.001>, <https://www.sciencedirect.com/science/article/pii/S2666307422000213>.
- [45] A.F. Aji, K. Heafield, Sparse communication for distributed gradient descent, in: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2017, <https://doi.org/10.18653/v1/d17-1045>.
- [46] D.J. Beutel, T. Topal, A. Mathur, X. Qiu, J. Fernandez-Marques, Y. Gao, L. Sani, K.H. Li, T. Parcollet, P.P.B. de Gusmão, N.D. Lane, Flower: A friendly federated learning research framework, arXiv preprint [arXiv:2007.14390](https://arxiv.org/abs/2007.14390), 2022.
- [47] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: an imperative style, High-Performance deep learning library, arXiv preprint [arXiv:1912.01703](https://arxiv.org/abs/1912.01703), 2019.
- [48] A. Celaya, B. Riviere, D. Fuentes, A generalized surface loss for reducing the hausdorff distance in medical imaging segmentation, arXiv preprint [arXiv:2302.03868](https://arxiv.org/abs/2302.03868), 2024.
- [49] J. Li, Y. Zhang, Y. Li, X. Gong, W. Wang, Fedsparse: a communication-efficient federated learning framework based on sparse updates, *Electronics* 13 (2024) 5042, <https://doi.org/10.3390/electronics13245042>.
- [50] J. Jia, J. Liu, C. Zhou, H. Tian, M. Dong, D. Dou, Efficient asynchronous federated learning with sparsification and quantization, *Concurr. Comput. Pract. Exp.* 36 (2024) e8002, <https://doi.org/10.1002/cpe.8002>.

Author biography

Ciro Benito Raggio, M.Sc., is a Research Associate at the Karlsruhe Institute of Technology, Germany, where he is pursuing a PhD in Decentralized AI for Healthcare and Medical Imaging. He received his B.Sc. in Computer Science and Biomedical Engineering and his M.Sc. in Biomedical Engineering from the Magna Graecia University of Catanzaro, Italy. His research interests include federated learning and privacy-preserving AI for decentralized medical image-to-image translation, with applications in synthetic CT generation and computer-aided diagnosis. He has contributed to open-source tools, peer-reviewed publications, and international AI collaborations.

Asaf Raza received his PhD in Information and Communication Technologies from the University of Calabria (UniCal), Italy, where his research focused on Federated Learning and Radiomics for privacy-preserving medical image analysis, particularly in brain tumor imaging. He previously earned his Master's degree in Software Engineering from the University of Engineering and Technology (UET) Taxila, Pakistan, with a research focus on brain tumor classification using deep learning. His research interests include federated learning, radiomics, and AI-driven medical imaging.

Antonella Guzzo is an Associate Professor of Computer Engineering at the University of Calabria, Italy. She holds a Ph.D. in Computer Engineering (2005) and previously worked as a research fellow at ICAR-CNR. Her research focuses on process mining, data mining, AI, and deep learning. She has published over 75 papers, serves as a reviewer for numerous international journals, and is active in program committees for major conferences. Since 2005, she has contributed to several national and international research projects. She is a cofounder of the IEEE Task Force on Process Mining and a member of the IEEE Working Group for the XES standard.

Maria Francesca Spadea is Chair of Modelling and Simulation for Medical Engineering at Karlsruhe Institute of Technology (Germany), where she co-leads the Institute of Biomedical Engineering. She earned her MSc in Mechanical Engineering from the University of Calabria (2002) and her PhD in Bioengineering from Politecnico di Milano (2006). Between 2006 and 2015, she was a Visiting Professor at the Massachusetts Institute of Technology and Harvard Medical School, where she was also a Fulbright Research Scholar (2010–2011), focusing on metal artefact reduction in CT imaging. In the past 15 years, she has established extensive national and international collaborations across academia, clinical institutions, and industry. Her research focuses on developing and validating computational methods to support diagnosis and therapy, including medical image processing and analysis, surgical navigation and advanced image guided radiotherapy.

Giancarlo Fortino is a Full Professor of Computer Engineering at the University of Calabria, Italy, and a distinguished professor at multiple Chinese universities. He is a senior research fellow at ICAR-CNR, a CAS PIFI visiting scientist, and an IEEE Sensors Council Distinguished Lecturer. His research focuses on wearable systems, e-health, IoT, agent-based computing, and AI-driven human-machine systems. Prof. Fortino has chaired 130+ conferences/workshops, served on 700+ program committees, and guest-edited over 80 special issues. He has also organized numerous high-impact events, including recent workshops on GenAI for IoT (WF-IoT 2024), vehicle formation control (IEEE CASE 2024), and distributed collective intelligence (ICDCS 2022).

PART IV

TOOLING, REPRODUCIBILITY AND TRANSLATION

ImageAugmenter: A user-friendly 3D Slicer tool for medical image augmentation

9.1 Introduction

This chapter presents *ImageAugmenter*, a user-friendly and open-source module for the 3D Slicer Imaging Computing Platform [107, 146], designed to make medical image augmentation widely accessible. Augmentation has emerged as a primary strategy to address one of the most persistent challenges in medical imaging (i.e. the limited availability of annotated data for training Deep Learning (DL) models). By generating diverse and realistic variations of existing images, augmentation improves model generalisation, reduces overfitting, and supports more robust performance across a range of clinical tasks, including anatomical segmentation, disease classification, image-to-image translation and computer-aided diagnosis [146].

While existing DL frameworks –such as PyTorch [147] and MONAI [148]– offer powerful augmentation pipelines, they require programming expertise, creating a barrier for researchers and clinicians who lack coding skills. ImageAugmenter overcomes this limitation by providing a graphical interface for applying a variety of transformations, covering spatial, intensity, and noise-based transformations.

The functionality of ImageAugmenter has been evaluated using multiple public medical imaging datasets of different dimensions, covering diverse imaging modalities. The validation focused on three main aspects: (i) usability and effectiveness in applying transformations, (ii) comparison with other available tools, and (iii) computational performance.

In detail, ImageAugmenter was compared to the TorchIO 3D Slicer extension, revealing several key distinctions. Unlike the TorchIO extension, which provides only 10 basic transformations and requires loading manually each volume into 3D Slicer, ImageAugmenter offers more than 20 transforms applicable simultaneously to entire datasets, including images

and segmentation masks. Augmented datasets can be directly exported on the operating system, preserving the original file hierarchy. Users can preview augmented images in real-time, customise transformation parameters, and export the resulting datasets for immediate use in DL and Federated Learning (FL) workflows. Furthermore, the ImageAugmenter extension supports GPU acceleration.

This design allows a broad audience to leverage state-of-the-art augmentation techniques without writing code, fostering a more inclusive and efficient research environment.

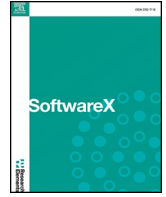
Performance benchmarks for previewing and augmenting images were measured across datasets of increasing size and complexity, using combinations of up to eight transformations applied simultaneously. Results demonstrated consistent output generation proportional to the number of transformations, with preview times remaining short due to application on a single representative volume. Execution times increased with experimental complexity, showing no unexpected behaviors. GPU acceleration provided benefits mainly when processing a large number of images and transformations, whereas for simpler tasks CPU-only execution was competitive.

Beyond its immediate utility in single-site DL studies, *ImageAugmenter* also holds potential for FL applications. In healthcare FL, distributed training across institutions often faces challenges such as data heterogeneity, imbalance, and scarcity –particularly in smaller clinics– which can limit model performance and generalisation. By enabling straightforward, local augmentation of medical images, ImageAugmenter allows each federated client to generate richer and more diverse datasets, enhancing the robustness of the federated model while maintaining strict data privacy.

It should be noted that the manuscript included in this chapter describes an early version of ImageAugmenter (1.0.4), which has since evolved to incorporate additional functionality, performance and user interface improvements. Nevertheless, the original publication captures the essential methodological contributions, usability principles, and the potential impact on both conventional and federated DL pipelines.

The content of this chapter has been adapted from a manuscript previously published in the peer-reviewed journal “SoftwareX”. Portions of the text are reproduced verbatim from the published article [146]. The author of this thesis led the conceptualisation, design, implementation, and evaluation of the extension. Furthermore, the author led the drafting and revision of the original manuscript in close collaboration with the co-authors.

9.2 Manuscript



ImageAugmenter: A user-friendly 3D Slicer tool for medical image augmentation

Ciro Benito Raggio^{a,*}, Paolo Zaffino^{b,1}, Maria Francesca Spadea^{a,1}

^a Karlsruhe Institute of Technology, Fritz-Haber-Weg 1, Karlsruhe 76131, Germany

^b Magna Graecia University of Catanzaro, Viale Europa, Catanzaro 88100, Italy

ARTICLE INFO

Keywords:
Medical imaging
Augmentation
3D Slicer
Deep learning

ABSTRACT

Limited medical image data hinders the training of deep learning (DL) models in the biomedical field. Image augmentation can reduce the data-scarcity problem by generating variations of existing images. However, currently implemented methods require coding, excluding non-programmer users from this opportunity.

We therefore present *ImageAugmenter*, an easy-to-use and open-source module for 3D Slicer imaging computing platform. It offers a simple and intuitive interface for applying over 20 simultaneous MONAI Transforms (spatial, intensity, etc.) to medical image datasets, all without programming.

ImageAugmenter makes accessible medical image augmentation, enabling a wider range of users to improve the performance of DL models in medical image analysis by increasing the number of samples available for training.

Metadata

Nr	Code metadata description	Please fill in this column
C1	Current code version	v1.0.4 (2024/09/03)
C2	Permanent link to code/repository used for this code version	https://github.com/ciroraggio/SlicerImageAugmenter/releases/tag/v1.0.4
C3	Permanent link to reproducible capsule	
C4	Legal code license	MIT License
C5	Code versioning system used	Git
C6	Software code languages, tools and services used	Python, CMake
C7	Compilation requirements, operating environments and dependencies	3D Slicer, PyTorch, MONAI, SimpleITK, munch
C8	If available, link to developer documentation/manual	https://ciroraggio.github.io/SlicerImageAugmenter/
C9	Support email for questions	ciro.raggio@kit.edu

1. Motivation and significance

In the field of medical image analysis, deep learning (DL) models have achieved remarkable success in tasks like disease classification, tissue and organ segmentation, and image registration [1]. However, two major barriers in the development of these models are the limited

availability of medical image data, and the increasingly stringent regulations that make the sharing of medical images, even for research purposes, more complex. This scarcity of data hinders the training process, often leading to models that overfit the training data and perform poorly on unseen data. Data augmentation has been proposed by several authors to address this issue [2–5].

Image augmentation is a technique that synthetically expands the training dataset by manipulating existing images or by generating artificial samples from scratch [6]. In the first, and most common case, the augmentation can include geometric transformations (translations, rotations, flips, scaling, deformation, zooming), intensity variations (histogram based operations), and noise injection [7].

Augmented data helps the DL model to learn features that are invariant to these alterations and improves the generalization capabilities of the model. Current augmentation techniques are often based on custom code written in programming languages such as Python. While there are existing tools and frameworks that address medical image augmentation, such as Medical Imaging Toolkit (MITK) [8,9], Pillow [10], PyTorch [11], TorchIO [12], and MONAI [13], they all have one thing in common: they require programming knowledge to be effectively used. This excludes researchers and medical professionals, who may lack programming skills, from exploiting the benefits of image augmentation. It is also time consuming for investigators to implement

* Corresponding author at: Institute of Biomedical Engineering, Karlsruhe Institute of Technology, Fritz-Haber-Weg 1, Karlsruhe 76131, Germany.

E-mail address: ciro.raggio@kit.edu (C.B. Raggio).

¹ These authors contributed equally to this work and share senior authorship.

different kinds of augmentation before finding the optimal set that maximizes model performance.

ImageAugmenter addresses this challenge by providing a user-friendly tool for medical image augmentation and enabling the creation of richer and more diverse training datasets, which can directly improve the performance of DL models.

This translates to increased accuracy, generalizability, and robustness for tasks like cancer detection, segmentation of anatomical structures, and computer-aided diagnosis.

ImageAugmenter is integrated within the open source 3D Slicer [14] platform as a module of a new extension, designated as "Slicer-ImageAugmenter." 3D Slicer is widely used in scientific and clinical settings. Its popularity is evidenced by a high number of downloads (over 1.5 million since 2011)¹ and a large community of active users.

Using the module is simple and intuitive. The user must:

1. install the extension through 3D Slicer's extension management module;
2. select the data to be augmented, where the module supports several medical image formats, including DICOM and NIfTI with two different types of folder hierarchies;
3. select the MONAI Transforms to be applied;
4. perform data augmentation; here, several parameters to customize data augmentation can be selected;
5. visualize and analyze the augmented data.

The integration with 3D Slicer makes the tool easily accessible to a wide range of users, facilitating the use of advanced data augmentation techniques for medical image retrieval and analysis. Moreover, augmented images can be also visualized through 3D Slicer panels for easily checking the artificial data space.

2. Software description

The module leverages the capabilities of PyTorch [11] and MONAI [13] to perform data augmentation on medical image datasets. This process involves applying a series of specific MONAI transformations to each image within the dataset, creating a larger and more diverse collection of samples for training DL models. In light of the mentioned points, all forthcoming references to the term "transformation" should be understood in a more precise manner as "MONAI Transform".

2.1. Software architecture

The ImageAugmenter module is structured into several classes, including those provided by the 3D Slicer infrastructure for building scripted modules². These include the ImageAugmenterWidget, which is responsible for managing the graphical user interface (GUI) of the module, and the ImageAugmenterLogic, which handles the general logic of the module. Additionally, a set of custom classes and Python modules are grouped in the library named ImageAugmenterLib, which allows for the creation of a modular architecture and serves as the core of the module.

The ImageAugmenterTransformationParser class is responsible for coordinating the controller classes and collecting the list of MONAI Transforms in accordance with the user inputs.

The ImageAugmenterTransformControllerInterface, located in the ImageAugmenterLib, is the abstract class upon which three distinct classes are then built: ImageAugmenterSpatialController, ImageAugmenterIntensityController and ImageAugmenterCropController. The aim of the controller classes is to be specialized in mapping different types of MONAI Transforms, ensuring compatibility with the dataset class. The ImageAugmenterDataset class implements the Dataset abstract class from PyTorch and handles dataset loading and MONAI Transforms application.

In addition to the mentioned classes, two further Python modules are

included in the ImageAugmenterLib. The ImageAugmenterUtils Python module contains essential functions, such as saving to disk and previewing the result, plus other minor utility functions. The ImageAugmenterValidator Python module, on the other hand, contains functions used in the initial phase to validate the fields in the GUI.

This modular approach, which combines the standard Slicer module structure with a set of custom classes, provides a foundation for future enhancements and additions to the module's capabilities.

2.2. Software functionalities

ImageAugmenter enables users to define and apply different image augmentation transformations to their own medical image datasets, with the ability to maintain the same input hierarchy and ensure that spatial correspondence between image and mask is maintained through the applied transformations.

With a user-friendly graphical interface, the module allows users to select input and output directories, specify image and mask prefixes, define desired transformations through a comprehensive set of options, and choose the device (CPU or PyTorch compatible GPUs) on which to execute the computation. It also provides real-time progress visualization and informative status updates.

The module includes a "preview" functionality. Users can select a subset of transformations and visualize the resulting augmented images alongside their original counterparts. This facilitates the refinement of the selected transformations to achieve the optimal outcome for their specific use case. Furthermore, this approach avoids occupying memory on the disk and facilitates the user's understanding of the model's input during training. As a consequence, time can be invested in further stages of research because a data sanity check has already been performed.

In the current version, 7 spatial transformations, 12 intensity transformations, 2 crop transformations and 2 pad transformations are available. Nevertheless, our efforts have been concentrated on ensuring that the integration of further MONAI Transforms is straightforward, largely due to the code design.

Moreover, the majority of parameters that users are required to configure for each transformation are identical in terms of nomenclature and functionality to those found in the MONAI library. This approach streamlines the referencing of MONAI documentation and standardizes the augmentation pipeline. It ensures that every detail of the transformations in the module is reflected and directly accessible in the official MONAI documentation.

3. Illustrative examples

Fig. 1 illustrates the *ImageAugmenter* module. The user is first prompted to specify the path to the images to be augmented, file hierarchy information, and file name patterns that will allow the detection of images and masks (if any) within the specified path. The user is then prompted to specify a path where the augmented images will be stored.

Subsequently, the user is given the option to select any MONAI Transforms to be applied, as well as tune their parameters. If the underlying hardware provides a PyTorch-compatible GPU, the user can also select the device on which the augmentation will be performed.

Finally, the user can choose to preview or save the augmented images.

Fig. 2 shows an example of how the 3D Slicer scene looks like when using the "Preview" feature of *ImageAugmenter*. The inclusion of a preview function within the 3D Slicer module increases the user's control over the augmentation process. By allowing real-time tuning of parameters within the augmentation panel, users will be able to establish the result of the MONAI Transforms used by checking the images slice by slice from the 3D Slicer infrastructure.

Finally, Listing 1 shows how *ImageAugmenter* exports augmented images with different dataset types. It is appreciable how the module maintains the input hierarchy, clearly adding the type of MONAI

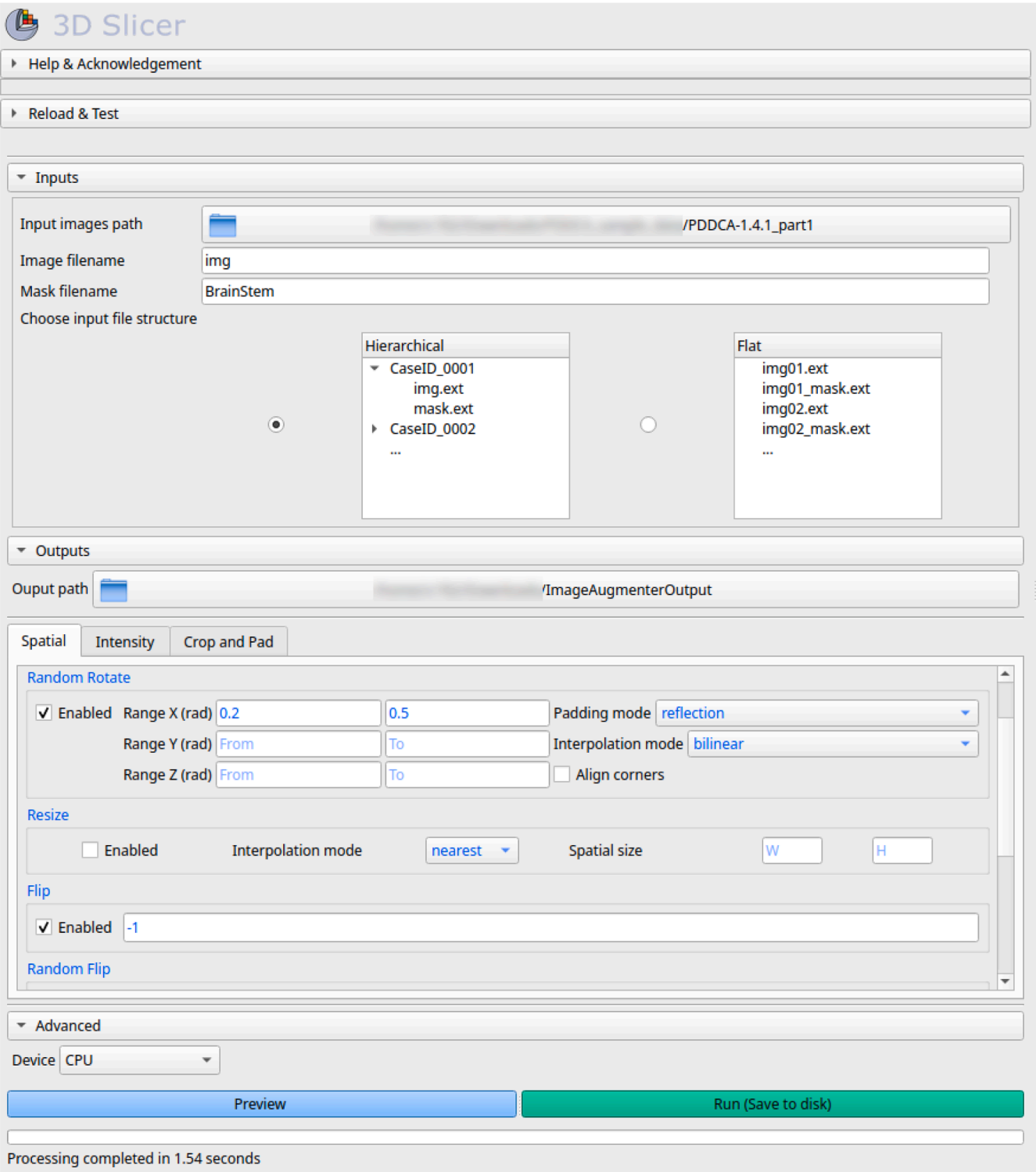


Fig. 1. ImageAugmenter graphical user interface.

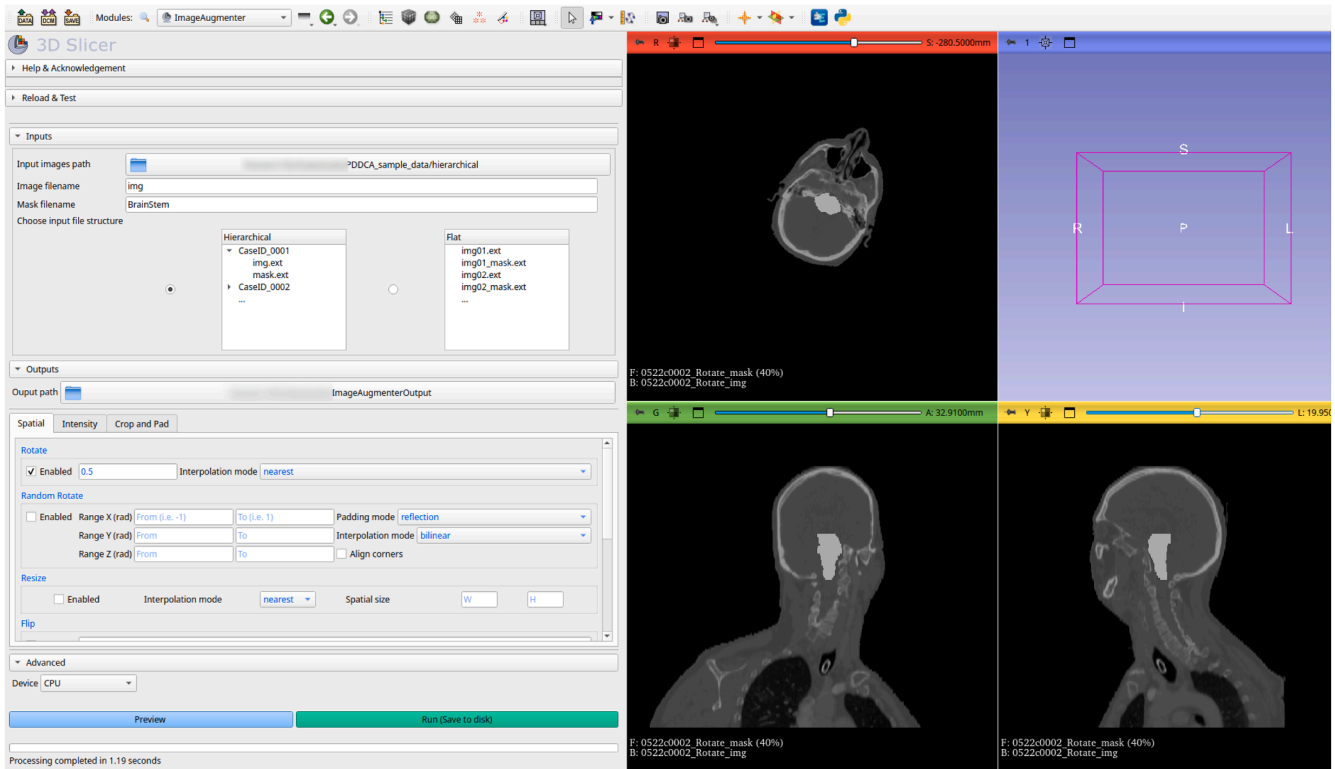


Fig. 2. Preview mode.

Transform applied to a specific case.

```
HierarchicalDataset/
├── 0522c0013/
│   ├── img.nrrd
│   └── BrainStem.nrrd
├── 0522c0014/
│   ├── img.nrrd
│   └── BrainStem.nrrd
└── ...

ImageAugmenterOutput/
├── 0522c0013_Flip/
│   ├── img.nrrd
│   └── BrainStem.nrrd
├── 0522c0013_Rotate/
│   ├── img.nrrd
│   └── BrainStem.nrrd
├── 0522c0014_Flip/
│   ├── img.nrrd
│   └── BrainStem.nrrd
├── 0522c0014_Rotate/
│   ├── img.nrrd
│   └── BrainStem.nrrd
└── ...

FlatDataset/
├── img001.nrrd
├── label001.nrrd
├── img002.nrrd
├── label002.nrrd
└── ...

ImageAugmenterOutput/
├── img001_Flip.nrrd
├── label001_Flip.nrrd
├── img001_Rotate.nrrd
├── label001_Rotate.nrrd
├── img002_Flip.nrrd
├── label002_Flip.nrrd
├── img002_Rotate.nrrd
├── label002_Rotate.nrrd
└── ...
```

Listing 1. Input and output example. The two hierarchies of inputs that are supported are illustrated on the left, and the respective generated outputs are shown on the right.

4. Impact

Medical data augmentation has emerged as a powerful technique to enhance the accuracy and robustness of various deep learning-based tasks in medical image analysis [15].

The application of ImageAugmenter can provide a meaningful contribution in addressing key research questions and challenges in medical deep learning. As evidenced by studies referenced in [16,17], data augmentation is an effective method for enhancing the generalization and robustness of medical image analysis models across different imaging modalities, scanner types, and patient demographics. Addressing this issue is essential to ensure that models perform effectively on previously unseen data and are not susceptible to bias. Furthermore, medical imaging datasets often suffer from class imbalance, where certain disease classes are underrepresented. Thoughtful application of augmentation methods can mitigate this issue, particularly in tasks like lesion detection or segmentation [18]. By augmenting images from minority classes, the dataset becomes more balanced, leading to improved model performance.

It is also important to consider that medical imaging is prone to artifacts such as motion blur, noise, or intensity inhomogeneities [19]. A key consideration is how augmentation techniques can be used to simulate these imaging artifacts, thereby training models that are more resilient to such common challenges in real-world medical imaging. This will improve the ability of the model to deal effectively with noisy or imperfect data.

Finally, in order to make it clear the impact that ImageAugmenter module can have on the research community, we would like to point out how that Goceri’s review [15] examined very comprehensively augmentation techniques that have been applied to improve the performance of deep learning-based diagnosis on different organs (brain, lung, breast, and eye), using different imaging modalities (MRI, CT, mammography, and fundoscopy). Results reported in Tables 1–4 of the review [15], showed that 56 out of the 77 studies proposed in the tables, involved the use of transformations (such as translation, scaling, flipping, brightness changing, rotation, mirroring, zooming, resizing, etc), as augmentation methods, improving performance across the wide range of tasks to which it was applied.

In light of this, the positive implications of ImageAugmenter for the scientific community and upcoming research efforts are obvious. Moreover, a strong foundation for the adoption of this module is the widespread use of 3D Slicer, both within and beyond the intended user group.

4.1. Comparison with other available tools

The ImageAugmenter module within the 3D Slicer environment is only comparable to the TorchIO 3D Slicer module³ in terms of functionality. TorchIO is a Python package based on PyTorch for reading, preprocessing, augmenting, and writing 3D medical images for deep learning applications. It offers intensity and spatial transforms for data augmentation and preprocessing. The developers of TorchIO library describe the TorchIO 3D Slicer module as a tool that allows users to rapidly visualize the impact of each transformation parameter, thereby

providing an intuitive understanding of the output without the need for coding⁴. Although the underlying concept may seem similar, several important differences emerge between the TorchIO 3D Slicer module and ImageAugmenter. The following section will describe the distinctions between the two modules and illustrate the enhancements that we have implemented to enhance the overall user experience.

The first distinction between the two modules is related to their foundational frameworks. Our module is based on MONAI, while TorchIO 3D Slicer module is based on the TorchIO package. MONAI provides a comprehensive suite of high-level Transform classes, particularly in the domain of spatial transformations, which are not present in TorchIO. Consequently, this represents an important drawback for the 3D Slicer TorchIO module, as it necessitates additional effort from users to implement functionality comparable to the advanced Transform classes available in MONAI.

Another difference lies in the number of transformations that the two modules offer to the user and the usability of these transformations. Our module provides more than 20 MONAI Transforms customizable and applicable to an entire dataset of images and masks simultaneously. The 3D Slicer TorchIO module offers approximately 10 TorchIO transformations, the majority of which are random and basic, such as RandomAffine.

It is also crucial to highlight the distinction in the manner in which our module is integrated in the 3D Slicer environment and how it interacts with the operating system. Unlike the TorchIO 3D Slicer module, the user is not required to load the images into the 3D Slicer environment to apply the augmentation. Our module enables the loading of datasets directly from the operating system, supporting two distinct hierarchies. It is possible to apply multiple transformations simultaneously, preview the transformations, and then automatically save the results with the same hierarchy and file names as the input dataset, enhancing and accelerating the user experience.

On the other hand, the TorchIO 3D Slicer module enables the application of a single TorchIO transformation to a single volume at a time and it is not possible to save the augmented volume directly on the operating system with the same input hierarchy. Consequently, further interaction with 3D Slicer is required to export the augmented volume and repeat the process for the entire dataset.

To illustrate more clearly the differences between ImageAugmenter and the TorchIO 3D Slicer module, a comparison has been proposed in Table 1. The table highlights the key features and capabilities of both modules, providing an overview of their respective strengths and limitations.

4.2. Performance

The analysis of the tool’s performance is based on a series of different benchmarks. The benchmarks evaluated the time required to preview and save images as the image size, the number of images, and the number of image transformations increased. The indicated preview and save times are averaged on three consecutive runs. The tests were conducted with both CPU and GPU to provide a more comprehensive evaluation and perspective of the tool.

Table 1
Differences between ImageAugmenter and the TorchIO 3D Slicer module.

Feature	ImageAugmenter	TorchIO 3D Slicer Module
Base Framework	MONAI	TorchIO
Transforms #	24 MONAI transforms	10 TorchIO transforms
Transforms Parameters	Customizable	Customizable
Dataset Loading	Direct from operating system, supports two different hierarchies	Requires loading the data into 3D Slicer as Volumes in order to apply the augmentation
Augmentation	Can apply multiple transforms to the entire dataset, on the image and the related segmentation/mask (if present) simultaneously	Applies a single transformation to a single volume at a time
Preview	Available for all the MONAI Transforms chosen on a volume of the dataset	Available for the volume on which the transform is being applied
Saving	The augmented dataset is automatically exported with the same hierarchy and file names as the original, directly to the operating system	Additional interactions with 3D Slicer are required for the exportation of augmented volumes to the operating system manually

The hardware used for these tests included:

- CPU: Intel(R) Xeon(R) Gold 6230 CPU @ 2.10GHz
- GPU: NVIDIA Corporation GA104GL [RTX A4000] (16GB)
- RAM: 12 × 16GB DDR4 DIMM

Furthermore, the number of cases and the amount of input files, as

The following section provides a summary of the datasets (Table 2), the types of MONAI Transforms applied (Table 3), and the parameters used to run the benchmarks. The results obtained for each benchmark are presented in Table 4.

Dataset

Table 2

Description of the datasets and input files used in the benchmark.

DatasetID	Description	Size	Input image file	Input segmentation file	Patients (#)	Files in total (#)
D1	PDDCA v1.4.1 (part 1)	512 × 512 × 138 circa	img.nrrd	BrainStem.nrrd	12	48
D2	SynthRAD2023 - Task1 - Brain - Validation set	220 × 220 × 200 circa	mr.nii.gz	mask.nii.gz	30	60
D3	SynthRAD2023 - Task1.2 - Brain - Train set	220 × 220 × 200 circa	ct.nii.gz	mask.nii.gz	179	359

well as the corresponding cases and output files, are presented to assess the consistency of the savings.

To ensure reproducibility, three different public datasets of different

Transformations

Table 3

Description of the transformations and specific parameters used in the benchmark.

TransformsID	Description	Settings
T1	A. Rotate B. RandomFlip	A. Angle: 0.5rad; interpolation mode: nearest B. -
T2	A. Rotate (angle: 0.5rad, interpolation mode: nearest) B. RandomFlip C. Zoom	A. Angle: 0.5rad; interpolation mode: nearest B. - C. Factor: 1.2; interpolation mode: area; padding mode: edge
T3	D. BorderPad A. Rotate B. RandomFlip C. Zoom D. BorderPad E. CenterSpatialCrop F. GaussianSmooth G. RandomZoom H. RandomRotate	D. Spatial border: 10; mode: symmetric A. Angle: 0.5rad; interpolation mode: nearest B. - C. Factor: 1.2; interpolation mode: area; padding mode: edge D. Spatial border: 10; mode: symmetric E. Height: 200; width: 200 F. Factor: 1.0 G. From 0.5 to 0.9; interpolation mode :area; padding mode:edge H. On the X axis; from 0.6rad to 0.9rad; interpolation mode: nearest; padding mode:reflection

sizes were used:

- PDDCA v1.4.1 [20] with 12 images averaging 512 × 512 × 138 in size.
- SynthRAD2023 [21] challenge dataset for Task 1, using only the validation set with 30 images of size 220 × 220 × 200 (average) .
- SynthRAD2023 challenge dataset for Task 1.2, using only the training set with 179 images of size 220 × 220 × 200 (average).

5. Results

It is important to note that the results obtained (Table 4) depend on the hardware used and should therefore be understood as a non-absolute indicator of the tool performance.

The results prove the consistency in terms of output, with a number

Table 4

Summary of tests performed, devices used, time taken and files exported.

Test #	Device	DatasetID	TransformsID (number of actual transformations)	Preview time (s)	Saving time (s)	Augmented patients (#)
1	CPU	D1	T1 (2)	2.37	24.97	24
2	CPU	D1	T2 (4)	3.46	36.06	48
3	CPU	D1	T3 (8)	5.55	57.87	96
4	CPU	D2	T1 (2)	1.05	21.2	60
5	CPU	D2	T2 (4)	1.55	28.98	120
6	CPU	D2	T3 (8)	2.47	44.99	240
7	CPU	D3	T1 (2)	1.07	122.58	358
8	CPU	D3	T2 (4)	1.72	188.61	716
9	CPU	D3	T3 (8)	2.55	274.12	1432
10	GPU	D1	T1 (2)	2.48	24.17	24
11	GPU	D1	T2 (4)	3.89	40.13	48
12	GPU	D1	T3 (8)	6.04	56.31	96
13	GPU	D2	T1 (2)	1.05	20.02	60
14	GPU	D2	T2 (4)	1.6	30.9	120
15	GPU	D2	T3 (8)	2.62	45.26	240
16	GPU	D3	T1 (2)	1	132.23	358
17	GPU	D3	T2 (4)	1.73	217.11	716
18	GPU	D3	T3 (8)	2.5	265.95	1432

of new cases and files generated that are proportional to the transformations used. With regard to the preview times, it can be observed that they are relatively short. This is because the transformations are applied to a single exemplary volume, which allows the parameters to be promptly fine-tuned until the user is satisfied, thus precisely obtaining the desired behavior.

The execution time increases in proportion to the complexity of the experiments, exhibiting no unexpected behaviors. Additionally, the results indicate that the use of the GPU does not consistently outperform the use of the CPU. This is because, for a few transformations, the time

required for the transfer of data between devices may exceed the benefits gained, and the system still relies on the CPU for disk I/O operations. Therefore, for simple tasks, using the GPU may result in a time penalty.

This is visible in the comparison between experiment #9 and experiment #18, executed respectively on CPU and GPU. In this case, the computation on the CPU takes 10 s less than the GPU implementation. This outcome suggests that the benefits of GPU usage are only evident when an important number of images and transformations are involved.



Fig. 3. Original image (bottom) compared to Rotate result (middle) and to Random Flip result (top).

Table 5
Comparison of original and transformed volumes.

	Dimensions	Spacing	Origin	Scan Order	Scalar type	Scalar Range
Original image	512 × 512 × 156	0.98mm x 0.98mm x 2.50mm	250.39mm x -155.61mm x -1178.00mm	Axial IS	float	-1024 to 2117.14
Rotated image	512 × 512 × 156	0.98mm x 0.98mm x 2.50mm	250.39mm x -155.61mm x -1178.00mm	Axial IS	float	-1024 to 2117.14
Flipped image	512 × 512 × 156	0.98mm x 0.98mm x 2.50mm	250.39mm x -155.61mm x -1178.00mm	Axial IS	float	-1024 to 2117.14

5.1. Exemplary output

This section illustrates how the MONAI Transforms applied with ImageAugmenter maintain anatomical consistency by comparing the original image (bottom) and the result obtained (top) by applying the T1 transformations (rotation and flip) to a case of the D1 dataset (Fig. 3). The original and the transformed volumes are then manually compared using the volume geometry metadata available in the "Volumes" module of 3D Slicer (Table 5).

6. Conclusions

In this paper, we presented *ImageAugmenter*, a 3D Slicer module that provides an intuitive user interface for executing data augmentation on medical images. It is designed to be flexible and versatile, supporting a wide range of MONAI Transforms and offering the users precise control over the transformation parameters. In addition, it guarantees the application of the same random transforms to each pair of anatomical and segmentation volumes, as well as the support of PyTorch compatible GPUs to decrease computation time.

The performance of ImageAugmenter was evaluated through a series of benchmarks that measured the time required for previewing and saving images as the image size, number of images, and transformations increased. These tests, conducted on both the CPU and the GPU, demonstrated that the number of generated cases and files was consistently proportional to the applied transformations. The preview times were short, allowing for quick parameter adjustments. It was found that the module does not alter dimensions, spacing, origin, scan order, scalar type and the scalar range of the volumes. Therefore, the consistency of the results depends on the user's parameter choices.

The tool allows users to visually assess the effect of data augmentation transformations and to save the augmented data to be used for analysis and AI models training. This 3D Slicer module would represent a powerful tool for researchers, clinicians, biologists, and any user who intends to use data augmentation techniques for improving the accuracy, generalizability, and robustness of their medical DL models, in a broad accessible way, without writing code and by simply using a graphical user interface.

Statement

During the preparation of this work the author(s) used Google's tool Gemini in order to improve language and readability. After using this tool, the author(s) reviewed and edited the content as needed and take (s) full responsibility for the content of the publication.

CRediT authorship contribution statement

Ciro Benito Raggio: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Paolo Zaffino:** Writing – review & editing, Visualization, Validation, Supervision, Software, Conceptualization, Methodology. **Maria Francesca Spadea:** Writing – review & editing, Supervision, Resources, Investigation, Data curation.

Declaration of competing interest

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

Acknowledgements

None.

Data availability

All data used to test the software are public and cited in the article.

References

[1] Chan HP, Samala RK, Hadjiiski LM, Zhou C. Deep learning in medical image analysis. *Adv Exp Med Biol* 2020;1213:3–21. https://doi.org/10.1007/978-3-030-33128-3_1.

[2] Yang YY, Ho MY, Tai CH, Wu RM, Kuo MC, Tseng YJ. FastEval Parkinsonism: an instant deep learning–assisted video-based online system for Parkinsonian motor symptom evaluation. *NPJ Digit Med* 2024;7(1). <https://doi.org/10.1038/s41746-024-01022-x>.

[3] Paciorek AM, Claudio FSC, et al. Automated assessment of cardiac pathologies on cardiac MRI using T1-mapping and late gadolinium phase sensitive inversion recovery sequences with deep learning. *BMC Med Imaging* 2024;24(1). <https://doi.org/10.1186/s12880-024-01217-4>.

[4] Lee SJ, Oh HJ, Son YD, et al. Enhancing deep learning classification performance of tongue lesions in imbalanced data: mosaic-based soft labeling with curriculum learning. *BMC Oral Health* 2024;24(1). <https://doi.org/10.1186/s12903-024-03898-3>.

[5] Kaur BP, H S, Hans R, et al. An augmentation aided concise CNN based architecture for COVID-19 diagnosis in real time. *Sci Rep* 2024;14(1). <https://doi.org/10.1038/s41598-024-51317-y>.

[6] Garcea F, Serra A, Lamberti F, Morra L. Data augmentation for medical imaging: a systematic literature review. *Comput Biol Med* 2023;152:106391. <https://doi.org/10.1016/j.cmpbiomed.2022.106391>.

[7] Rainio O, Klén R. Comparison of simple augmentation transformations for a convolutional neural network classifying medical images. *Signal Image Video Process* 2024. <https://doi.org/10.1007/s11760-024-02998-5>.

[8] Nolden M, Zelzer S, Seitel A, et al. The medical imaging interaction toolkit: challenges and advances. *Int J Comput Assist Radiol Surg* 2013;8(4):607–20. <https://doi.org/10.1007/s11548-013-0840-8>.

[9] Beare R, Lowekamp B, Yaniv Z. Image segmentation, registration and characterization in R with SimpleITK. *J Stat Softw* 2018;86(8). <https://doi.org/10.18637/jss.v086.i08>.

[10] A. Murray, H. Kemenade, wiredfool, et al. "Python-pillow/pillow": 10.3.0. Zenodo. 2024. <https://zenodo.org/records/10903255>.

[11] A. Paszke, S. Gross, F. Massa, et al. "PyTorch: an imperative style, high-performance deep learning library."; 2019. <https://dl.acm.org/doi/10.5555/3454287.3455008>.

[12] Pérez-García F, Sparks R, Ourselin S. TorchIO: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Comput Methods Programs Biomed* 2021;208:106236. <https://doi.org/10.1016/j.cmpb.2021.106236>.

[13] M. Jorge Cardoso, W. Li, R. Brown, et al." MONAI: an open-source framework for deep learning in healthcare." 2022. 10.48550/arxiv.2211.02701.

[14] Fedorov A, Beichel R, Kalpathy-Cramer J, et al. 3D Slicer as an image computing platform for the quantitative imaging network. *Magn Reson Imaging* 2012;30(9): 1323–41. <https://doi.org/10.1016/j.mri.2012.05.001>.

[15] Gocer E. Medical image data augmentation: techniques, comparisons and interpretations. *Artif Intell Rev* 2023. <https://doi.org/10.1007/s10462-023-10453-z>.

[16] Dadras AA, Jaziri A, Frodl E, Vogl TJ, Dietz J, Bucher AM. Lightweight techniques to improve generalization and robustness of U-net based networks for pulmonary lobe segmentation. *Bioengineering* 2023;11(1). <https://doi.org/10.3390/bioengineering11010021>. 21–21.

[17] Azizi S, Culp L, Freyberg J, et al. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nat Biomed Eng* 2023;7(6): 756–79. <https://doi.org/10.1038/s41551-023-01049-7>.

[18] Selvaraj KM, Gnanagurusubbiah S, Roy R, Hephzipah J, Balu S. Enhancing skin lesion classification with advanced deep learning ensemble models: a path towards

- accurate medical diagnostics. *Curr Probl Cancer* 2024. <https://doi.org/10.1016/j.crrproblecancer.2024.101077>. 101077-101077.
- [19] Qu G, Lu B, Shi J, et al. Motion-artifact-augmented pseudo-label network for semi-supervised brain tumor segmentation. *Phys Med Biol* 2024;69(5). <https://doi.org/10.1088/1361-6560/ad2634>. /*Physics in medicine and biology*055023-055023.
- [20] Raudaschl PF, et al. Evaluation of segmentation methods on head and neck CT: auto-segmentation challenge 2015. *Med Phys* 2017;44(5):2020–36. <https://doi.org/10.1002/mp.12197>. Apr.
- [21] Thummerer A, et al. SynthRAD2023 grand challenge dataset: generating synthetic CT for radiotherapy. *Med Phys* 2023;50(7):4664–74. <https://doi.org/10.1002/mp.16529>. June.

A Quality Control Tool to Predict Synthetic Computed Tomography Hounsfield Unit Accuracy

10.1 Introduction

The clinical translation of Deep Learning (DL) models for medical Image-to-Image (I2I) translation requires not only high predictive performance, but also reliable and independent quality control mechanisms. For instance, Synthetic Computed Tomography (sCT) images generated from Magnetic Resonance Imaging (MRI) or Cone Beam Computed Tomography (CBCT) acquisitions have demonstrated the potential to replace conventional CT scans for dose calculation, enabling MR-only and adaptive workflows in radiotherapy [1, 57]. However, the clinical adoption of sCT remains constrained by a fundamental limitation: once deployed, the quality of a synthetic image cannot be directly assessed in the absence of ground truth (GT) CT data [55].

This chapter presents a study aimed at addressing this critical gap by introducing an independent DL-based framework to predict the conversion accuracy of sCT images in terms of slice-wise Mean Absolute Error (MAE), without requiring access to the corresponding GT CT. The proposed approach operates as an external quality control layer, fully decoupled from the image translation algorithm. Rather than estimating model uncertainty within the generative pipeline, the framework predicts the expected slice-wise conversion error itself, thereby providing a complementary and redundant safety mechanism for clinical decision-making [55].

Methodologically, the study introduces a cascading multi-model architecture composed of an initial MAE interval classification stage followed by a class-specific regression step. This design decomposes the complex MAE prediction task into two tractable subproblems,

improving robustness and predictive stability. The workflow was trained and evaluated on two independent cohorts of head and neck cancer patients, including both MR-based and CBCT-based sCT images. Quantitative results demonstrated small absolute deviations between predicted and true MAE values, with no evidence of systematic bias. Importantly, the inference time was compatible with routine clinical workflows, supporting the feasibility of real-time deployment [55].

Although the study was conducted in a centralised setting, it addresses a fundamental prerequisite for the broader deployment of synthetic imaging pipelines, including those developed under decentralised and privacy-preserving paradigms throughout this thesis. Federated Learning (FL) strategies for MR-to-CT and CBCT-to-CT translation, presented in Chapter 4 and Chapter 5), aim to enhance generalisation and data privacy across institutions. Nevertheless, even highly generalisable models require trustworthy, patient-specific quality assessment tools once deployed in real clinical environments. In this sense, the work described here complements decentralised training approaches by focusing on post-training reliability and safety. This contribution represents an essential step toward closing the loop in DL-based I2I conversion for radiotherapy: generation, evaluation, and informed clinical acceptance of synthetic images. The conceptual separation between image synthesis and its quality estimation laid the methodological foundation for subsequent translational developments presented in Chapter 11.

The content of this chapter has been adapted from a manuscript previously published in the peer-reviewed journal “Journal of Imaging”. Portions of the text are reproduced verbatim from the published article [55]. The author of this thesis contributed to the comparative evaluation of multiple classification architectures for the first stage of the proposed pipeline, supporting the selection of the most suitable model for MAE interval prediction. In addition, the author contributed to the revision of the manuscript as a co-author.

10.2 Manuscript

Article

Toward Closing the Loop in Image-to-Image Conversion in Radiotherapy: A Quality Control Tool to Predict Synthetic Computed Tomography Hounsfield Unit Accuracy

Paolo Zaffino ^{1,2,*} , ¹ , ¹ , ¹ , ¹ , ¹ , ¹

convenient to acquire a real CT. In particular, radiotherapy is one of the main fields where sCT can play a fundamental role in improving treatment planning, allowing MR and CBCT to be integrated into the imaging pipeline without the need for additional CT scans [5–11]. Indeed, MR offers better soft tissue contrast to contour the target tissue and the organs at risk during treatment simulation. However, MR intensities do not reveal electron density properties, and additional CT acquisitions are necessary for dose calculation purposes. CBCT, on the other hand, is widely used for assessing inter-fraction geometric deviations in adaptive radiotherapy. But, again, CBCT cannot be used to calculate the dose due to the poor image quality; this forces the clinician to re-plan the treatment on new CT scans when it is necessary. The conversion of MR and CBCT in sCT images can reduce radiation toxicity among patients and decrease the time required for the treatment [2,12]. To date, the best-performing strategies to remap image intensities rely on supervised artificial intelligence algorithms [3,4], in particular those based on deep learning (DL). However, the accuracy of this kind of algorithm strongly depends on the dataset used for the training, and they can underperform in case of strong image artifacts and atypical anatomy. Moreover, the translation quality is assessed by comparing the sCT with the real CT, which represents the ground truth (GT). One of the most used metrics to assess conversion accuracy is the pixel-to-pixel Mean Absolute Error (MAE) [3], defined as Equation (1).

$$MAE = \frac{\sum_{i=1}^n |GT_i - sCT_i|}{n} \quad (1)$$

Mean Absolute Error formula.

i ranges from 1 to n, number of pixels/voxels both for GT and sCT.

In the case of clinical use, the absence of GT makes it impossible to evaluate the conversion accuracy, which has a direct impact on dose calculation. As a consequence, clinicians have to trust the pre-trained DL algorithms, being unable, except for eye-catching macroscopic mistakes, to detect conversion errors that could dramatically affect the treatment. For the clinical use of sCT for treatment adaptation [13], the implementation of online quality control procedures is mandatory. Some recent works [14–16] propose a DL model for synthetic CT generation with uncertainty predictions. The conversion pipeline provides an sCT alongside a predicted uncertainty map. A shortcoming of this approach is the information generation within one algorithm. In case of conversion failure, the uncertainty estimation could also be unreliable. Moreover, it is worth underlining the fundamental difference between “uncertainty” and “error”: a system (as well as a human) can make totally wrong predictions while being, at the same time, absolutely certain about them. For this reason, it is important to predict, preferably by using an independent tool, the conversion error of a process rather than its confidence. In this paper, we implemented a strategy to estimate the quality of sCT in terms of MAE without the need for the corresponding GT and independent of the translation model. The method is based on deep learning algorithms that were trained and tested on two patient cohorts including MR- and CBCT-based sCT images.

2. Materials and Methods

2.1. Dataset

Two datasets of head and neck cancer patients, previously described in Thummerer et al. [17,18], were used in this study. All patients were treated with intensity-modulated proton therapy at the University Medical Center Groningen, The Netherlands. The first set included 33 patients scanned both with CBCT and CT. The second set included 27 patients with MR and corresponding CT images. All CBCT/CT and MR/CT pairs were carefully aligned using deformable image registration to match the anatomy. For each patient of both datasets, an sCT was generated by using a basic version of the algorithm described in Spadea et al. [19]. The original Spadea et al. method, relying on an encoder/decoder architecture to execute 2D-based image translation, takes advantage of a multi-plane

approach, in which volumes obtained along axial, sagittal, and coronal directions are voted to generate the final image. In this work, only the synthetic CTs reconstructed by stacking the translated 2D images along the craniocaudal axis were used. From now on, the sCTs derived from CBCT will be referred to as sCT_{CBCT} and the ones obtained from MR as sCT_{MR} .

2.2. General Pipeline

The proposed quality control approach is based on a 2D DL model architecture pipeline. The main idea is to provide an axial sCT slice as the input to the DL scheme and to obtain the predicted MAE for that image as the output (see Figure 1). To achieve this, the DL model is trained on retrospective data where ground truth CT and the corresponding sCT are available. Having GT images, the actual MAE can be computed slice-by-slice according to equation 1. During the training, the pipeline receives the sCT slice (input) and the actual MAE (output to predict). Afterwards, during deployment, the trained models only receive the sCT slices and predict an MAE scalar for the synthetic 2D image.



Figure 1. Representation of the general MAE prediction pipeline. An axial sCT slice is given as input, and the associated MAE scalar for the image slice is predicted by using a DL pipeline.

More specifically, as depicted in Figure 2, the workflow consists of a cascade of 2 series of DL models, used in a sequential way to handle the complexity of the MAE prediction task. The rationale is to first execute a raw prediction of the MAE interval (classification), followed by a more precise estimation of the MAE value (regression). This approach was inspired by a similar and successful strategy to predict segmentation accuracy in multi-atlas-based segmentation [20]. This strategy allows the initial problem to be split into two subproblems, both of which are easier to solve than the first one.

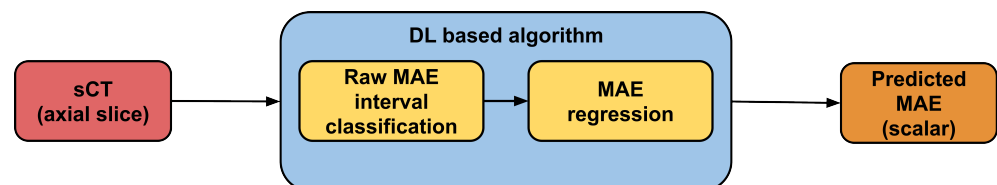


Figure 2. A more detailed graphical representation of the MAE prediction pipeline. The final MAE prediction is obtained as a result of two DL steps: First a raw MAE interval classification is performed, followed by a more precise MAE estimation based on a regression algorithm.

2.3. Output Visualization

The output of a single prediction is a scalar value representing the estimated MAE for the entire given axial slice. In envisioning a clinical use, it is fundamental to visualize the predicted conversion error slice-by-slice and overlaid on the sCT. To achieve this, a volume (named “Predicted MAE volume”, $pMAE_{volume}$), having the same shape as the corresponding sCT, is generated. During the sCT evaluation step, the predicted MAE of the i -th axial slice is assigned to the i -th axial slice of $pMAE_{volume}$, which will be visualized by taking advantage of a dedicated color map. Figure 3 shows an example of an sCT overlaid to its $pMAE_{volume}$.

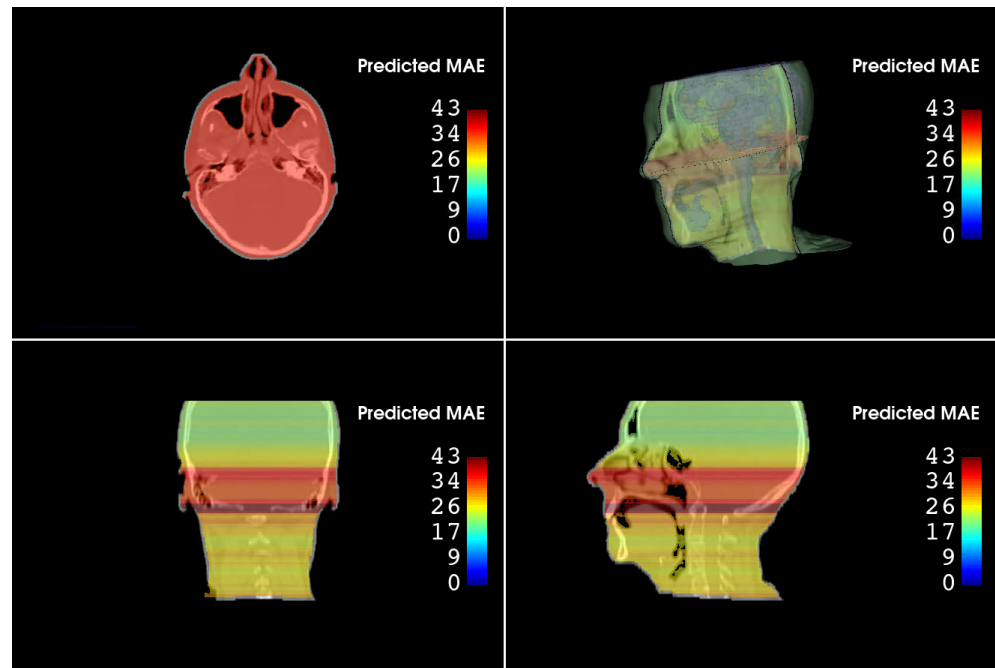


Figure 3. Exemplary sCT_{CBCT} overlaid with its $pMAE_{volume}$. In addition to the 2D views (axial, sagittal, and coronal planes), the 3D representation is also shown.

2.4. Pipeline Details

As already reported, the first step for MAE estimation is a raw classification of the conversion error. More specifically, 4 classes are used in the proposed pipeline (low, low-medium, medium-high, and high MAE). In order to define the MAE range for each interval, the actual MAE distribution of the entire dataset is computed. Then, the first class of the prediction is associated with an MAE ranging from 0 to the 1st quartile of the total distribution, the second class from the 1st to the 2nd quartile, the third class from the 2nd to the 3rd quartile, and finally the fourth class from the 3rd percentile to the highest MAE values (in order to deal with an open interval, this last bin was saturated to an upper bound HU value). This allows for the balancing of the number of examples per class. The same values are also used as an initialization to define the MAE boundaries for the regression step (the second one of the pipeline), where a separate model is trained for each of the four MAE classes. More specifically, each of the four regression models is trained by showing only the slices having an MAE value included in the same range of the connected classification class. However, in order to take into account possible mistakes in the classification step, the boundaries of the regression models are relaxed by also including cases with GT MAE slightly lower/higher than the quartiles (± 5 HU). Table 1 reports the computed bin ranges, both for the classification and regression steps. In total, 5 models have to be trained (1 for classification and 4 for regression), but, in the inference mode, each sCT slice will be processed just by 2 of them (classification followed by the class-specific regression). For the sake of clarity, the entire prediction pipeline is shown in Figure 4.

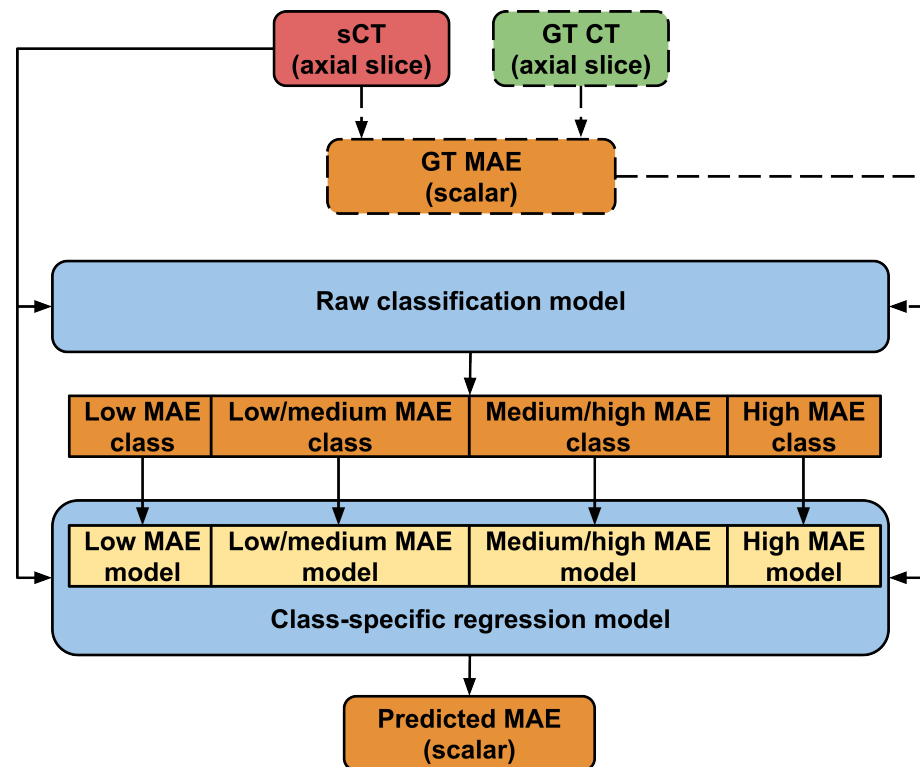


Figure 4. Detailed workflow of MAE prediction. A single sCT axial slice is fed firstly into a DL model that classifies it as belonging to a specific MAE class. According to this prediction, the 2D image is then provided as input to a connected DL regression model, specifically trained to operate on a restricted range of MAE values. As a result, the MAE of a single sCT slice can be forecasted. In order to train the different models with a GT MAE, the ground truth CT is needed (dashed lines are needed only to train the models).

2.5. DL Architecture and Training

The DL architecture to be used in this work was chosen after testing several models, both from VGG [21] and ResNet [22] families. Our tests revealed that VGG-16 provides the best prediction accuracies, so it was used as the baseline architecture for all the models. VGG (Visual Geometry Group) is a family of high-performance convolutional neural networks developed for classification and regression tasks. Even though it was created for general-purpose computer vision applications, it has also been widely used in the biomedical environment. The backbone of VGG-16 includes 13 convolutional layers (that serve as feature extraction engines), followed by 3 fully connected layers (that are used to obtain the output label starting from the computed feature map). The loss functions to optimize for classification and regression are, respectively, cross entropy and mean squared error. The best epoch for the classification task is not chosen on the basis of the cross entropy value but by considering the accuracy weighted according to the misclassification magnitude. In practical terms, this means that a classification error of 1 class is preferred over 2 and 3 classes of misclassification. This choice is made to let the regression step, trained on slightly larger MAE boundaries than the connected classification bin, recover the potential classification error. The entire workflow is written in Python by taking advantage of PyTorch library [23]. The learning rate is set to 5×10^{-5} , and $L1$ and $L2$ regularizations are included (both weights equal to 4×10^{-4}). The image augmentation scheme includes mirroring and translations, as defined by Spadea et al [19]. The dataset is split into 75% training, 10% validation, and 15% testing. K-fold cross-validation is used to evaluate the model generalization performance on the entire dataset.

2.6. Experiments

To assess the learning generalization capability of the approach toward different image modalities, 3 pipelines are implemented:

1. Models trained only with sCT_{CBCT} as input (named $Pipeline_{CBCT}$),
2. Models trained only with sCT_{MR} as input ($Pipeline_{MR}$),
3. Models trained with both sCT_{CBCT} and sCT_{MR} as input ($Pipeline_{MIXED}$).

Once all the models of each pipeline are trained, multiple testing schemes are run to evaluate the prediction performance. In particular, the following experiments are executed:

- $Pipeline_{CBCT}$ is used to predict the MAE only for sCT_{CBCT} data
- $Pipeline_{MR}$ is used to predict the MAE only for sCT_{MR} data
- $Pipeline_{MIXED}$ is used to predict the MAE only for sCT_{CBCT} data
- $Pipeline_{MIXED}$ is used to predict the MAE only for sCT_{MR} data

Table 1 shows the MAE binning used to train both classification and regression models in each pipeline.

Table 1. MAE ranges for classification and regression, both for MR, CBCT, and MIXED pipelines.

Pipeline	Low MAE	Medium-Low MAE	Medium-High MAE	High MAE
MR	0–47 (classification)	47–54 (classification)	54–68 (classification)	68–100 (classification)
	0–52 (regression)	42–59 (regression)	49–73 (regression)	63–100 (regression)
CBCT	0–27 (classification)	27–32 (classification)	32–42 (classification)	42–70 (classification)
	0–32 (regression)	22–37 (regression)	27–47 (regression)	37–70 (regression)
MIXED	0–32 (classification)	32–44 (classification)	44–56 (classification)	56–90 (classification)
	0–37 (regression)	27–49 (regression)	39–61 (regression)	51–90 (regression)

2.7. Prediction Pipeline Evaluation

The accuracy of the prediction workflow is quantified in terms of signed and unsigned deviation from the GT MAE. More specifically, both the Prediction Deviation (PD) and the Absolute Prediction Deviation (APD) are computed for each slice as defined in Equations (2) and (3):

$$PD = MAE_{GT} - MAE_{predicted} \quad (2)$$

Signed predicted deviation between GT and inferred MAE.

$$APD = |MAE_{GT} - MAE_{predicted}| \quad (3)$$

Absolute predicted deviation between GT and inferred MAE.

Both MAE_{GT} and $MAE_{predicted}$ are saturated to the upper bound values reported in Table 1.

3. Results

Figures 5 and 6 show, for each DL pipeline, the PD and APD distributions computed over all the 2D slices of all the patients.

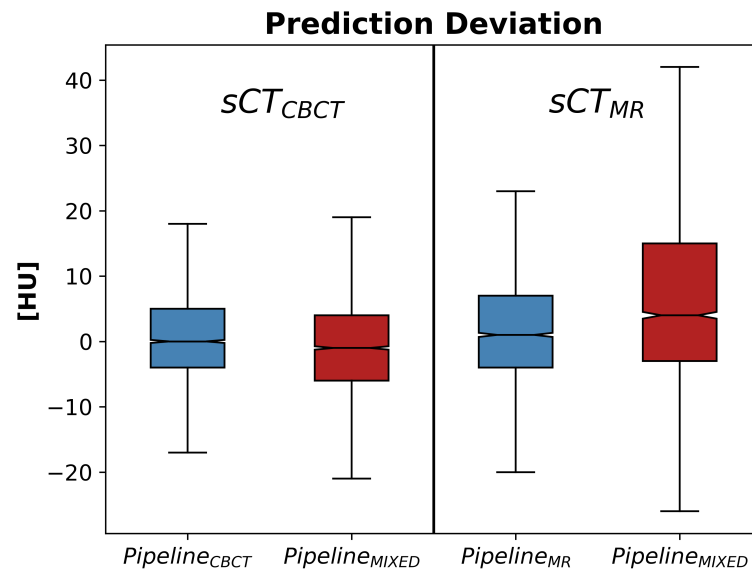


Figure 5. PD distributions for modality-specific and mixed pipelines. Results for sCT_{CBCT} and sCT_{MR} are reported, respectively, in the left and in the right panel.

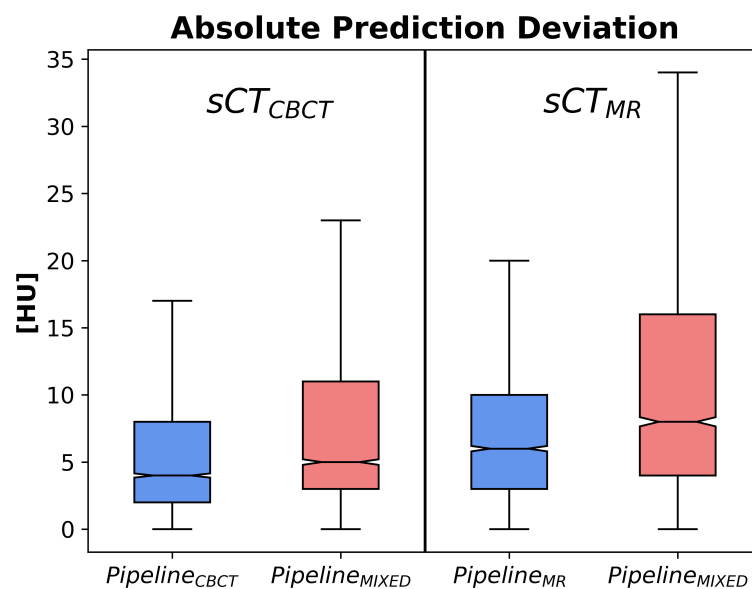


Figure 6. APD distributions for modality-specific and mixed pipelines. Results for sCT_{CBCT} and sCT_{MR} are reported, respectively, in the left and in the right panel.

Tables containing the median, the 5th, the 25th, the 75th, and the 95th percentiles of both PD and APD are reported in Appendix A. To assess whether a statistical difference exists between the predictions obtained by using the mixed-trained workflow ($Pipeline_{MIXED}$) versus the specific modality pipeline ($Pipeline_{CBCT}$ and $Pipeline_{MR}$), the Wilcoxon rank-sum test is run over all the pairs of MAE predictions obtained for all the patients fed into the different pipelines. These tests reveal that both single-modality options performed significantly better than the mixed solution (p -values < 0.01). Figure 3 shows an example of the overlay of the sCT and its $pMAE_{volume}$. As can be seen, for slices depicting small and movable anatomical structures (e.g., nasal cavities), higher MAE values are predicted. This behavior was expected and desirable, confirming the effectiveness of the proposed strategy.

Despite the necessity for a more extended training phase to enable the training of 5 different models, the inference step is completed in a few seconds. As a result, the required time to evaluate an sCT is totally compatible with a clinical scenario and will not slow down the radiotherapy planning workflow.

4. Discussion

In this work, we introduce a new method to predict the accuracy of sCT generation when the GT is not available (i.e. use of sCT in a clinical setting). The technique is based on using a DL model that is independent from any other model used to generate the sCT.

Being the first implemented workflow for the autonomous MAE prediction of DL-based sCT, there are no closely related studies to compare it with. The most similar articles in the literature are those that generate uncertainty maps alongside sCT [14–16]. The idea behind such works is to have a distribution of values for each voxel rather than a single intensity (by using Bayesian neural networks or by executing multiple inferences with active dropout layers). The averages of the obtained distributions represent the final sCT, while the variabilities are informative about the uncertainty of the prediction (with the uncertainty map correlating with the GT intensity and dose errors). However, in this regard, it is fundamental to underline the conceptual difference existing between our proposed workflow (conversion error estimation) and the uncertainty prediction proposed by others, since being confident about something does not necessarily mean being accurate. The main concern about these strategies is the possibility that the conversion model provides, with high assurance, a synthetic image that is actually wrong. Because of this, the main pillars of the proposed workflow are the separation between the image translation step and its quality control, and the prediction of the actual error rather than the confidence. Predicting the conversion error of sCT, in fact, is of paramount importance in contexts such as radiotherapy, where wrong HU assignments lead to wrong dose estimation to the target and organs at risk. Specifically, for the clinical introduction of MR-guided radio and proton therapy [24], an autonomous and independent quality control tool (an information with an additional level of redundancy and with a rationale totally different from the uncertainty assessment produced by the conversion algorithm itself) would be required to ensure reliable clinical decision-making based on deep-learning generated medical image data. We would like to highlight that the aim of the proposed tool is to intercept conversion errors that are not expected and not detectable by the human operator. Clinicians, in fact, well know the anatomical regions that can be more likely affected by translation error and are able to identify evident anomalous conversions, so our intention is to catch unforeseeable failures. The findings of our experiments show that it is possible to predict the accuracy of sCT in terms of MAE, with small deviations between inferred and GT MAE values. In fact, considering the HU range, an absolute prediction deviation of $4 \pm 3HU$ (for CBCT-derived sCT) and $6 \pm 3.5HU$ (for MR-derived sCT) has no impact on the dose estimation, even in the case of proton therapy, known to be extremely sensitive to HU variations. Signed prediction deviations demonstrate the absence of systematic errors. Results are also consistent with real-case expectations. Referring to Figure 3, worse conversion error is predicted for axial slices included between the nasal cavities and the mandibular bone. This result is very common in sCT generation, since the image quality in this anatomical area is deteriorated by the presence of dental fillings and motion artifacts. In a typical clinical setting, the user would receive a warning on which part of the sCT is reliable and which is not. In light of the magnitude of the predicted error and its spatial localization with respect to the tumor and the tissues to be spared, clinicians may elect to exclude the use of the sCT. The comparison between models trained by using single-modality versus mixed-modality dataset reveals that better results are obtained with single-modality training. The mixed setup was implemented to test the capability of the workflow to be agnostic to the initial image modality. The entire workflow is based on the VGG-16 model, which has been demonstrated to be the most effective architecture for predicting the MAE in our tests. In addition to its high predictive accuracy, this model requires fewer computational resources than deeper networks, allowing for fast sCT evaluation even in the absence of dedicated high-end GPUs, which is a crucial advantage in a real clinical setting.

Regarding the potential error propagation between the classification and regression steps, the proposed workflow is also robust to this possibility. In fact, as can be deduced from the results (Figures 5 and 6), the prediction deviations are much smaller than the

differences that would be generated if the raw MAE bin was misclassified and if this error was not recovered in the regression phase.

The proposed approach could be further improved by replacing the global axial MAE scalars currently predicted as quality control metrics with a more informative index from a clinical point of view. Such an index could, for example, include dosimetric and/or voxel-level prediction. In the future, it would be interesting to evaluate the proposed workflow on images collected in different institutions and converted by using different translation strategies (including cycle-GAN architectures, to get rid of the error introduced by the image registration step). We envision pursuing work in that direction, executing more complex and more computationally demanding experiments. The proposed idea, in addition to enabling a real-time synthetic image evaluation in the clinic, can be also used to implement synthetic image generation algorithms, when CT-MRI or CT-CBCT paired data are not available. The pipeline described here, in fact, can directly replace the computation of the MAE-based loss when the ground truth CT is not available.

5. Conclusions

In conclusion, we demonstrated that an independent prediction of the performance of an algorithm for sCT generation is possible and, most importantly, we hope to start a debate about usable strategies in real clinical environments for assessing the quality of synthetic medical images.

Author Contributions: P.Z.: Conceptualization, Investigation, Methodology, Software, Visualization, Formal analysis, and Writing—original draft; C.B.R.: Software, Writing—review and editing; A.T.: Data curation, Resources, and Writing—review and editing; G.G.M.: Data curation, Resources, and Writing—review and editing; J.A.L.: Supervision, Writing—review and editing; A.P.: Visualization, Writing—review and editing; C.C.: Supervision, Visualization, and Writing—review and editing; J.S.: Supervision, Writing—review and editing; A.C.K.: Supervision, Visualization, and Writing—review and editing; S.B.: Supervision, Resources, and Writing—review and editing; M.F.S.: Supervision, Methodology, Formal analysis, and Writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: All the patients were enrolled in the UMCG Radiation Oncology standardized follow-up program, approved by the medical ethics committee. Being a retrospective study, no additional images were acquired.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The datasets used in this article are not readily available because they contain sensitive information. Requests to access the datasets should be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Tables A1 and A2 show, for each DL pipeline, the median and the percentiles of PD and APD distributions computed over all the 2D slices of all the patients.

Table A1. Statistics about the PD for each conducted experiment.

Experiment	Prediction Deviation (PD)				
	5th Percentile [HU]	25th Percentile [HU]	Median [HU]	75th Percentile [HU]	95th Percentile [HU]
<i>Pipeline_{CBCT}</i> predicting on <i>sCT_{CBCT}</i>	−10	−4	0	5	20
<i>Pipeline_{MR}</i> predicting on <i>sCT_{MR}</i>	−12	−4	1	7	16
<i>Pipeline_{MIXED}</i> predicting on <i>sCT_{CBCT}</i>	−23	−6	−1	4	21
<i>Pipeline_{MIXED}</i> predicting on <i>sCT_{MR}</i>	−11	−3	4	15	32

Table A2. Statistics about the APD for each conducted experiment.

Experiment	Absolute Prediction Deviation (APD)				
	5th Percentile [HU]	25th Percentile [HU]	Median [HU]	75th Percentile [HU]	95th Percentile [HU]
<i>Pipeline_{CBCT}</i> predicting on <i>sCT_{CBCT}</i>	0	2	4	8	21
<i>Pipeline_{MR}</i> predicting on <i>sCT_{MR}</i>	1	3	6	10	17
<i>Pipeline_{MIXED}</i> predicting on <i>sCT_{CBCT}</i>	0	3	5	11	29
<i>Pipeline_{MIXED}</i> predicting on <i>sCT_{MR}</i>	1	4	8	16	32

References

1. Arabi, H.; Dowling, J.A.; Burgos, N.; Han, X.; Greer, P.B.; Koutsouvelis, N.; Zaidi, H. Comparative study of algorithms for synthetic CT generation from MRI: Consequences for MRI-guided radiation planning in the pelvic region. *Med. Phys.* **2018**, *45*, 5218–5233. [[CrossRef](#)] [[PubMed](#)]
2. Johnstone, E.; Wyatt, J.J.; Henry, A.M.; Short, S.C.; Sebag-Montefiore, D.; Murray, L.; Kelly, C.G.; McCallum, H.M.; Speight, R. Systematic review of synthetic computed tomography generation methodologies for use in magnetic resonance imaging-only radiation therapy. *Int. J. Radiat. Oncol. Biol. Phys.* **2018**, *100*, 199–217. [[CrossRef](#)] [[PubMed](#)]
3. Spadea, M.F.; Maspero, M.; Zaffino, P.; Seco, J. Deep learning based synthetic-CT generation in radiotherapy and PET: A review. *Med. Phys.* **2021**, *48*, 6537–6566. [[CrossRef](#)] [[PubMed](#)]
4. Boulanger, M.; Nunes, J.C.; Chourak, H.; Largent, A.; Tahri, S.; Acosta, O.; De Crevoisier, R.; Lafond, C.; Barateau, A. Deep learning methods to generate synthetic CT from MRI in radiotherapy: A literature review. *Phys. Medica* **2021**, *89*, 265–281. [[CrossRef](#)] [[PubMed](#)]
5. Dinkla, A.M.; Florkow, M.C.; Maspero, M.; Savenije, M.H.; Zijlstra, F.; Doornaert, P.A.; van Stralen, M.; Philippens, M.E.; van den Berg, C.A.; Seevinck, P.R. Dosimetric evaluation of synthetic CT for head and neck radiotherapy generated by a patch-based three-dimensional convolutional neural network. *Med. Phys.* **2019**, *46*, 4095–4104. [[CrossRef](#)]
6. Liu, Y.; Lei, Y.; Wang, Y.; Wang, T.; Ren, L.; Lin, L.; McDonald, M.; Curran, W.J.; Liu, T.; Zhou, J.; et al. MRI-based treatment planning for proton radiotherapy: Dosimetric validation of a deep learning-based liver synthetic CT generation method. *Phys. Med. Biol.* **2019**, *64*, 145015. [[CrossRef](#)]
7. Liu, Y.; Lei, Y.; Wang, T.; Fu, Y.; Tang, X.; Curran, W.J.; Liu, T.; Patel, P.; Yang, X. CBCT-based synthetic CT generation using deep-attention cycleGAN for pancreatic adaptive radiotherapy. *Med. Phys.* **2020**, *47*, 2472–2483. [[CrossRef](#)]
8. Maspero, M.; Savenije, M.H.; Dinkla, A.M.; Seevinck, P.R.; Intven, M.P.; Jurgensliemk-Schulz, I.M.; Kerkmeijer, L.G.; Van Den Berg, C.A. Dose evaluation of fast synthetic-CT generation using a generative adversarial network for general pelvis MR-only radiotherapy. *Phys. Med. Biol.* **2018**, *63*, 185001. [[CrossRef](#)]
9. Maspero, M.; Bentvelzen, L.G.; Savenije, M.H.; Guerreiro, F.; Seravalli, E.; Janssens, G.O.; van den Berg, C.A.; Philippens, M.E. Deep learning-based synthetic CT generation for paediatric brain MR-only photon and proton radiotherapy. *Radiother. Oncol.* **2020**, *153*, 197–204. [[CrossRef](#)]
10. Dai, X.; Lei, Y.; Wynne, J.; Janopaul-Naylor, J.; Wang, T.; Roper, J.; Curran, W.J.; Liu, T.; Patel, P.; Yang, X. Synthetic CT-aided multiorgan segmentation for CBCT-guided adaptive pancreatic radiotherapy. *Med. Phys.* **2021**, *48*, 7063–7073. [[CrossRef](#)]
11. Gao, L.; Xie, K.; Wu, X.; Lu, Z.; Li, C.; Sun, J.; Lin, T.; Sui, J.; Ni, X. Generating synthetic CT from low-dose cone-beam CT by using generative adversarial networks for adaptive radiotherapy. *Radiat. Oncol.* **2021**, *16*, 202. [[CrossRef](#)] [[PubMed](#)]
12. Kazemifar, S.; McGuire, S.; Timmerman, R.; Wardak, Z.; Nguyen, D.; Park, Y.; Jiang, S.; Owringi, A. MRI-only brain radiotherapy: Assessing the dosimetric accuracy of synthetic CT images generated using a deep learning approach. *Radiother. Oncol.* **2019**, *136*, 56–63. [[CrossRef](#)] [[PubMed](#)]
13. van Elmpt, W.; Taasti, V.T.; Redalen, K.R. Current and future developments of synthetic computed tomography generation for radiotherapy. *Phys. Imaging Radiat. Oncol.* **2023**, *28*, 100521. [[CrossRef](#)]
14. Hemsley, M.; Chugh, B.; Ruschin, M.; Lee, Y.; Tseng, C.L.; Stanisiz, G.; Lau, A. Deep generative model for synthetic-CT generation with uncertainty predictions. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, 4–8 October 2020; Proceedings, Part I 23; Springer: Cham, Switzerland, 2020; pp. 834–844.
15. Li, X.; Bellotti, R.; Meier, G.; Bachtiary, B.; Weber, D.; Lomax, A.; Buhmann, J.; Zhang, Y. Uncertainty-aware MR-based CT synthesis for robust proton therapy planning of brain tumour. *Radiother. Oncol.* **2024**, *191*, 110056. [[CrossRef](#)] [[PubMed](#)]
16. Galapon, A.V., Jr.; Thummerer, A.; Langendijk, J.A.; Wagenaar, D.; Both, S. Feasibility of Monte Carlo dropout-based uncertainty maps to evaluate deep learning-based synthetic CTs for adaptive proton therapy. *Med. Phys.* **2024**, *51*, 2499–2509. [[CrossRef](#)]
17. Thummerer, A.; Zaffino, P.; Meijers, A.; Marmitt, G.G.; Seco, J.; Steenbakkers, R.J.; Langendijk, J.A.; Both, S.; Spadea, M.F.; Knopf, A.C. Comparison of CBCT based synthetic CT methods suitable for proton dose calculations in adaptive proton therapy. *Phys. Med. Biol.* **2020**, *65*, 095002. [[CrossRef](#)]
18. Thummerer, A.; De Jong, B.A.; Zaffino, P.; Meijers, A.; Marmitt, G.G.; Seco, J.; Steenbakkers, R.J.; Langendijk, J.A.; Both, S.; Spadea, M.F.; et al. Comparison of the suitability of CBCT-and MR-based synthetic CTs for daily adaptive proton therapy in head and neck patients. *Phys. Med. Biol.* **2020**, *65*, 235036. [[CrossRef](#)]

19. Spadea, M.F.; Pileggi, G.; Zaffino, P.; Salome, P.; Catana, C.; Izquierdo-Garcia, D.; Amato, F.; Seco, J. Deep convolution neural network (DCNN) multiplane approach to synthetic CT generation from MR images—Application in brain proton therapy. *Int. J. Radiat. Oncol. Biol. Phys.* **2019**, *105*, 495–503. [[CrossRef](#)]
20. Zaffino, P.; Ciardo, D.; Raudaschl, P.; Fritscher, K.; Ricotti, R.; Alterio, D.; Marvaso, G.; Fodor, C.; Baroni, G.; Amato, F.; et al. Multi atlas based segmentation: Should we prefer the best atlas group over the group of best atlases? *Phys. Med. Biol.* **2018**, *63*, 12NT01. [[CrossRef](#)]
21. Simonyan, K. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
22. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2016; pp. 770–778.
23. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, Canada, 8–14 December 2019; pp. 8026–8037.
24. Hoffmann, A.; Oborn, B.; Moteabbed, M.; Yan, S.; Bortfeld, T.; Knopf, A.; Fuchs, H.; Georg, D.; Seco, J.; Spadea, M.F.; et al. MR-guided proton therapy: A review and a preview. *Radiat. Oncol.* **2020**, *15*, 129. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

ModalityConverter: Medical Image-to-Image Translation in 3D Slicer

11.1 Motivation and significance

This chapter presents *ModalityConverter*, an open-source 3D Slicer [107] extension that, for the first time within the 3D Slicer ecosystem, provided free and open access to pre-trained Artificial Intelligence (AI) models for medical Image-to-Image (I2I) translation and synthetic images assessment. The extension constitutes a direct translational outcome of the research activities developed throughout this doctoral thesis.

Indeed, within this thesis, two Federated Learning (FL) frameworks have been developed to enable privacy-preserving and decentralised training of Synthetic Computed Tomography (sCT) generation models, addressing both brain T1w MRI-to-CT and head and neck CBCT-to-CT translation tasks. These contributions, described in Chapter 4 and Chapter 5, demonstrated how FL enables collaborative multi-institutional model training while preserving data privacy, regulatory compliance, and robust model generalisation.

However, beyond algorithmic innovation and methodological validation, a fundamental challenge in research remains the effective translation of scientific advances into tools that are accessible, reproducible, and usable by the broader research and clinical communities. Therefore, by embedding the pre-trained FedSynthCT models within 3D Slicer, the research contributions of this thesis are transformed into practical and reusable tools that can be accessed by researchers without requiring specialised machine learning infrastructure or advanced technical expertise.

In addition to the FedSynthCT models, the extension also integrates an I2I translation model for synthetic PET (sPET) generation from lung CT, developed by Salehjahreni et al. [77] and lately integrated in collaboration with the authors of the original study. Although this model was not trained using the FL paradigm, its inclusion is motivated by a coherent

scientific vision: the creation of a unified, open, and accessible ecosystem for medical I2I within 3D Slicer.

In its current version (v. 1.4.6), the ModalityConverter extension comprises two complementary modules: (i) the Modality Converter module for image-to-image translation, and (ii) the Synthetic Image Quality Assurance module for the evaluation of synthetic image reliability. This dual-module structure extends the original translational objective of the thesis beyond model deployment, enabling integrated generation and quality assessment of synthetic medical images within a unified research environment.

Overall, this chapter positions the ModalityConverter extension not as a standalone software artifact, but as the translational interface of the thesis: a concrete and operational bridge between centralised and decentralised models, deployment, quality assessment of synthetic outputs and real-world research usability, supporting the full lifecycle of synthetic medical image applications.

11.2 Software architecture and module ecosystem

11.2.1 ModalityConverter module

The ModalityConverter module integrates multiple Deep Learning (DL) models within a unified and user-friendly interface in the 3D Slicer environment. It provides GPU-accelerated inference to enable efficient processing of high-dimensional medical imaging data and is explicitly designed to support the integration of additional custom models, ensuring extensibility and long-term sustainability. The extension follows the standard “*Scripted Loadable Module*” architectural pattern adopted by 3D Slicer. Its structure is based on a clear separation of responsibilities between graphical interaction, computational logic, and model management.

In detail, the key architectural components of the ModalityConverter module are:

- **ModalityConverterWidget**

This class manages the graphical user interface and user interactions. It provides input and output selectors, model selection menus, parameter configuration panels, and progress and status indicators, fully integrated within the 3D Slicer environment.

- **ModalityConverterLogic**

This class encapsulates the computational workflow of the extension. It is responsible for parsing the selected model identifier, loading the corresponding pre-trained model parameters, executing the inference pipeline on the selected input image, and writing the generated output to the target volume, which is subsequently visualised in the 3D Slicer multi-panel view.

- **ModalityConverterLib**

This library includes a collection of Python modules that implement the core functional infrastructure of the extension. In particular, an abstract `BaseModel` class and a `@register_model` Python decorator are defined to formalise and abstract the internal model logic. All specific model implementations inherit from this base class and are required to implement the methods `loadModelFromPath()` and `runInference()`, thereby enforcing a consistent interface across heterogeneous model architectures.

Moreover, a JavaScript Object Notation (JSON) metadata file is used to define the available models, including unique identifiers, download URLs, display names, and descriptive information. This modular design ensures that the integration of new models is achieved through the creation of a dedicated subclass of `BaseModel` and the corresponding metadata entry, following a documented and reproducible convention.

Figure 11.1 provides a visual representation of the graphical user interface, highlighting the main functional areas for input selection, model configuration, and inference execution within the `ModalityConverter` module. The interface guides the user through all essential steps of the translation pipeline, from data selection to model execution.

The typical usage workflow of the module is structured as follows:

1. **Input selection:** the user selects an existing volume in the 3D Slicer scene as the input image. A button has been provided to give the user quick and direct access to the 3D Slicer “Download Sample Data” module, which allows users to download sample medical images such as MRI, CT, and CBCT scans of different anatomical areas. For specific models, an optional binary mask can be provided in order to restrict the translation process to a defined region of interest.
2. **Model selection:** a drop-down menu lists all available pre-trained models. Selecting a model displays detailed descriptive information, including modality mapping, anatomical scope, and bibliographic references, as defined in the associated metadata.

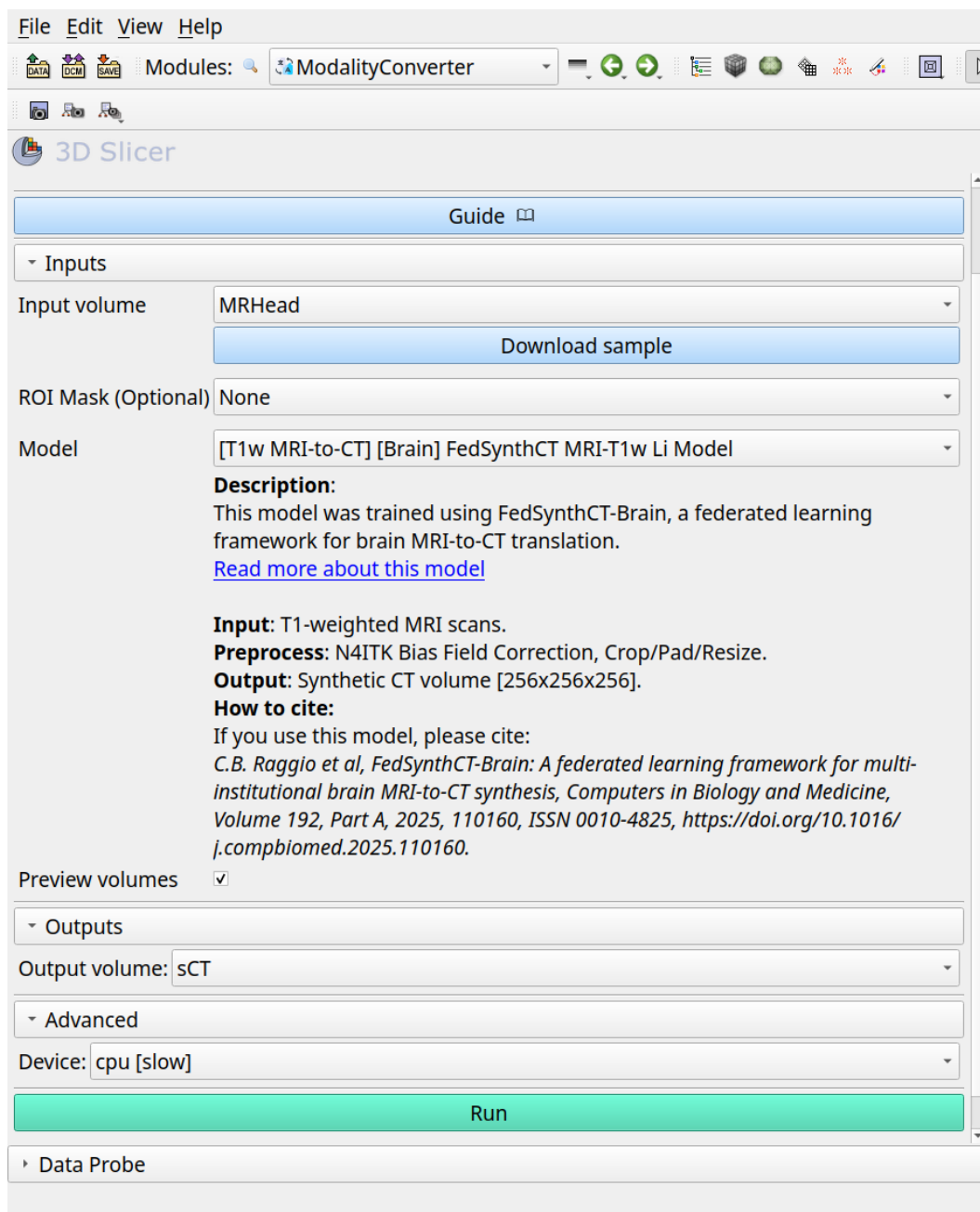


Figure 11.1: Graphical user interface of the ModalityConverter module.

3. **Preview volumes:** the “Preview volumes” option enables real-time visualisation of intermediate volumes generated throughout the inference pipeline, including preprocessed inputs and automatically generated masks, within the 3D Slicer scene.
4. **Output and device configuration:** the user specifies the output volume in which the synthetic image will be stored. Within the advanced configuration options, the

computational device can be selected, allowing execution either on the CPU or on a CUDA-compatible GPU.

5. **Inference execution:** the inference process can be started by the “Run” button. Informative status messages and progress indicators provide feedback during execution. Upon completion, the generated synthetic image is automatically visualised in the 3D Slicer scene.

By adhering to established 3D Slicer development conventions, leveraging object-oriented programming paradigm, and standard Python-based DL frameworks, the module ensures reproducible and well-documented behaviour. Model-specific parameters, inference pipelines, and transformations remain consistent with their original training configurations, while detailed execution feedback supports debugging, validation, and auditing in research workflows.

11.2.2 Synthetic Image Quality Assurance module

Beyond synthetic image generation, reliable deployment of I2I translation models in research and clinical environments requires more objective and reproducible quality assessment mechanisms [55]. While traditional validation is typically performed offline using paired reference data and global quantitative metrics, real-world deployment scenarios lack access to ground-truth images. Consequently, integrated quality estimation becomes a critical component for responsible use of synthetic medical images.

To address this need, the ModalityConverter extension includes a dedicated module, named Synthetic Image Quality Assurance (SIQA). This module maintains the same architectural principles adopted in the ModalityConverter module, presented in Section 11.2.1.

The same modular structure (Widget, Logic, and Library components) is employed, ensuring consistency, maintainability, and extensibility. Thus, the primary distinction lies not in the infrastructure, but in the nature of the integrated models. While the ModalityConverter module makes available pre-trained I2I translation DL models that generate synthetic volumes, the SIQA module integrates DL models designed to evaluate the quality and reliability of synthetic outputs. Instead of producing a new modality, these models generate quantitative indicators or spatially resolved error estimations that support interpretation and validation of synthetic images.

As the first implementation within the SIQA module, the author of the original study [55] integrated a DL model for slice-wise Mean Absolute Error (MAE) prediction of the sCT volumes, to which the author of this thesis also contributed. The methodological details and a detailed description of the output for this model were described in Chapter 10.

Integrating this quality assessment model within the same deployment framework as the generative models enables a closed-loop workflow. For instance, sCT generation through FedSynthCT models can be immediately followed by local quality estimation within the same environment.

11.2.3 Custom model integration

To demonstrate the practical usability and extensibility of the extension, a complete example of custom model integration is reported. This example illustrates how a pre-trained model can be incorporated into ModalityConverter following the modular design principles described in the previous sections, without requiring modifications to the core architecture of the extension. This example can be replicated in the context of the SIQA module.

The integration workflow is structured around three fundamental components: (i) the implementation of a dedicated model class, (ii) the registration of the model through metadata, and (iii) the dynamic loading of the pre-trained model file. This design enforces a strict separation between model-specific logic and the general inference infrastructure, ensuring maintainability, reproducibility, and long-term extensibility.

Model class implementation

A new Python class is created within the *ModalityConverterLib/ModelsImpl/* directory and inherits from the abstract `BaseModel` class, reported in Listing 11.1. The model is registered through the `@register_model` decorator (see Listing 11.1), which associates a unique identifier to the implementation and enables automatic discovery by the framework.

```

1 from abc import ABC, abstractmethod
2 import os
3 import slicer
4 from slicer import vtkMRMLScalarVolumeNode
5 from ModalityConverterLib.UI.utils import PRINT_MODULE_SUFFIX
6
7 # Global registry for model classes
8 MODEL_REGISTRY = {}
9
10 def register_model(key):
11     """Decorator to register a model class with a specific key in the
12     metadata.json."""
13     def decorator(cls):
14         print(f"{PRINT_MODULE_SUFFIX} Registering model '{key}' with
15         class {cls.__name__}")
16         MODEL_REGISTRY[key] = cls
17         return cls
18     return decorator
19
20 """Base class for all models in the ModalityConverter library."""
21 class BaseModel(ABC):
22     def __init__(self, modelKey: str, device: str = "cpu"):
23         self.modelKey = modelKey
24         self.model = None
25         self.baseModelsDir = os.path.join(os.path.dirname(__file__), '../
26         Resources/Models')

```

```

24     self.modelsDir = os.path.abspath(self.baseModelsDir)
25     self.device = device.lower()
26
27     def loadModel(self):
28         import json
29         import requests
30         from urllib.parse import urlparse
31
32         modelMetadataPath = os.path.join(self.modelsDir, "metadata.json")
33
34         # Ensure the models directory exists
35         if not os.path.exists(self.modelsDir):
36             os.makedirs(self.modelsDir)
37
38         # Search the key file by unique key
39         for file in os.listdir(self.modelsDir):
40             if file.startswith(self.modelKey):
41                 self.modelPath = os.path.join(self.modelsDir, file)
42                 break
43         else:
44             self.modelPath = None # No files found with the specified
key
45
46         # Download model if not present locally
47         if self.modelPath is None or not os.path.exists(self.modelPath):
48             # Load model metadata
49             if not os.path.exists(modelMetadataPath):
50                 raise FileNotFoundError(f"Model metadata file not found
at {modelMetadataPath}")
51
52             with open(modelMetadataPath, "r") as f:
53                 modelMetadata = json.load(f)
54
55             if self.modelKey not in modelMetadata:
56                 raise ValueError(f"Model key '{self.modelKey}' not found
in metadata file.")
57
58             print(f"{PRINT_MODULE_SUFFIX} Model '{self.modelKey}' not
found locally. Downloading...")
59             slicer.app.processEvents()
60             url = modelMetadata[self.modelKey]["url"]
61
62             _, ext = os.path.splitext(urlparse(url).path)
63
64             # Use the metadata key as a filename, keeping the original
extension
65             self.modelPath = os.path.join(self.modelsDir, self.modelKey +
ext)
66
67             response = requests.get(url, stream=True)
68             response.raise_for_status()
69

```

```

70         with open(self.modelPath, "wb") as modelFile:
71             for chunk in response.iter_content(chunk_size=8192):
72                 modelFile.write(chunk)
73
74         # Load the model with custom subclasses method
75         self.model = self._loadModelFromPath(self.modelPath)
76
77         @abstractmethod
78         def _loadModelFromPath(self, modelPath: str):
79             """Load the model from the given path. Will be called in the
80             load_data method. To be implemented in subclasses for loading
81             customization."""
82             pass
83
84         @abstractmethod
85         def runInference(self, inputVolume: vtkMRMLScalarVolumeNode,
86                         outputVolume: vtkMRMLScalarVolumeNode, inputMask:
87                         vtkMRMLScalarVolumeNode=None, showAllFiles: bool=True):
88             pass

```

Listing 11.1: The BaseModel class implementation.

An example of a minimal model implementation is reported in Listing 11.2. After registering the model with a unique key, the user implements the model loading logic within the `_loadModelFromPath` method. This loading procedure may vary depending on the underlying framework used to train the model; a common and interoperable choice is the adoption of standardised model distribution formats such as Open Neural Network Exchange (ONNX).

Subsequently, the user develops the model inference logic within the `runInference` method, which by design provides all the core elements required for the conversion process, including the input image, the output node for result storage, and the optional region-of-interest mask, as defined by the user through the graphical interface. This method therefore encapsulates the complete processing pipeline, where preprocessing, forward inference, and output visualisation logic are explicitly defined.

```

1 from ModalityConverterLib.ModelBase import BaseModel, register_model
2
3 @register_model("example_model_key")
4 class ExampleModel(BaseModel):
5     def __init__(self, modelKey: str, device: str = "cpu"):
6         super().__init__(modelKey, device)
7
8     def _loadModelFromPath(self, modelPath):
9         # Load the pre-trained model (e.g. torch, onnxruntime, etc.)
10        model = ...
11        return model
12
13    def runInference(self, inputVolume, outputVolume,
14                    inputMask=None, showAllFiles=True):
15        # Custom preprocessing logic
16        # ...

```

```

17         # Custom inference logic
18         # ...
19
20         # Show output in Slicer scene
21         # ...
22

```

Listing 11.2: Example of custom model class implementation.

This design enforces a standardised interface across heterogeneous model architectures, allowing different DL frameworks and model formats to coexist within the same execution environment.

Metadata-driven model registration

Model availability and configuration are defined through a JSON metadata file. Each model is described by a unique identifier, a download URL for the pre-trained weights, a display name, and a structured description including input modality, preprocessing steps, output modality, and citation information. An example metadata entry is reported in Listing 11.3. Notably, the model key corresponds to the one reported in the `@register_model` decorator (see Listing 11.2).

```

1 "example_model_key": {
2   "url": "https://example.com/example_model.onnx",
3   "display_name": "Example Model",
4   "description": "Example model for image-to-image translation.",
5   "module_name": "ExampleModel",
6   "mask_required": false
7 }

```

Listing 11.3: Example of model metadata registration.

This metadata-driven approach enables automatic population of the graphical interface and dynamic linking between the user interface and the corresponding backend implementation, without requiring manual modification of the GUI code.

11.2.3.1 Dynamic loading and execution

When a model is selected through the graphical interface, the extension automatically resolves the corresponding metadata entry, loads the model class into memory using the associated implementation class, and downloads the pre-trained model file if it is not already present locally. The inference pipeline of the selected model is subsequently executed, generating synthetic volumes that are directly instantiated within the 3D Slicer scene. This functionality is encapsulated within the `ModalityConverterLogic` class, as reported in Listing 11.4.

```

1 class ModalityConverterLogic(ScriptedLoadableModuleLogic):
2     def __init__(self) -> None:

```

```

3         """Called when the logic class is instantiated. Can be used for
initializing member variables."""
4         ScriptedLoadableModuleLogic.__init__(self)
5         self.model = None
6
7     def getParameterNode(self):
8         return ModalityConverterParameterNode(super().getParameterNode())
9
10    def getModelInstance(self, selectedModelModuleName, selectedModelKey,
basePackage="ModalityConverterLib.ModelsImpl", device="cpu"):
11        """
12        Dynamically loads and returns an instance of a model class based
on the provided module name and model key.
13
14        Parameters:
15        - selectedModelModuleName (str): The name of the module
containing the model class.
16        - selectedModelKey (str): The key identifying the model in the
model registry.
17        - basePackage (str): The base package path where the model
modules are located. Defaults to "ModalityConverterLib.ModelsImpl".
18
19        Returns:
20        - object: An instance of the model class corresponding to the
provided key.
21
22        Raises:
23        - ImportError: If the specified module cannot be imported.
24        - ValueError: If the model key is not found in the model registry
.
25        """
26
27        import importlib
28        from ModalityConverterLib.ModelBase import MODEL_REGISTRY
29
30        # Importing the model module dynamically will trigger the
registration of the model class in the model registry
31        fullModulePath = f"{basePackage}.{selectedModelModuleName}"
32        try:
33            logging.info(f"Attempting to import module: {fullModulePath}"
)
34            importlib.import_module(fullModulePath)
35        except ImportError as e:
36            raise ImportError(f"Could not import module '{fullModulePath
}' for model '{selectedModelKey}': {str(e)}")
37
38        if selectedModelKey not in MODEL_REGISTRY:
39            raise ValueError(f"Model key '{selectedModelKey}' not found
in the model registry. Please check the metadata.json file and ensure
the key is correctly defined.")
40
41        model_class = MODEL_REGISTRY[selectedModelKey]

```

```

42         return model_class(selectedModelKey, device)
43
44
45     def process(self,
46                 inputVolume: vtkMRMLScalarVolumeNode,
47                 outputVolume: vtkMRMLScalarVolumeNode,
48                 maskVolume: vtkMRMLScalarVolumeNode,
49                 selectedModelKey: str = "",
50                 selectedModelModuleName: str = None,
51                 showAllFiles: bool = True,
52                 device: str = "cpu") -> None:
53
54         if not inputVolume or not outputVolume:
55             raise ValueError("Input or output volume is invalid")
56
57         modelInstance = self.getModelInstance(selectedModelModuleName,
58                                                selectedModelKey, device=device)
59         modelInstance.loadModel()
60
61         modelInstance.runInference(inputVolume=inputVolume, inputMask=
62                                   maskVolume, outputVolume=outputVolume, showAllFiles=showAllFiles)

```

Listing 11.4: ModalityConverter logic class. Upon running the inference, the `process` method orchestrates the execution by invoking the backend implementation of the selected custom model, leveraging the modular architecture of the extension.

This integration example demonstrates that the proposed framework supports a fully modular and scalable model deployment strategy. New models can be incorporated through localised additions of code and metadata, without affecting existing implementations or the global software structure.

From a translational perspective, this design enables the rapid dissemination of research models into a usable research tool, effectively bridging the gap between algorithmic development and real-world usability.

11.3 Illustrative example

To demonstrate the practical use of the ModalityConverter extension, an illustrative example is presented. The MRHead volume used in this example is a T1w MRI, which is distributed with 3D Slicer and can be downloaded via the “Download Sample Data” module. This MRHead volume provides a useful test case as it was not included in the training data used for the FedSynthCT-Brain models. Therefore, it enables a qualitative assessment of the model’s ability to apply to previously unseen data, while also demonstrating a realistic usage scenario of the ModalityConverter extension.

Figure 11.2 shows the input configuration within the 3D Slicer environment. The MRHead volume was selected as the input volume, in combination with a user-defined segmentation representing the region-of-interest (ROI) mask. The specified ROI constrained the inference process, such that the sCT output was generated within the selected anatomical region. In this particular instance, the ROI was confined to the brain, corresponding to the anatomical region targeted by the FedSynthCT-Brain models.

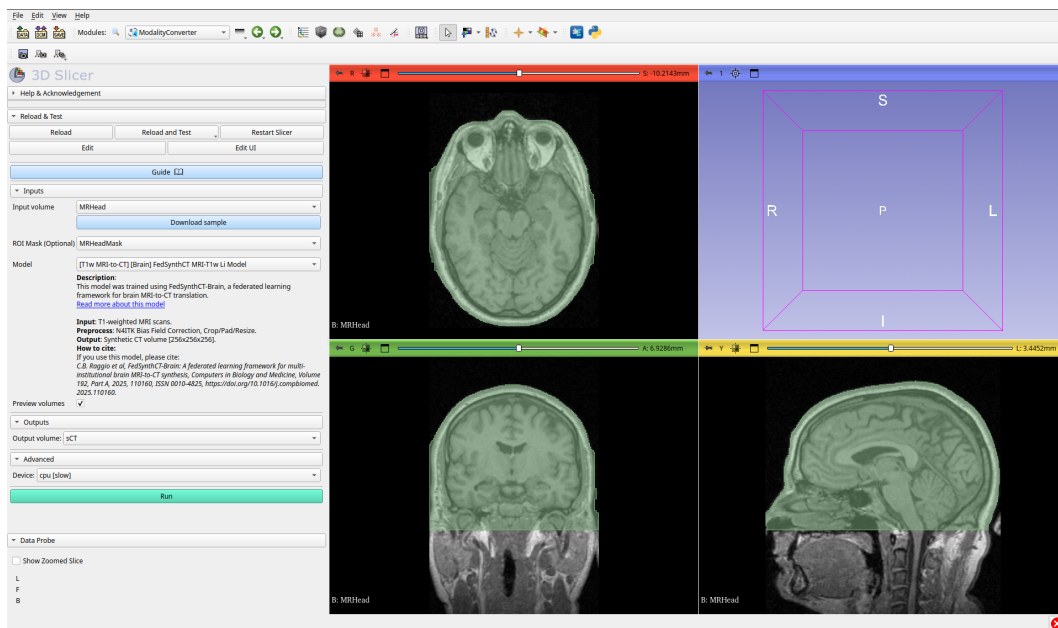


Figure 11.2: Example input to the ModalityConverter extension in 3D Slicer. The MRHead volume and a segmentation representing the region-of-interest mask are provided as inputs to the extension for inference with a FedSynthCT-Brain model.

The resulting sCT volume produced by the FedSynthCT-Brain model is presented in Figure 11.3. As demonstrated in Figure 1, the generated sCT was correctly constrained to the ROI specified in the input selectors. Upon completion of the inference process, the output volume was displayed in the 3D Slicer environment, including the corresponding process

messages in the integrated terminal. Despite the MRHead dataset not being included in the model training set, the generated sCT demonstrated anatomically coherent structures and intensity patterns that were consistent with the anticipated CT appearance.

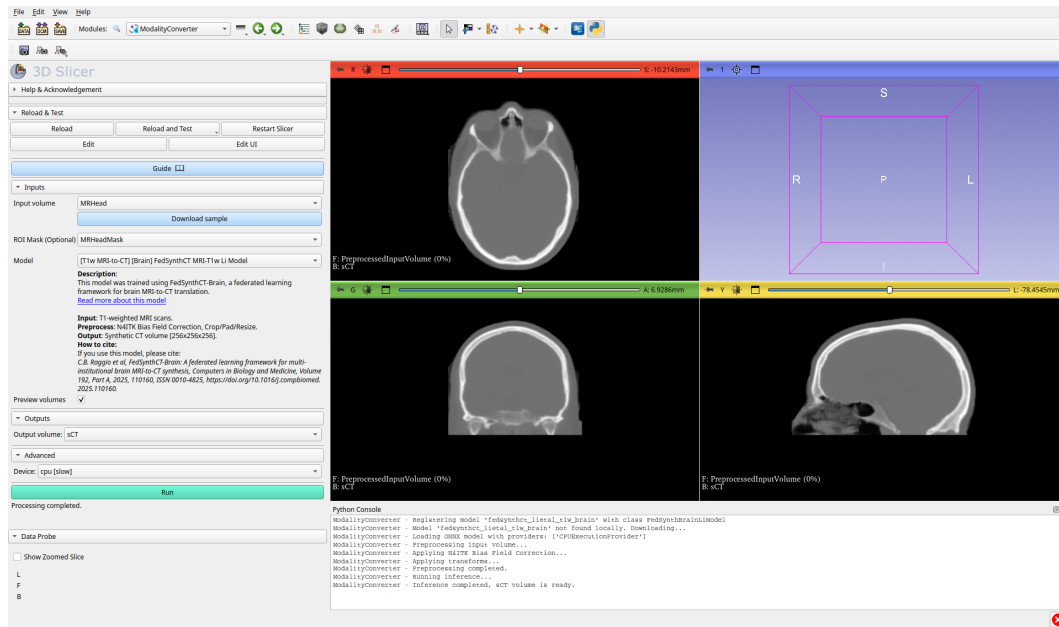


Figure 11.3: Example of MRHead conversion. The sCT output volume generated by the FedSynthCT-Brain models is shown. The generated sCT is limited to the region of interest specified in the input field.

Furthermore, a side-by-side comparison between the original MRHead input and the corresponding sCT is provided in Figure 11.4. This qualitative comparison highlighted the structural correspondence between modalities and illustrated the capability of the model to preserve relevant anatomical features during cross-modality translation.

Finally, Figure 11.5 shows three-dimensional renderings of the original MRHead volume and the generated sCT skull in 3D Slicer. Both renderings were automatically generated using the Volume Rendering module in 3D Slicer. This visualisation further emphasises the spatial consistency of the synthetic volume and ModalityConverter’s effective integration into the standard 3D Slicer workflow.

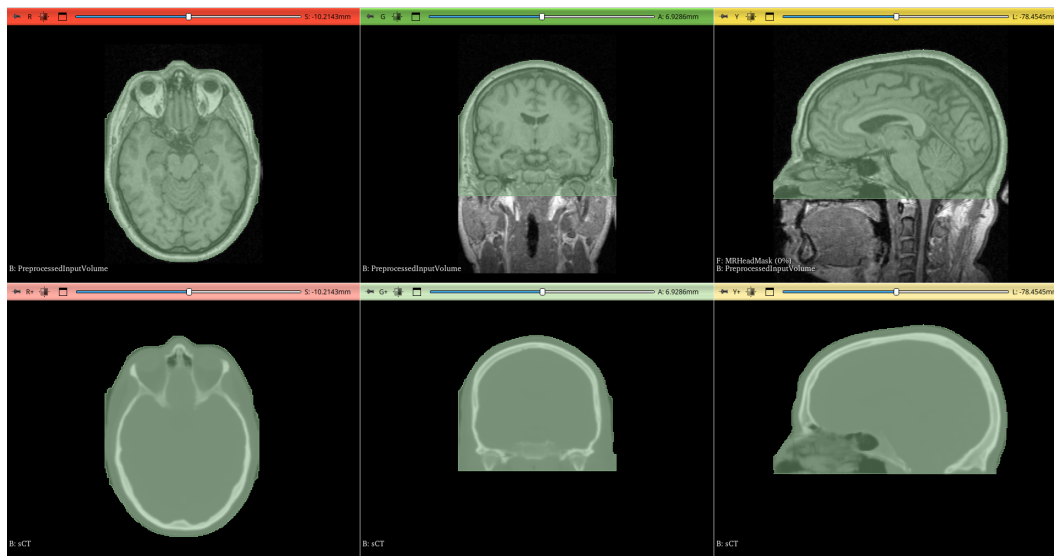


Figure 11.4: Comparison between the MRHead input volume and the corresponding sCT generated by the FedSynthCT-Brain model.

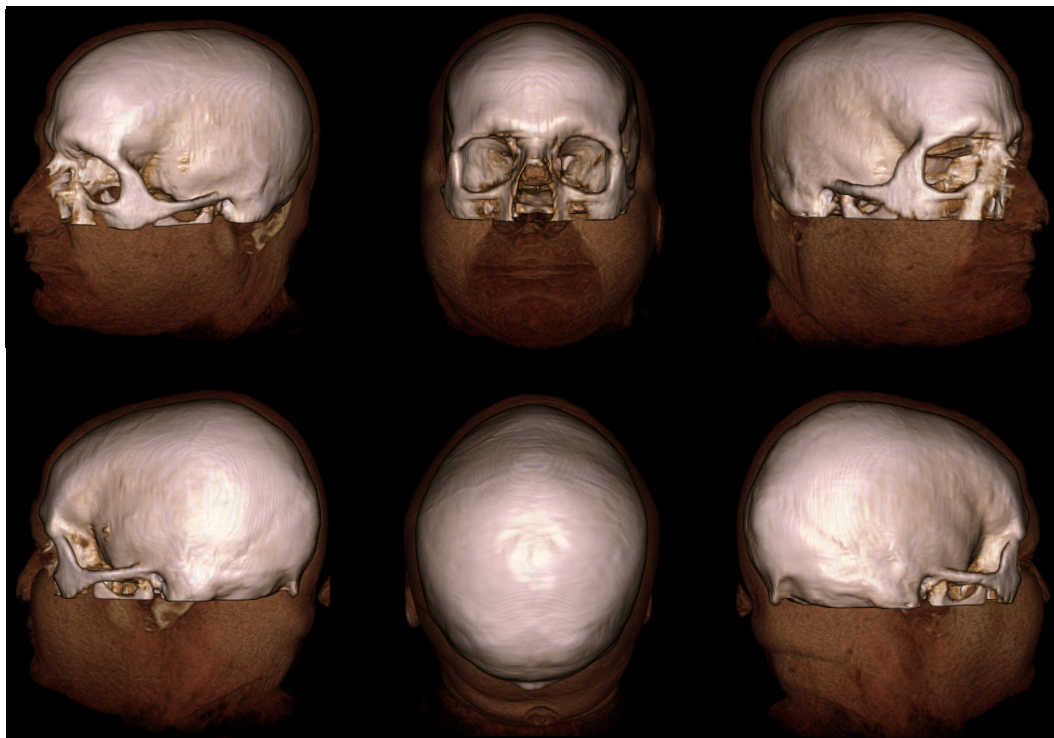


Figure 11.5: Rendering of the MRHead volume and the skull from the generated sCT. Both were automatically generated via the Volume Rendering 3D Slicer module.

11.4 Conclusions

The ModalityConverter extension presented in this chapter represents the translational application of the decentralised models developed throughout this research. While the core scientific contributions of this thesis lie in the design and validation of FL frameworks, privacy-preserving optimisation strategies, and decentralised infrastructures for healthcare, the extension operationalises these methodological advances into an accessible and reproducible research tool.

The creation of the extension and integration of FL-trained MRI-to-sCT and CBCT-to-sCT models into the 3D Slicer environment represents a key contribution of the thesis, marking the evolution of algorithmic research prototypes into deployable scientific instruments. Furthermore, the integration of supplementary centralised I2I translation models, such as CT-to-sPET, retains the conceptual integrity of the thesis, situating the federated contributions within a broader and practically relevant research ecosystem. The extension, in essence, functions as a unified deployment framework for heterogeneous medical image translation models, thereby enabling comparative evaluation, modular expansion, and community-driven development. In this sense, decentralised AI is not regarded as an isolated methodological construct; rather, it is conceived as a foundational component of an open and interoperable scientific infrastructure.

The evolution of the extension into a dual-module architecture, comprising both the ModalityConverter module for image-to-image translation and the Synthetic Image Quality Assurance module for quality assessment, strengthens its translational impact. By adopting a unified abstraction layer, the infrastructure enables heterogeneous AI models—such as centralised, federated, generative and evaluative—to coexist within a consistent and scalable deployment strategy. Importantly, the inclusion of quality estimation models establishes a closed-loop workflow in which synthetic image generation and reliability assessment are seamlessly connected and accessible in a single extension. This design extends the concept of decentralisation beyond collaborative training, deployment, accessibility, reproducibility, and responsible evaluation.

Overall, the ModalityConverter ecosystem represents the culmination of the conceptual trajectory outlined in this thesis: from decentralised data governance and federated optimisation to decentralised scientific impact. By ensuring that federated models and associated quality assessment tools become accessible research assets within a globally adopted imaging platform, this work contributes not only to methodological innovation in healthcare AI, but also to its sustainable and transparent integration into real-world research practice.

PART V

KEY FINDINGS AND FUTURE DIRECTIONS

Discussion and Conclusion

This thesis investigated the application of decentralised learning strategies in healthcare, with a particular focus on the use of Federated Learning (FL) for medical Image-to-Image (I2I) translation and Synthetic Computed Tomography (sCT) generation. As outlined in Chapter 1, the objective was not only to assess the feasibility of FL in clinically relevant tasks, but also to address key barriers to its real-world adoption, including data heterogeneity, privacy constraints, computational efficiency, and robustness.

Across Parts II–IV, the work followed a structured progression: from establishing feasibility (Chapters 4–5), to addressing efficiency and reliability (Chapters 6–8), and finally to tooling and translational outcomes (Chapters 9–11). Each of these phases builds upon the previous one: the feasibility studies define the scope of application; the efficiency and security analyses identify and address the practical barriers to deployment; and the tooling contributions operationalise the methodological advances into accessible and reproducible instruments. Taken together, these studies provide a comprehensive perspective on the potential and the current limitations of federated approaches for medical image analysis.

The following sections summarise and discuss the main findings, focusing on their broader implications, limitations and relevance to the field.

12.1 Enabling Federated Learning for sCT Generation

A primary contribution of this thesis lies in demonstrating the feasibility of applying FL for sCT synthesis, a domain that remained comparatively underexplored with respect to FL-based classification, segmentation and cross-modality Magnetic Resonance Imaging (MRI) translation at the time of writing. In particular, the studies on MRI-to-sCT (Chapter 4) and Cone Beam Computed Tomography (CBCT)-to-sCT generation (Chapter 5) collectively show that federated training can be successfully applied to complex prediction tasks across multiple institutions, within realistic cross-silo settings.

Across both investigated settings, the federated models exhibited stable training behaviour despite heterogeneous data distributions, and maintained consistent performance when evaluated on independent external datasets. These findings suggest that collaborative training across institutions is not only viable, but may also contribute to mitigating domain shift effects, which is a major limitation of Artificial Intelligence (AI) in healthcare. Notably, the extension to CBCT-to-sCT translation in the head and neck region (Chapter 5) confirmed that the proposed framework is not restricted to a specific modality or acquisition protocol, and that consistent cross-site generalisation can be achieved without introducing centre-specific adaptations, even when using a different architectural paradigm (e.g. a Conditional Generative Adversarial Network (cGAN)).

At the same time, these results require careful interpretation. The MRI-to-sCT study involved five centres, while the CBCT-to-sCT study involved four publicly available institutional datasets, implementing a realistic cross-silo federation of four to five institutions. This configuration was sufficient to demonstrate feasibility; it constitutes, however, an initial step towards larger collaborative frameworks, providing a methodological and experimental foundation upon which more extensive federations can be built, with the potential for stronger generalisation as more institutions contribute data.

From a methodological perspective, both studies highlight that the feasibility of federated training relies on factors that are often implicit, such as data curation, preprocessing consistency, model architecture, and aggregation strategy selection. In decentralised settings, where direct access to data across sites is restricted, these aspects become even more critical and require careful coordination, as they cannot be easily inspected or corrected centrally.

Furthermore, qualitative analyses of the generated sCTs revealed that, despite favourable quantitative performance, localised errors and artefacts persist, particularly in anatomically complex regions. This observation aligns with the broader literature on sCT generation and indicates that FL does not inherently resolve the underlying modelling challenges associated with image synthesis, rather, it inherits these limitations while enabling collaborative training. Addressing these residual artefacts remains an open challenge and is likely to require architectural innovations or data-driven solutions that go beyond the choice of training paradigm.

A common limitation of both studies is the absence of associated radiotherapy planning data and analysis. However, this thesis primarily aims to establish the feasibility of a decentralised approach to sCT generation that has not previously been investigated in a decentralised setting. Indeed, an important contribution of this research work lies in extending the federated paradigm from cross-modality MRI translation to MRI-to-sCT and CBCT-to-sCT translation. Although dosimetric evaluation represents a prerequisite for radiotherapy applications, its feasibility will rely on the technical foundation established in this research. Therefore, a dosimetric analysis is the most immediate direction for subsequent work. Future studies based on the current framework will be able to focus on clinical validation, with the federated training methodology already in place.

A further remark concerns the absence of a centralised training baseline in the experimental comparisons. FL and centralised learning are complementary approaches suited to

different constraints, rather than competing paradigms. When data centralisation is feasible, centralised training may be preferable in order to avoid FL disadvantages. In contrast, the federated approach is designed for scenarios where regulatory, ethical or practical barriers render centralisation impractical, which are the conditions that motivated this research. Therefore, a performance comparison against a centralised model would be misleading as it would combine two fundamentally different problem settings and would not constitute a meaningful evaluation criterion for the proposed framework. In this context, the important question is not whether FL matches centralised performance (a challenging achievement given the emergent application), but whether it produces models of sufficient quality to be clinically and scientifically useful when centralisation is not an option.

Overall, the findings of both studies establish sCT generation as a viable application domain for FL, while emphasising that effectiveness remains contingent on both data quality and experimental design.

12.2 Sustainability, Efficiency, and Security of Federated Learning

Beyond feasibility, this thesis investigated key practical aspects that govern the scalability and reliability of FL in real-world deployments, namely computational efficiency, communication constraints, and security against adversarial threats. While these topics are treated in separate chapters, they share a common underlying concern: ensuring that decentralised systems remain not only accurate, but also practical and trustworthy under realistic conditions.

Computational Efficiency A central observation arising from Chapter 4 is that FL for I2I translation is computationally demanding, due to the complexity of the models involved and the prolonged training times required to reach stable convergence. This has direct implications for real-world deployments, where resource limitations and potential system instability may hinder participation by institutions with constrained infrastructure, thereby undermining the inclusive nature of federated collaboration.

The strategy based on adaptive parameter freezing, proposed in Chapter 6, demonstrated that it is possible to substantially reduce computational load while maintaining stable performance. However, the reliance on predefined freezing configurations rather than on dynamic adaptation to client-specific characteristics, represents a limitation of the study. Developing approaches capable of responding to client heterogeneity without increasing local computational overhead remains an open challenge, and constitutes a viable direction for future investigation.

Security The security analysis presented in Chapter 7 challenged a common assumption in the FL literature, namely that decentralisation inherently provides privacy guarantees. The results demonstrated that, while gradient inversion attacks were not effective in reconstructing training data within the FedSynthCT framework, other attack vectors remain feasible. In

particular, membership inference attacks (e.g. FedMIA [125]) and poisoning strategies pose realistic threats that could compromise data confidentiality, model integrity, or both. The introduction of mitigation mechanisms, including anomaly detection and model update validation, demonstrated potential to enhance robustness, albeit at the cost of increased system complexity and potential impacts on convergence.

Thus, the main insight from this analysis is that trustworthy deployment in clinical environments requires additional infrastructural layers beyond the federated algorithm itself. These include authentication protocols, secure communication channels, certificate-based communication, and pre-aggregation validation mechanisms. Importantly, these components fall outside the scope of the FL algorithm by design. They are system-level solutions that do not affect the performance or the feasibility of the federated training procedure, but rather define the broader secure infrastructure within which it must operate. In healthcare, where regulatory requirements are stringent and failures can have serious consequences, robustness and security should be regarded as integral design criteria rather than as post hoc considerations.

Communication Efficiency Chapter 8 shifts focus towards the optimisation of communication overhead in FL, a problem of particular relevance in settings characterised by limited bandwidth or heterogeneous network infrastructure. The analysis of FedWSOcomp [143]—combining compression, sparsification, and weight clustering—demonstrated that moderate levels of these techniques can substantially reduce communication overhead while preserving segmentation accuracy. Conversely, more aggressive compression leads to performance degradation and reduced convergence stability, delineating a practical operational range for the method.

Notably, this chapter focuses on a segmentation task rather than image translation. Segmentation represents a mature application domain for FL, where foundational feasibility questions have already been settled in the literature. Consequently, this makes it a controlled setting in which to isolate and systematically evaluate communication efficiency strategies. Furthermore, when federating segmentation models, practical bottlenecks are the communication overhead and convergence trend rather than computational demand, making optimisation of these the most meaningful target. Nevertheless, the findings are directly relevant to I2I translation settings, in which communication efficiency will become increasingly important as federated frameworks incorporate more institutions.

A key limitation of this study lies in the exclusive investigation of consistent but fixed configurations. The potential benefits of adaptive strategies, capable of dynamically adjusting sparsification and clustering parameters during training, remain unexplored and represent a promising direction for future research.

Taken together, the findings across these three areas highlight that efficiency and security are not secondary concerns, but integral components of any federated system intended for clinical deployment. Ensuring reliable and scalable operation requires not only robust models,

but also a holistic design approach that encompasses computational, communicational, and security considerations simultaneously.

12.3 Tooling, Reproducibility, and Research Translation

In addition to methodological contributions, this thesis addressed the gap between research development and practical usability through the design and release of open-source tools integrated into the 3D Slicer platform [107]. These contributions (presented in Chapters 9-11) reflect a coherent translational vision: ensuring that the methodological advances developed throughout this thesis are not confined to experimental codebases, but are made accessible, reproducible, and usable by the research community. Indeed, the three contributions in this part of the thesis address complementary aspects of the research-to-practice pipeline: data preparation, quality control, and model deployment.

ImageAugmenter [146] (Chapter 9) provides a user-friendly graphical interface for applying more than 20 spatial, intensity, and noise-based transformations to medical imaging datasets, including support for segmentation masks, GPU acceleration, and dataset-level batch processing. In contrast with existing solutions such as the TorchIO 3D Slicer extension, which offers a limited set of basic transformations and requires manual volume loading, ImageAugmenter enables simultaneous application of multiple transforms to entire datasets, with real-time preview and direct export preserving the original file hierarchy. By removing the requirement for programming expertise, the tool lowers the barrier to adopting state-of-the-art augmentation techniques, promotes methodological consistency across studies and institutions, and is directly applicable to FL settings, where local data augmentation can partially compensate for data scarcity and class imbalance at individual clients.

The sCT quality control framework [55] (Chapter 10) addresses one of the main barriers to the clinical adoption of Deep Learning (DL)-based image synthesis: the inability to assess the quantitative error of synthetic outputs in the absence of ground-truth Computed Tomography (CT) data. Results demonstrated small absolute deviations between predicted and true Mean Absolute Error (MAE) values, with no evidence of systematic bias, and inference times compatible with routine clinical workflows. Although developed in a centralised setting, this framework directly complements the federated synthesis models proposed earlier in the thesis, providing patient-specific reliability estimation for models deployed in real clinical environments where reference data are unavailable.

ModalityConverter (Chapter 11) integrates both the federated synthesis models and the quality assessment tool into a unified 3D Slicer extension, thereby operationalising the full research pipeline within a widely adopted and open-source medical research platform. The extension comprises two complementary modules: the *ModalityConverter* module, which provides GPU-accelerated inference for I2I translation across multiple modalities

(including MRI-to-sCT, CBCT-to-sCT, and CT-to-sPET); and the *Synthetic Image Quality Assurance* module, which integrates the MAE prediction framework to enable immediate quality estimation of sCT outputs. This design definitely a closed-loop workflow in which synthetic image generation and quality assessment are seamlessly connected within a single environment, without requiring specialised infrastructure or programming expertise.

The distinction between these three tools is important. ImageAugmenter operates at the data preparation stage, before model training; the quality control framework operates at the deployment stage, after model inference; and ModalityConverter provides the deployment infrastructure itself. Together, they cover the full lifecycle of a synthetic medical image application, from dataset construction to model deployment and output validation.

A limitation of the tooling contributions concerns the current scope of the integrated models. ModalityConverter currently embeds a limited number of pre-trained models, and the quality assurance module supports a single modality and a single error metric (MAE) for head and neck sCT. Extending both the model coverage and the quality assessment capabilities to additional modalities and anatomical regions represents a possible direction for future development.

12.4 Limitations and Future Directions

While the individual studies provide focused insights into specific aspects of FL for medical imaging, emerging challenges and opportunities define a coherent path for future research across the overall work. As noted in Section 12.1, the absence of radiotherapy planning data in the investigated datasets prevented a direct dosimetric assessment of the generated sCTs. This remains the most immediate limitation of the studies proposed in this thesis and should be the primary focus of near-term follow-up work. Integrating treatment planning information would enable evaluation in terms of dose calculation accuracy and clinical acceptability, which are the ultimate criteria for the future adoption in radiotherapy workflows.

Scalability to larger federations represents a second priority. The present studies involved four to five participating institutions. This configuration was sufficient to demonstrate feasibility, but not to characterise the overall behaviour of federated systems under greater institutional heterogeneity. Expanding the federation to a larger number of centres, with more diverse acquisition protocols and patient populations, would allow a more systematic assessment of generalisation and robustness, and would strengthen the translational relevance of the findings.

Furthermore, the reliance on fixed configurations for parameter freezing, sparsification, and clustering (identified across Chapters 6 and 8) represents a limiting factor that can be addressed in future studies. Developing adaptive strategies capable of responding dynamically to client-specific characteristics and communication constraints would improve robustness in heterogeneous settings and reduce the need for manual configurations.

Finally, the integration of federated models into clinical workflows requires standardised evaluation protocols, interoperability with hospital information systems, and community-

adopted benchmarks specific to federated sCT evaluation. The open-source tools developed in this thesis represent a concrete step towards this goal, providing accessible infrastructure for model deployment and output quality assessment. Extending their scope, in terms of supported modalities, anatomical regions, and quality metrics, constitutes a parallel and complementary development track.

Overall, these directions define a clear research path progressing from technical feasibility, already established by the present work, towards clinical validation and operational integration, which are necessary conditions for the successful and responsible deployment of FL in healthcare.

12.5 Conclusion

The present work has been motivated by the identification of a gap in the existing literature. While DL models for medical imaging require large and heterogeneous training datasets to generalise reliably, the regulatory, institutional, and practical barriers to data centralisation in healthcare are substantial and unlikely to diminish.

FL was identified as a potential solution to this problem. However, its application to medical I2I translation, and to sCT generation in particular, had not been systematically investigated prior to the current work.

The studies presented in this thesis provide evidence that federated training is a viable approach for this class of tasks, and that the challenges specific to I2I translation do not preclude effective decentralised collaboration across institutions. Furthermore, the work clarifies that feasibility is a necessary but insufficient condition for deployment, and that computational efficiency, communication overhead, and security against adversarial threats must be addressed as integral components of the system design rather than as secondary considerations.

The open opportunities and challenges represent the natural extension of this research. In order to address them, it is necessary to implement both methodological advances and multi-institutional collaboration, which FL itself is designed to enable. The present thesis contributes the methodological, analytical, and software foundation upon which this next phase can be built, and demonstrates that decentralised learning is not only a theoretically appealing but also a practically viable paradigm for collaborative medical image analysis.

References

- [1] M. F. Spadea, M. Maspero, P. Zaffino, and J. Seco, “Deep learning based synthetic-CT generation in radiotherapy and PET: A review,” *Medical Physics*, vol. 48, pp. 6537–6566, 2021. doi:<https://doi.org/10.1002/mp.15150>
- [2] L. Pinto-Coelho, “How Artificial Intelligence Is Shaping Medical Imaging Technology: A Survey of Innovations and Applications,” *Bioengineering*, vol. 10, 2023. doi:10.3390/bioengineering10121435
- [3] S. Dayarathna, K. T. Islam, S. Uribe, G. Yang, M. Hayat, and Z. Chen, “Deep learning based synthesis of MRI, CT and PET: Review and analysis,” *Medical Image Analysis*, vol. 92, p. 103046, 2024. doi:<https://doi.org/10.1016/j.media.2023.103046>
- [4] A. S. Panayides, H. Chen, N. D. Filipovic, et al., “Position Paper: Artificial Intelligence in Medical Image Analysis: Advances, Clinical Translation, and Emerging Frontiers,” *IEEE Journal of Biomedical and Health Informatics*, pp. 1–16, 2025. doi:10.1109/JBHI.2025.3649496
- [5] “Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance),” pp. 1–88, 2016. <http://data.europa.eu/eli/reg/2016/679/oj>
- [6] D. Peloquin, M. DiMaio, B. Bierer, and M. Barnes, “Disruptive and avoidable: GDPR challenges to secondary research uses of data,” *European Journal of Human Genetics*, vol. 28, pp. 697–705, 2020. doi:<https://doi.org/10.1038/s41431-020-0596-x>
- [7] M. Kutylowski, A. Lauks-Dutka, and M. Yung, “GDPR – Challenges for Reconciling Legal Rules with Technical Reality,” in *Computer Security – ESORICS 2020*, L. Chen, N. Li, K. Liang, and S. Schneider, Eds. Cham: Springer International Publishing, 2020, pp. 736–755.
- [8] B. Wouters, D. Shaw, C. Sun, et al., “Putting the GDPR into practice: Difficulties and uncertainties experienced in the conduct of big data health research,” *European Data Protection Law Review*, vol. 7, pp. 206–216, 2021.
- [9] A. Bernier, F. Molnár-Gábor, B. M. Knoppers, et al., “Reconciling the biomedical data commons and the GDPR: three lessons from the EUCAN ELSI collaboratory,” *European Journal of Human Genetics*, vol. 32, pp. 69–76, 2024.
- [10] Y. Liu, A. Chen, H. Shi, et al., “CT synthesis from MRI using multi-cycle GAN for head-and-neck radiation therapy,” *Computerized Medical Imaging and Graphics*, vol. 91, p. 101953, 2021. doi:<https://doi.org/10.1016/j.compmedimag.2021.101953>
- [11] R. Graf, J. Schmitt, S. Schlaeger, et al., “Denoising diffusion-based MRI to CT image translation enables automated spinal segmentation,” *European Radiology Experimental*, vol. 7, p. 70, 2023. doi:10.1186/s41747-023-00385-2

- [12] M. Özbey, O. Dalmaz, S. U. Dar, et al., “Unsupervised medical image translation with adversarial diffusion models,” *IEEE Transactions on Medical Imaging*, vol. 42, pp. 3524–3539, 2023. doi:10.1109/TMI.2023.3290149
- [13] M. A. Bahloul, S. Jabeen, S. Benoumhani, H. A. Alsaleh, Z. Belkhatir, and A. Al-Wabil, “Advancements in synthetic CT generation from MRI: A review of techniques, and trends in Radiation therapy planning,” *Journal of Applied Clinical Medical Physics*, vol. 25, 2024. doi:10.1002/acm2.14499
- [14] O. Dalmaz, M. U. Mirza, G. Elmas, et al., “One model to unite them all: Personalized federated learning of multi-contrast MRI synthesis,” *Medical Image Analysis*, vol. 94, p. 103121, 2024. doi:https://doi.org/10.1016/j.media.2024.103121
- [15] M. Li, P. Xu, J. Hu, Z. Tang, and G. Yang, “From challenges and pitfalls to recommendations and opportunities: Implementing federated learning in healthcare,” *Medical Image Analysis*, vol. 101, p. 103497, 2025. doi:https://doi.org/10.1016/j.media.2025.103497
- [16] E. Markodimitrakakis, E. Trivizakis, G. S. Ioannidis, et al., “Federated learning on magnetic resonance imaging: a critical review,” *Artificial Intelligence Review*, 2026.
- [17] M. Adnan, S. Kalra, J. C. Cresswell, G. W. Taylor, and H. R. Tizhoosh, “Federated learning and differential privacy for medical image analysis,” *Scientific reports*, vol. 12, p. 1953, 2022.
- [18] H. Guan, P.-T. Yap, A. Bozoki, and M. Liu, “Federated learning for medical image analysis: A survey,” *Pattern Recognition*, p. 110424, 2024.
- [19] H. Mahmood, Z. Alamgir, S. T. Javed, S. Karim, and M. Awais, “Federated Generative Models in Medical Imaging: Current Advances, Challenges, and Future Directions,” *IEEE Access*, vol. 14, pp. 5197–5217, 2026. doi:10.1109/ACCESS.2026.3650810
- [20] M. Boulanger, J.-C. Nunes, H. Chourak, et al., “Deep learning methods to generate synthetic CT from MRI in radiotherapy: A literature review,” *Physica Medica*, vol. 89, p. 265–281, 2021. doi:10.1016/j.ejmp.2021.07.027
- [21] S. Hussain, I. Mubeen, N. Ullah, et al., “Modern Diagnostic Imaging Technique Applications and Risk Factors in the Medical Field: A Review,” *BioMed Research International*, vol. 2022, p. 5164970, 2022. doi:https://doi.org/10.1155/2022/5164970
- [22] X. Ou, X. Chen, X. Xu, et al., “Recent Development in X-Ray Imaging Technology: Future and Challenges,” *Research*, vol. 2021, 2021. doi:10.34133/2021/9892152
- [23] G. Poludniowski, P. M. Evans, and S. Webb, “Rayleigh scatter in kilovoltage x-ray imaging: is the independent atom approximation good enough?” *Physics in Medicine Biology*, vol. 54, p. 6931, 2009. doi:10.1088/0031-9155/54/22/012
- [24] A. C. Kak and M. Slaney, *Principles of Computerized Tomographic Imaging*. Society for Industrial and Applied Mathematics, 2001. doi:10.1137/1.9780898719277
- [25] T. M. Buzug, *Computed Tomography: From Photon Statistics to Modern Cone-Beam CT*. Berlin, Germany: Springer, 2008.
- [26] W. A. Kalender, *Computed Tomography: fundamentals, system technology, image quality, applications*, 3rd ed. Publicis MCD Verlag, 2011.
- [27] C. H. McCollough, S. Leng, L. Yu, and J. G. Fletcher, “Dual- and Multi-Energy CT: Principles, Technical Approaches, and Clinical Applications,” *Radiology*, vol. 276, pp. 637–653, 2015, pMID: 26302388. doi:10.1148/radiol.2015142631
- [28] L. Zhu, J. Wang, and L. Xing, “Noise suppression in scatter correction for cone-beam CT,” *Medical Physics*, vol. 36, pp. 741–752, 2009. doi:https://doi.org/10.1118/1.3063001

- [29] L. Zhu, Y. Xie, J. Wang, and L. Xing, "Scatter correction for cone-beam CT in radiation therapy," *Medical Physics*, vol. 36, pp. 2258–2268, 2009. doi:<https://doi.org/10.1118/1.3130047>
- [30] I. Fotina, J. Hopfgartner, M. Stock, T. Steininger, C. Lütgendorf-Caucig, and D. Georg, "Feasibility of CBCT-based dose calculation: Comparative analysis of HU adjustment techniques," *Radiotherapy and Oncology*, vol. 104, pp. 249–256, 2012. doi:<https://doi.org/10.1016/j.radonc.2012.06.007>
- [31] T. Kaasalainen, M. Ekholm, T. Siiskonen, and M. Kortensniemi, "Dental cone beam CT: An updated review," *Physica Medica*, vol. 88, pp. 193–217, 2021. doi:<https://doi.org/10.1016/j.ejmp.2021.07.007>
- [32] H. Liu, D. Schaal, H. Curry, et al., "Review of cone beam computed tomography based online adaptive radiotherapy: current trend and future direction," *Radiat. Oncol.*, vol. 18, p. 144, 2023.
- [33] A. Thummerer, E. van der Bijl, A. J. Galapon, et al., "Synthrad2025 grand challenge dataset: Generating synthetic cts for radiotherapy from head to abdomen," *Medical Physics*, vol. 52, p. e17981, 2025. doi:<https://doi.org/10.1002/mp.17981>
- [34] D. W. Townsend, "Multimodality imaging of structure and function," *Physics in Medicine & Biology*, vol. 53, p. R1, 2008. doi:[10.1088/0031-9155/53/4/R01](https://doi.org/10.1088/0031-9155/53/4/R01)
- [35] M. E. Phelps, "PET: The Merging of Biology and Imaging into Molecular Imaging," *Journal of Nuclear Medicine*, vol. 41, pp. 661–681, 2000. <https://jnm.snmjournals.org/content/41/4/661>
- [36] M. M. Khalil, *Positron Emission Tomography (PET): Physics and Instrumentation*. Cham: Springer International Publishing, 2021, pp. 289–318. doi:[10.1007/978-3-030-65245-6_12](https://doi.org/10.1007/978-3-030-65245-6_12)
- [37] J. W. Fletcher, B. Djulbegovic, H. P. Soares, et al., "Recommendations on the Use of 18F-FDG PET in Oncology," *Journal of Nuclear Medicine*, vol. 49, pp. 480–508, 2008. doi:[10.2967/jnumed.107.047787](https://doi.org/10.2967/jnumed.107.047787)
- [38] C. D. Fuller, H. Elhalawani, and A. Mohamed, "MICCAI 2018 – Computational Precision Medicine Challenge: 18F-FDG PET Radiomics Risk Stratifiers in Head and Neck Cancer," 2021. doi:[10.6084/m9.figshare.15075195.v2](https://doi.org/10.6084/m9.figshare.15075195.v2)
- [39] J. A. Thie, "Understanding the Standardized Uptake Value, Its Methods, and Implications for Usage," *Journal of Nuclear Medicine*, vol. 45, pp. 1431–1434, 2004. <https://jnm.snmjournals.org/content/45/9/1431>
- [40] P. E. Kinahan and J. W. Fletcher, "Positron Emission Tomography-Computed Tomography Standardized Uptake Values in Clinical Practice and Assessing Response to Therapy," *Seminars in Ultrasound, CT and MRI*, vol. 31, pp. 496–505, 2010, pET-CT in RT Planning and Assessment. doi:<https://doi.org/10.1053/j.sult.2010.10.001>
- [41] ICRP, "The 2007 recommendations of the international commission on radiological protection. ICRP publication 103," *Ann. ICRP*, vol. 37, pp. 1–332, 2007.
- [42] J. Han, J. Zuo, G. Zillante, R. Chang, and L. Du, "A systematic review of PET circularity technologies and management strategies: Challenges and future directions," *Resources, Conservation and Recycling*, vol. 219, p. 108280, 2025. doi:<https://doi.org/10.1016/j.resconrec.2025.108280>
- [43] I. Hilschenz, R. Körber, H.-J. Scheer, et al., "Magnetic resonance imaging at frequencies below 1 khz," *Magnetic Resonance Imaging*, vol. 31, pp. 171–177, 2013. doi:<https://doi.org/10.1016/j.mri.2012.06.014>
- [44] R. W. Brown, Y. C. Norman Cheng, E. M. Haacke, M. R. Thompson, and R. Venkatesan, *enMagnetic resonance imaging*, 2nd ed. Hoboken, NJ: Wiley-Blackwell, 2014.

- [45] T. S. Sindhu, N. Kumaratharan, and P. Anandan, "A review of magnetic resonance imaging and its clinical applications," in *2022 6th International Conference on Devices, Circuits and Systems (ICDCS)*, 2022, pp. 38–42. doi:10.1109/ICDCS54290.2022.9780834
- [46] D. J. Pennell, "Cardiovascular Magnetic Resonance," *Circulation*, vol. 121, pp. 692–705, 2010. doi:10.1161/CIRCULATIONAHA.108.811547
- [47] R. Bammer, "Basic principles of diffusion-weighted imaging," *European Journal of Radiology*, vol. 45, pp. 169–184, 2003. doi:https://doi.org/10.1016/S0720-048X(02)00303-0
- [48] G. H. Glover, "Overview of Functional Magnetic Resonance Imaging," *Neurosurgery Clinics of North America*, vol. 22, pp. 133–139, 2011, functional Imaging. doi:https://doi.org/10.1016/j.nec.2010.11.001
- [49] R. M. Mann, N. Cho, and L. Moy, "Breast MRI: State of the Art," *Radiology*, vol. 292, pp. 520–536, 2019, pMID: 31361209. doi:10.1148/radiol.2019182947
- [50] N. Iyad, M. S. Ahmad, S. G. Alkhatib, and M. Hjoui, "Gadolinium contrast agents- challenges and opportunities of a multidisciplinary approach: Literature review," *European Journal of Radiology Open*, vol. 11, p. 100503, 2023. doi:https://doi.org/10.1016/j.ejro.2023.100503
- [51] J. Jonsson, T. Nyholm, and K. Söderkvist, "The rationale for mr-only treatment planning for external radiotherapy," *Clinical and Translational Radiation Oncology*, vol. 18, pp. 60–65, 2019. doi:https://doi.org/10.1016/j.ctro.2019.03.005
- [52] P. Decazes, P. Hinault, O. Veresezan, S. Thureau, P. Gouel, and P. Vera, "Trimodality PET/CT/MRI and Radiotherapy: A Mini-Review," *Frontiers in Oncology*, vol. Volume 10 - 2020, 2021. doi:10.3389/fonc.2020.614008
- [53] E. M. Huijben, M. L. Terpstra, A. J. Galapon, et al., "Generating synthetic computed tomography for radiotherapy: Synthrad2023 challenge report," *Medical Image Analysis*, vol. 97, p. 103276, 2024. doi:https://doi.org/10.1016/j.media.2024.103276
- [54] X. Li, R. Bellotti, G. Meier, et al., "Uncertainty-aware MR-based CT synthesis for robust proton therapy planning of brain tumour," *Radiotherapy and Oncology*, vol. 191, p. 110056, 2024. doi:https://doi.org/10.1016/j.radonc.2023.110056
- [55] P. Zaffino, C. B. Raggio, A. Thummerer, et al., "Toward Closing the Loop in Image-to-Image Conversion in Radiotherapy: A Quality Control Tool to Predict Synthetic Computed Tomography Hounsfield Unit Accuracy," *Journal of Imaging*, vol. 10, 2024. doi:10.3390/jimaging10120316
- [56] C. Bojchko, P. Hua, W. Sumner, K. Guram, T. Atwood, and A. Sharabi, "Adaptive replanning using cone beam CT for deformation of original CT simulation," *Journal of Medical Radiation Sciences*, vol. 69, pp. 267–272, 2022. doi:https://doi.org/10.1002/jmrs.549
- [57] A. Altalib, S. McGregor, C. Li, and A. Perelli, "Synthetic CT Image Generation From CBCT: A Systematic Review," *IEEE Transactions on Radiation and Plasma Medical Sciences*, pp. 1–1, 2025. doi:10.1109/TRPMS.2025.3533749
- [58] J. Wang, S. Wang, and Y. Zhang, "Deep learning on medical image analysis," *CAAI Transactions on Intelligence Technology*, vol. 10, pp. 1–35, 2025. doi:https://doi.org/10.1049/cit2.12356
- [59] I. D. Mienye, T. G. Swart, G. Obaido, M. Jordan, and P. Ilono, "Deep Convolutional Neural Networks in Medical Image Analysis: A Review," *Information*, vol. 16, 2025. doi:10.3390/info16030195

- [60] P. K. Mall, P. K. Singh, S. Srivastav, et al., “A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities,” *Healthcare Analytics*, vol. 4, p. 100216, 2023. doi:<https://doi.org/10.1016/j.health.2023.100216>
- [61] B. Sistaninejhad, H. Rasi, and P. Nayeri, “A Review Paper about Deep Learning for Medical Image Analysis,” *Computational and Mathematical Methods in Medicine*, vol. 2023, p. 7091301, 2023. doi:<https://doi.org/10.1155/2023/7091301>
- [62] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 818–833.
- [63] A. A. Murat and M. S. Kiran, “A comprehensive review on yolo versions for object detection,” *Engineering Science and Technology, an International Journal*, vol. 70, p. 102161, 2025. doi:<https://doi.org/10.1016/j.jestch.2025.102161>
- [64] Y. Xu, R. Quan, W. Xu, Y. Huang, X. Chen, and F. Liu, “Advances in Medical Image Segmentation: A Comprehensive Review of Traditional, Deep Learning and Hybrid Approaches,” *Bioengineering*, vol. 11, 2024. doi:[10.3390/bioengineering11101034](https://doi.org/10.3390/bioengineering11101034)
- [65] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, LNCS, vol. 9351. Springer, 2015, pp. 234–241. doi:https://doi.org/10.1007/978-3-319-24574-4_28
- [66] F. Neha, D. Bhati, D. K. Shukla, S. M. Dalvi, N. Mantzou, and S. Shubbar, “An analytics-driven review of U-Net for medical image segmentation,” *Healthcare Analytics*, vol. 8, p. 100416, 2025. doi:<https://doi.org/10.1016/j.health.2025.100416>
- [67] R. Yousef, S. Khan, G. Gupta, et al., “U-Net-Based Models towards Optimal MR Brain Image Segmentation,” *Diagnostics*, vol. 13, 2023. doi:[10.3390/diagnostics13091624](https://doi.org/10.3390/diagnostics13091624)
- [68] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, “UNet++: A Nested U-Net Architecture for Medical Image Segmentation,” in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham: Springer International Publishing, 2018, pp. 3–11.
- [69] S. Mullan, L. Zhang, H. Zhang, and M. Sonka, “Chapter 18 - Deep learning medical image segmentation,” in *Medical Image Analysis*, The MICCAI Society book Series, A. F. Frangi, J. L. Prince, and M. Sonka, Eds. Academic Press, 2024, pp. 475–500. doi:<https://doi.org/10.1016/B978-0-12-813657-7.00042-X>
- [70] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation,” *Nature Methods*, vol. 18, p. 203–211, 2020. doi:[10.1038/s41592-020-01008-z](https://doi.org/10.1038/s41592-020-01008-z)
- [71] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251. doi:[10.1109/ICCV.2017.244](https://doi.org/10.1109/ICCV.2017.244)
- [72] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., “Generative Adversarial Networks,” 2014. <https://arxiv.org/abs/1406.2661>
- [73] Y. Heng, M. Yinghua, F. G. Khan, et al., “Survey: application and analysis of generative adversarial networks in medical images,” *Artificial Intelligence Review*, vol. 58, 2024. doi:[10.1007/s10462-024-10992-z](https://doi.org/10.1007/s10462-024-10992-z)
- [74] M. Ali, M. Ali, M. Hussain, and D. Koundal, “Generative Adversarial Networks (GANs) for Medical Image Processing: Recent Advancements,” *Archives of Computational Methods in Engineering*, vol. 32, p. 1185–1198, 2024. doi:[10.1007/s11831-024-10174-8](https://doi.org/10.1007/s11831-024-10174-8)

- [75] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5967–5976. doi:10.1109/CVPR.2017.632
- [76] 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, p. 2069–2078, 2021. doi:10.1007/s11548-021-02433-x
- [77] M. Salehjahromi, T. V. Karpinets, S. J. Sujit, et al., "Synthetic PET from CT improves diagnosis and prognosis for lung cancer: Proof of concept," *Cell Reports Medicine*, vol. 5, p. 101463, 2024. doi:https://doi.org/10.1016/j.xcrm.2024.101463
- [78] R. N. Pai, S. M. Shuhood, A. Shenoy, and P. G., "Image restoration using deep learning techniques: A comprehensive review," *ICT Express*, 2026. doi:https://doi.org/10.1016/j.ict.2026.03.002
- [79] S. Aburass, O. Dorgham, J. Al Shaqsi, M. Abu Rumman, and O. Al-Kadi, "Vision Transformers in Medical Imaging: a Comprehensive Review of Advancements and Applications Across Multiple Diseases," *Journal of Imaging Informatics in Medicine*, vol. 38, p. 3928–3971, 2025. doi:10.1007/s10278-025-01481-y
- [80] W. Wang, J. Xia, G. Luo, et al., "Diffusion model for medical image denoising, reconstruction and translation," *Computerized Medical Imaging and Graphics*, vol. 124, p. 102593, 2025. doi:https://doi.org/10.1016/j.compmedimag.2025.102593
- [81] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Scientific Reports*, vol. 14, 2024. doi:10.1038/s41598-024-56706-x
- [82] P. Civicioglu and E. Besdok, "Image discrimination robustness of classical image quality metrics: an analysis on MAE, MSE, ERGAS, LMSE, SSIM, SAM, UIQI, PSNR, NIQE, PIQE, SNR, VIF, FSIM, GMSD, and NCC," *Quality amp; Quantity*, 2025. doi:10.1007/s11135-025-02404-3
- [83] L. Maier-Hein, A. Reinke, P. Godau, et al., "Metrics reloaded: recommendations for image analysis validation," *Nature Methods*, vol. 21, p. 195–212, 2024. doi:10.1038/s41592-023-02151-z
- [84] A. Kebaili, J. Lapuyade-Lahorgue, and S. Ruan, "Deep Learning Approaches for Data Augmentation in Medical Imaging: A Review," *Journal of Imaging*, vol. 9, 2023. doi:10.3390/jimaging9040081
- [85] F. Garcea, A. Serra, F. Lamberti, and L. Morra, "Data augmentation for medical imaging: A systematic literature review," *Computers in Biology and Medicine*, vol. 152, p. 106391, 2023. doi:https://doi.org/10.1016/j.compbiomed.2022.106391
- [86] T. Islam, M. S. Hafiz, J. R. Jim, M. M. Kabir, and M. Mridha, "A systematic review of deep learning data augmentation in medical imaging: Recent advances and future research directions," *Healthcare Analytics*, vol. 5, p. 100340, 2024. doi:https://doi.org/10.1016/j.health.2024.100340
- [87] Y. Xu, Y. Li, F. Wang, Y. Zhang, and D. Huang, "Addressing the current challenges in the clinical application of AI-based radiomics for cancer imaging," *Frontiers in Medicine*, vol. Volume 12 - 2025, 2025. doi:10.3389/fmed.2025.1674397
- [88] B. Texier, C. Hémon, P. Lekieffre, et al., "Computed tomography synthesis from magnetic resonance imaging using cycle generative adversarial networks with multicenter learning," *Physics and Imaging in Radiation Oncology*, vol. 28, p. 100511, 2023. doi:https://doi.org/10.1016/j.phro.2023.100511
- [89] M. U. Suleman, M. Mursaleen, U. Khalil, et al., "Assessing the generalizability of artificial intelligence in radiology: a systematic review of performance across different clinical set-

- tings,” *Annals of Medicine and Surgery*, vol. 87, p. 8803–8811, 2025. doi:10.1097/ms9.0000000000004166
- [90] P. Kairouz, H. B. McMahan, B. Avent, et al., “Advances and Open Problems in Federated Learning,” *Foundations and Trends in Machine Learning*, vol. 14, p. 1–210, 2021. doi:10.1561/22000000083
- [91] S. S. Sandhu, H. T. Gorji, P. Tavakolian, K. Tavakolian, and A. Akhbardeh, “Medical imaging applications of federated learning,” *Diagnostics*, vol. 13, p. 3140, 2023. doi:https://doi.org/10.3390/diagnostics13193140
- [92] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-Efficient Learning of Deep Networks from Decentralized Data,” 2023. https://arxiv.org/abs/1602.05629
- [93] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, “Federated Optimization in Heterogeneous Networks,” 2020. https://arxiv.org/abs/1812.06127
- [94] H. Morsel, K. Samrouth, N. Bakir, and P. Vadim, “GAN for MRI Reconstruction in Federated Learning,” in *2024 International Conference on Smart Systems and Power Management (IC2SPM)*, 2024, pp. 29–33. doi:10.1109/IC2SPM62723.2024.10841358
- [95] S. Ahmed, F. Jinchao, M. A. Manan, M. Yaqub, M. U. Ali, and A. Raheem, “FedGraphMRI-net: A federated graph neural network framework for robust MRI reconstruction across non-IID data,” *Biomedical Signal Processing and Control*, vol. 102, p. 107360, 2025. doi:https://doi.org/10.1016/j.bspc.2024.107360
- [96] M. Pennisi, F. Proietto Salanitri, G. Bellitto, et al., “FedER: Federated Learning through Experience Replay and privacy-preserving data synthesis,” *Computer Vision and Image Understanding*, vol. 238, p. 103882, 2024. doi:https://doi.org/10.1016/j.cviu.2023.103882
- [97] J. Wang, G. Xie, Y. Huang, et al., “FedMed-GAN: Federated domain translation on unsupervised cross-modality brain image synthesis,” *Neurocomputing*, vol. 546, p. 126282, 2023. doi:https://doi.org/10.1016/j.neucom.2023.126282
- [98] J. Wang, G. Xie, Y. Huang, Y. Zheng, Y. Jin, and F. Zheng, “FedMed-ATL: Misaligned Unpaired Cross-Modality Neuroimage Synthesis via Affine Transform Loss,” in *Proceedings of the 30th ACM International Conference on Multimedia*, MM ’22. New York, NY, USA: Association for Computing Machinery, 2022, p. 1522–1531. doi:10.1145/3503161.3547762
- [99] C. B. Raggio, M. K. Zabaleta, N. Skupien, et al., “FedSynthCT-Brain: A federated learning framework for multi-institutional brain MRI-to-CT synthesis,” *Computers in Biology and Medicine*, vol. 192, p. 110160, 2025. doi:https://doi.org/10.1016/j.compbiomed.2025.110160
- [100] C. B. Raggio, P. Zaffino, and M. F. Spadea, “A privacy-preserving federated learning framework for generalizable cbct to synthetic ct translation in head and neck,” *Frontiers in Digital Health*, vol. Volume 8 - 2026, 2026. doi:10.3389/fdgth.2026.1812254
- [101] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, “Secure, privacy-preserving and federated machine learning in medical imaging,” *Nature Machine Intelligence*, vol. 2, pp. 305–311, 2020.
- [102] A. Thummerer, P. Zaffino, M. F. Spadea, A. Knopf, and M. Maspero, “Artificial intelligence to generate synthetic CT for adaptive particle therapy,” in *Imaging in Particle Therapy*, 2053–2563. IOP Publishing, 2024, pp. 8–1 to 8–16. doi:10.1088/978-0-7503-5117-1ch8
- [103] J. Grigo, J. Szkitsak, D. Höfler, R. Fietkau, F. Putz, and C. Bert, ““sCT-Feasibility” - a feasibility study for deep learning-based MRI-only brain radiotherapy,” *Radiation Oncology*, vol. 19, p. 33, 2024. doi:10.1186/s13014-024-02428-3

- [104] S. Tahri, B. Texier, J.-C. Nunes, et al., “A deep learning model to generate synthetic CT for prostate MR-only radiotherapy dose planning: a multicenter study,” *Frontiers in Oncology*, vol. 13, 2023. doi:10.3389/fonc.2023.1279750
- [105] S. Han, B. Texier, C. Hémon, et al., “Balancing data consistency and diversity: Preprocessing and online data augmentation for multi-center deep learning-based mr-to-ct synthesis,” *Pattern Recognition Letters*, 2025. doi:10.1016/j.patrec.2025.01.002
- [106] L. Han, T. Tan, Y. Huang, et al., “All-in-one medical image-to-image translation,” *Cell Reports Methods*, vol. 5, p. 101138, 2025. doi:https://doi.org/10.1016/j.crmeth.2025.101138
- [107] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, et al., “3D Slicer as an image computing platform for the Quantitative Imaging Network,” *Magnetic Resonance Imaging*, vol. 30, pp. 1323–1341, 2012, quantitative Imaging in Cancer. doi:https://doi.org/10.1016/j.mri.2012.05.001
- [108] L. Boldrini, A. D’Aviero, F. De Felice, et al., “Artificial intelligence applied to image-guided radiation therapy (IGRT): a systematic review by the Young Group of the Italian Association of Radiotherapy and Clinical Oncology (yAIRO),” *La Radiologia Medica*, vol. 129, p. 133–151, 2023. doi:10.1007/s11547-023-01708-4
- [109] M. Rabe, C. Kurz, A. Thummerer, and G. Landry, “Artificial intelligence for treatment delivery: image-guided radiotherapy,” *Strahlentherapie und Onkologie*, vol. 201, p. 283 – 297, 2025. doi:10.1007/s00066-024-02277-9
- [110] E. Mastella, F. Calderoni, L. Manco, et al., “A systematic review of the role of artificial intelligence in automating computed tomography-based adaptive radiotherapy for head and neck cancer,” *Physics and Imaging in Radiation Oncology*, vol. 33, p. 100731, 2025. doi:10.1016/j.phro.2025.100731
- [111] V. Bolón-Canedo, L. Morán-Fernández, B. Cancela, and A. Alonso-Betanzos, “A review of green artificial intelligence: Towards a more sustainable future,” *Neurocomputing*, p. 128096, 2024. https://doi.org/10.1016/j.neucom.2024.128096
- [112] S. Luccioni, Y. Jernite, and E. Strubell, “Power Hungry Processing: Watts Driving the Cost of AI Deployment?” in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’24. New York, NY, USA: Association for Computing Machinery, 2024, p. 85–99. doi:10.1145/3630106.3658542
- [113] G. Varoquaux, S. Luccioni, and M. Whittaker, “Hype, Sustainability, and the Price of the Bigger-is-Better Paradigm in AI,” in *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’25. New York, NY, USA: Association for Computing Machinery, 2025, p. 61–75. doi:10.1145/3715275.3732006
- [114] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, “Green AI,” *Commun. ACM*, vol. 63, p. 54–63, 2020. doi:10.1145/3381831
- [115] H. Liu, Y. Wang, W. Fan, et al., “Trustworthy AI: A computational perspective,” *ACM Transactions on Intelligent Systems and Technology*, vol. 14, pp. 1–59, 2022. doi:10.1145/3546872
- [116] A. Jobin, M. Ienca, and E. Vayena, “The global landscape of AI ethics guidelines,” *Nature machine intelligence*, vol. 1, pp. 389–399, 2019. https://doi.org/10.1038/s42256-019-0088-2
- [117] R. Raman, D. Pattnaik, H. H. Lathabai, C. Kumar, K. Govindan, and P. Nedungadi, “Green and sustainable AI research: an integrated thematic and topic modeling analysis,” *Journal of Big Data*, vol. 11, p. 55, 2024. https://doi.org/10.1186/s40537-024-00920-x
- [118] H. Siala and Y. Wang, “Shifting artificial intelligence to be responsible in healthcare: A systematic review,” *Social Science & Medicine*, vol. 296, p. 114782, 2022. doi:https://doi.org/10.1016/j.socscimed.2022.114782

- [119] C. Chen, T. Liao, X. Deng, Z. Wu, S. Huang, and Z. Zheng, “Advances in Robust Federated Learning: A Survey With Heterogeneity Considerations,” *IEEE Transactions on Big Data*, vol. 11, pp. 1548–1567, 2025. doi:10.1109/TBDDATA.2025.3527202
- [120] A. Rauniyar, D. H. Hagos, D. Jha, et al., “Federated learning for medical applications: A taxonomy, current trends, challenges, and future research directions,” *IEEE Internet of Things Journal*, 2023. doi:10.1109/JIOT.2023.3329061
- [121] C. B. Raggio, L. Migliorelli, N. Skupien, et al., “Energy-efficient federated learning via adaptive encoder freezing for MRI-to-CT conversion: A Green AI-guided research,” *Sustainable Computing: Informatics and Systems*, vol. 51, p. 101433, 2026. doi:https://doi.org/10.1016/j.suscom.2026.101433
- [122] K. Hu, S. Gong, Q. Zhang, C. Seng, M. Xia, and S. Jiang, “An overview of implementing security and privacy in federated learning,” *Artificial Intelligence Review*, vol. 57, 2024. doi:10.1007/s10462-024-10846-8
- [123] Y. Feng, Y. Guo, Y. Hou, et al., “A survey of security threats in federated learning,” *Complex & Intelligent Systems*, vol. 11, 2025. doi:10.1007/s40747-024-01664-0
- [124] L. Zhu, Z. Liu, and S. Han, *Deep leakage from gradients*. Red Hook, NY, USA: Curran Associates Inc., 2019.
- [125] G. Zhu, D. Li, H. Gu, Y. Yao, L. Fan, and Y. Han, “FedMIA: An Effective Membership Inference Attack Exploiting “All for One” Principle in Federated Learning,” in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, 2025, pp. 20 643–20 653. doi:10.1109/CVPR52734.2025.01922
- [126] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, “How To Backdoor Federated Learning,” *CoRR*, vol. abs/1807.00459, 2018. <http://arxiv.org/abs/1807.00459>
- [127] R. Jin and X. Li, “Backdoor attack and defense in federated generative adversarial network-based medical image synthesis,” *Medical Image Analysis*, vol. 90, p. 102965, 2023. doi:https://doi.org/10.1016/j.media.2023.102965
- [128] K. N. Kumar, C. K. Mohan, L. R. Cenkeramaddi, and N. Awasthi, “Minimal data poisoning attack in federated learning for medical image classification: An attacker perspective,” *Artificial Intelligence in Medicine*, vol. 159, p. 103024, 2025. doi:https://doi.org/10.1016/j.artmed.2024.103024
- [129] C. B. Raggio, L. Bucher, O. Blanck, F. Cicone, P. Zaffino, and M. F. Spadea, “Security analysis of a federated learning framework for medical image-to-image translation,” *Journal of Medical Systems*, vol. 50, 2026. doi:10.1007/s10916-026-02436-8
- [130] K. A. Bonawitz, V. Ivanov, B. Kreuter, et al., “Practical Secure Aggregation for Federated Learning on User-Held Data,” *CoRR*, vol. abs/1611.04482, 2016. <http://arxiv.org/abs/1611.04482>
- [131] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, “Byzantine-robust distributed learning: Towards optimal statistical rates,” in *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research, vol. 80. PMLR, 2018, pp. 5650–5659.
- [132] T. A. Soomro, L. Zheng, A. J. Afifi, et al., “Image segmentation for MR brain tumor detection using machine learning: a review,” *IEEE Reviews in Biomedical Engineering*, vol. 16, pp. 70–90, 2022. doi:https://doi.org/10.1109/RBME.2022.3185292
- [133] S. Karim, G. Tong, Y. Yu, et al., “Developments in Brain Tumor Segmentation Using MRI: Deep Learning Insights and Future Perspectives,” *IEEE Access*, vol. 12, pp. 26 875–26 896, 2024. doi:https://doi.org/10.1109/ACCESS.2024.3365048

- [134] S. Alphonse, F. Mathew, K. Dhanush, and V. Dinesh, "Federated learning with integrated attention multiscale model for brain tumor segmentation," *Scientific Reports*, vol. 15, p. 11889, 2025. doi:<https://doi.org/10.1038/s41598-025-96416-6>
- [135] W. Zhang, W. Jin, S. Rho, F. Jiang, and C.-f. Yang, "A federated learning framework for brain tumor segmentation without sharing patient data," *International Journal of Imaging Systems and Technology*, vol. 34, p. e23147, 2024. doi:<https://doi.org/10.1002/ima.23147>
- [136] K. S. Adewole, E. Alozie, H. Olagunju, et al., "A systematic review and meta-data analysis of clinical data repositories in Africa and beyond: recent development, challenges, and future directions," *Discover Data*, vol. 2, p. 8, 2024. doi:<https://doi.org/10.1007/s44248-024-00012-4>
- [137] C. Shiranthika, P. Saeedi, and I. V. Bajić, "Decentralized Learning in Healthcare: A Review of Emerging Techniques," *IEEE Access*, vol. 11, pp. 54 188–54 209, 2023. doi:10.1109/ACCESS.2023.3281832
- [138] C. Ren, H. Yu, H. Peng, et al., "Advances and Open Challenges in Federated Foundation Models," *IEEE Communications Surveys & Tutorials*, vol. 28, pp. 2087–2126, 2026. doi:10.1109/COMST.2025.3552524
- [139] A. N. Sahu, A. Dutta, A. M. Abdelmoniem, T. Banerjee, M. Canini, and P. Kalnis, "Rethinking gradient sparsification as total error minimization," in *NeurIPS 2021 - Advances in Neural Information Processing Systems*, 2021. <https://arxiv.org/abs/2108.00951>
- [140] D. Alistarh, D. Grubic, J. Z. Li, R. Tomioka, and M. Vojnovic, "QSGD: communication-efficient SGD via gradient quantization and encoding," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 1707–1718.
- [141] W. Wen, C. Xu, F. Yan, et al., "TernGrad: ternary gradients to reduce communication in distributed deep learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 1508–1518.
- [142] A. Yang, J. Liu, B. Yang, Z. Mo, and K. Li, "Balancing communication overhead and accuracy in compression integration: a survey," *The Journal of Supercomputing*, vol. 81, p. 964, 2025. doi:<https://doi.org/10.1007/s11227-025-07451-z>
- [143] A. N. Onaizah, Y. Xia, A. J. Obaid, and K. Hussain, "Deep learning based brain tumour architecture for weight sharing optimization in federated learning," *Expert Systems*, vol. 42, p. e13643, 2025. doi:<https://doi.org/10.1111/exsy.13643>
- [144] S. Bakas, A. Linardos, S. Pati, et al., "The Federated Tumor Segmentation (FeTS) Challenge 2024," 2024. doi:<https://doi.org/10.5281/ZENODO.10990500>
- [145] A. Raza, C. B. Raggio, A. Guzzo, M. F. Spadea, and G. Fortino, "Towards robust neurocomputing model in efficient federated brain tumour segmentation with sparsification and weights clustering," *Neurocomputing*, p. 133142, 2026. doi:<https://doi.org/10.1016/j.neucom.2026.133142>
- [146] C. B. Raggio, P. Zaffino, and M. F. Spadea, "ImageAugmenter: A user-friendly 3D Slicer tool for medical image augmentation," *SoftwareX*, vol. 28, p. 101923, 2024. doi:<https://doi.org/10.1016/j.softx.2024.101923>
- [147] A. Paszke, S. Gross, F. Massa, et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," 2019. <https://arxiv.org/abs/1912.01703>
- [148] M. J. Cardoso, W. Li, R. Brown, et al., "MONAI: An open-source framework for deep learning in healthcare," 2022. <https://arxiv.org/abs/2211.02701>

List of Publications and Supervised Theses

Journal Articles

- **Raggio C. B.**, Migliorelli L., Skupien N., Krohmer Zabaleta M., Blanck O., Cicone F., Cascini G. L., Zaffino P., Spadea M. F., *Energy-Efficient Federated Learning via Adaptive Encoder Freezing for MRI-to-CT Conversion: A Green AI-Guided Research*, Sustainable Computing: Informatics and Systems 2026; 51, <https://doi.org/10.1016/j.suscom.2026.101433>
- **Raggio C. B.***, Bucher L.* , Blanck O., Cicone F., Zaffino P., Spadea M. F., *Security Analysis of a Federated Learning Framework for Medical Image-to-Image Translation*; Journal of Medical Systems 2026; 50, 108, <https://doi.org/10.1007/s10916-026-02436-8>
(*) *Raggio C.B. and Bucher L. have contributed equally to this work*
- **Raggio C. B.**, Zaffino P., Spadea M. F., *A Privacy-Preserving Federated Learning Framework for Generalizable CBCT to Synthetic CT Translation in Head and Neck*, Frontiers in Digital Health 2026, 8:1812254, <https://doi.org/10.3389/fdgth.2026.1812254>
- Raza A.* & **Raggio C.B.***, Guzzo A., Spadea M. F., Fortino G., *Towards Robust Neurocomputing Model in Efficient Federated Brain Tumour Segmentation with Sparsification and Weights Clustering*, Neurocomputing 2026; 133142, 0925-2312, <https://doi.org/10.1016/j.neucom.2026.133142>
(*) *Raza A. and Raggio C.B. have contributed equally to this work*
- **Raggio C. B.**, Krohmer Zabaleta M., Skupien N., Blanck O., Cicone F., Cascini G. L., Zaffino P., Migliorelli L., Spadea M. F., *FedSynthCT-Brain: A federated learning framework for multi-institutional brain MRI-to-CT synthesis*, Computers in Biology and Medicine 2025; 110160, 0010-4825, <https://doi.org/10.1016/j.combiomed>.

2025.110160

- De Rose E. * & **Raggio C.B.** *, Rasheed A.R., Bruno P., Zaffino P., De Rosa S. & Calimeri F. & Spadea M.F., *A multi-model deep learning approach for the identification of coronary artery calcifications within 2D coronary angiography images* International Journal of Computer Assisted Radiology and Surgery 2025; 20, 1273–1281, <https://doi.org/10.1007/s11548-025-03382-5>
(*) De Rose E. and Raggio C.B. have contributed equally to this work
- **Raggio C. B.**, Zaffino P., Spadea M. F., *ImageAugmenter: A user-friendly 3D Slicer tool for medical image augmentation*, SoftwareX 2024; Volume 28, 2352-7110, <https://doi.org/10.1016/j.softx.2024.101923>
- Zaffino P., **Raggio C. B.**, Thummerer, A., Marmitt, G. G., Langendijk, J. A., Procopio, A., Cosentino, C., Seco, J., Knopf, A. C., Both, S., Spadea, M. F., *Toward Closing the Loop in Image-to-Image Conversion in Radiotherapy: A Quality Control Tool to Predict Synthetic Computed Tomography Hounsfield Unit Accuracy*, Journal of Imaging 2024; 10(12), 316, <https://doi.org/10.3390/jimaging10120316>

Refereed Conference Abstracts

- Romeo L., **Raggio C. B.**, Destito M., Cosentino C., Both S., Spadea M. F., Zaffino P., *How to fix a synthetic CT: a Deep Learning model to predict the voxel-wise conversion error*, European Society for Radiotherapy and Oncology (ESTRO) 2026; Stockholm, Sweden
- Krnjaca D., Heese H., **Raggio C. B.**, Spadea M. F., Nickisch H., *Connectivity-Preserving Losses for 3D Coronary Artery Segmentation*, IEEE 23rd International Symposium on Biomedical Imaging (ISBI) 2026; London, UK, https://www.researchgate.net/publication/403171912_Connectivity-Preserving_Losses_for_3D_Coronary_Artery_Segmentation

Invited Talks

- **Raggio C. B.**, *FedSynthCT: Enabling Collaborative Medical Image Translation Across Institutions*, International Symposium on Biomedical Research and Innovation for Digitalization and Global Exchange (BRIDGE Symposium), Nov 3-4 2025; Kyoto, Japan; <https://doi.org/10.1007/978-98-1-10-160-0>

[//sites.google.com/view/bridge-symposium-2025/home](https://sites.google.com/view/bridge-symposium-2025/home) (Last accessed: October 27, 2026)

Software and Libraries

- Riggio D., Ghodousipour K., Lichtlein L., **Raggio C. B.**, Spadea M. F., *SlicerHexaMesher*, 2026;
<https://github.com/DomenicoRiggio/SlicerHexaMesher> (Last accessed: October 27, 2026)
- **Raggio C. B.**, Zaffino P., Spadea M.F., *SlicerModalityConverter: An Open-Source 3D Slicer Extension for Medical Image-to-Image Translation*, 2025;
<https://github.com/ciroraggio/SlicerModalityConverter> (Last accessed: October 27, 2026)
- Giuliani A., Mitura J., **Raggio C. B.**, Namin M. J., Elkhechi N. G., Destito M., Cosentino C., Spadea M. F., Zaffino P., *Radiomics.jl: a Library for High-Performance Radiomic Features Extraction from Medical Images*, 2025,
<https://github.com/pzaffino/Radiomics.jl> (Last accessed: October 27, 2026)
- Romeo L., Veraldi R., **Raggio C. B.**, Barone S., Destito M., Giudice A., Cosentino C., Spadea M. F., Zaffino P., *ClassAnnotation: a 3D Slicer extension for patient-level annotation of medical imaging datasets*, 2025;
<https://github.com/lorenaromeo/SlicerClassAnnotation> (Last accessed: October 27, 2026)
- **Raggio C. B.**, Zaffino P., Spadea M.F., *ImageAugmenter: A user-friendly 3D Slicer tool for medical image augmentation*, 2024; <https://github.com/ciroraggio/SlicerImageAugmenter> (Last accessed: October 27, 2026)

Supervised Student Theses

- Bidner P., *Federated Training of EVA-I2I: Multi-Domain and Multi-Region Translation in Decentralised Scenarios*, Master Thesis, Institute of Biomedical Engineering, Karlsruhe Institute of Technology (KIT), 2025/2026

- Bucher L., *Validating the Privacy of a Decentralised Medical Image-to-Image Translation Framework through Security Attacks*, Master Thesis, Institute of Biomedical Engineering, Karlsruhe Institute of Technology (KIT) & Heidelberg University, 2025/2026
- Krnjaca D., *Impact of Connectivity-Preserving Loss Functions on the Segmentation of Thin Tubular Structures: Application to Coronary Arteries From CT Angiography Data*, Master Thesis, Institute of Biomedical Engineering, Karlsruhe Institute of Technology (KIT) & Philips Hamburg, 2026
- Schneider J., *Feasibility of CBCT-to-CT Synthesis in Adaptive Radiotherapy of Pancreatic Cancer*, Master Thesis, Karlsruhe Institute of Technology (KIT) & Heidelberg University & Kyoto University Hospital, 2025/2026
- Lenhard A., *Validating the Robustness of a Decentralized MRI to sCT Translation Framework on Different Anatomical Sites*, Bachelor Thesis, Institute of Biomedical Engineering & Department of Mathematics, Karlsruhe Institute of Technology (KIT), 2025
- Krohmer Zabaleta M., *Federated image translation in radiotherapy: a real-world feasibility study*, Bachelor Thesis, Institute of Biomedical Engineering, Karlsruhe Institute of Technology (KIT), 2024
- Skupien N., *Federated model aggregation methods for image translation in radiotherapy*, Master Thesis, Institute of Biomedical Engineering, Karlsruhe Institute of Technology (KIT), 2024
- Hong C., *Deep Learning Based Detection, Segmentation and Classification of the Third Molar from CBCT Images*, Master Thesis, Institute of Biomedical Engineering, Karlsruhe Institute of Technology (KIT), 2023/2024