KU LEUVEN

DOCTORAL SCHOOL FOR THE HUMANITIES AND SOCIAL SCIENCES
Faculty of Economics and Business

Federated Learning in Healthcare

Towards privacy enhancing, AI-driven decision support systems.

Vincent Scheltjens

Supervisors:
Prof. dr. ir. B. De Moor
Prof. dr. ir. W. Verbeke

Chair:
Prof. dr. Y. Wautelet

Dissertation presented in partial fulfillment
of the requirements for the degree of
Doctor in Business Economics (PhD)

March 2026

FEDERATED LEARNING IN HEALTHCARE

TOWARDS PRIVACY ENHANCING, AI-DRIVEN DECISION SUPPORT SYSTEMS.

Vincent SCHELTJENS

Supervisors:
Prof. dr. ir. B. De Moor
Prof. dr. ir. W. Verbeke

Members of the
Examination Committee:
Prof. dr. Y. Wautelet, chair
Prof. dr. ir. Y. Moreau
Prof. dr. J. De Smedt
Em. Prof. dr. ir. M.-F. Moens
Prof. dr. Jeroen Berrevoets
(University of Cambridge)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Business Economics (PhD)

March 2026

© 2026 Vincent Scheltjens
Uitgegeven in eigen beheer, Vincent Scheltjens, Leuven (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotokopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm, electronic or any other means without written permission from the publisher.

Preface

About four years ago, I embarked on a PhD journey that is now coming to an end. Looking back, these years have been marked by substantial personal and professional development. This period offered me the opportunity to explore new research directions, refine my scientific thinking, and work collaboratively within an environment that consistently challenged me to grow.

Scientific work is often presented as a sequence of clear results and well-defined conclusions. In practice, the process is far less linear. Ideas evolve gradually through experimentation, discussion, and reflection. Many attempts lead to new questions rather than immediate answers, and progress often emerges from moments of uncertainty. Looking back, the work presented in this dissertation reflects not only the outcome of research but also the many conversations, collaborations, and insights that shaped it along the way.

Although the PhD journey is, by nature, an individual one, the work presented here is undoubtedly the result of the guidance, support, and contributions of many people who stood by my side throughout these years.

First and foremost, I would like to express my sincere appreciation to Prof. Bart De Moor for granting me the opportunity to pursue a PhD under his supervision, an opportunity I have never taken for granted. His vision, the scientific environment he has cultivated, and the trust he placed in me provided the foundation upon which this work was built. The freedom he gave me to explore ideas independently, combined with his ability to frame research questions within a broader scientific context, has been invaluable.

I am equally grateful to Prof. Wouter Verbeke for his guidance throughout this process. Our interactions often served as the first validation for ideas that were still taking shape, and those discussions played an important role in refining both the direction and the quality of this work. The complementary guidance of both supervisors created an environment that allowed independence while ensuring direction when it was most needed.

I would also like to express my gratitude to the members of the examination committee, who dedicated their time and expertise to reviewing this thesis. Their careful consideration and valuable feedback contributed to improving both the clarity and the academic rigor of the final manuscript.

I wish to extend my gratitude to my colleagues at ESAT and within the broader research environment. Working within these groups provided a stimulating atmosphere in which ideas could be exchanged openly. The shared offices, spontaneous discussions, and collaborative spirit created an environment that made the daily process of research both engaging and rewarding.

In particular, I would like to acknowledge Naomi, with whom I shared the early stages of my research. The many discussions we had played an important role in shaping my understanding of what research truly involves: persistence, curiosity, and the willingness to continuously refine ideas. As she once mentioned in her own dissertation, "the best collaborator I have had", I would hereby like to confirm, that feeling is entirely mutual.

Working within the research group provided a stimulating and collaborative environment throughout the PhD. It has been a pleasure sharing offices and day-to-day research life with many colleagues, including Christof, Hans, Sarthak, Sem, Giulia, Fernando, Katrien, Mauricio, Willem, Lola, Bram, Alexandre, Axel, and Zahra.

Among this group, several colleagues accompanied similar stages of the doctoral journey. Sibren and Lukas started their PhDs around the same time, each working on different topics but sharing many of the same milestones along the way. Wishing both of you the best in your upcoming endeavours. Thank you, Thomas, for consistently attending my intermediate presentations, a gesture of support that never went unnoticed.

Beyond the academic environment, several people have shaped the broader context in which this work developed.

Jeroen has been part of that journey long before the PhD began. Our early entrepreneurial initiatives taught me valuable lessons about persistence, creativity, and building things from scratch — lessons that proved surprisingly relevant to the research process itself. Jens represents the continuation of that spirit in a different form. Our ongoing collaborations demonstrate that academic research does not have to remain confined to publications, but can also translate into practical applications and real-world impact.

My friends, especially David, Lorenz, and Gianni, and by extension all close friends, deserve a special mention. Over the years we have developed a strong and lasting bond. Having friends who show genuine interest in what you are

doing, makes a difference. I am deeply grateful for every single minute I get to spend by your sides.

David, becoming the godfather of your son Raphaël during this period is something I consider a true honor. Some milestones in life carry a meaning that goes well beyond academic achievements.

Aan mijn ouders en zus, Hilde, Patrick en Emma. Jullie hebben mij altijd gesteund, ook wanneer het waarschijnlijk niet altijd duidelijk was waar ik precies mee bezig was. Een doctoraat is geen traject met een duidelijk pad of een voorspelbaar resultaat, en toch hebben jullie mij altijd de ruimte gegeven om mijn eigen keuzes te maken en mijn eigen richting te volgen.

Wat voor mij misschien nog het meest betekend heeft, is dat jullie nooit hebben getwijfeld aan die vrijheid. Jullie vertrouwen en de waarden die jullie mij hebben meegegeven vormen nog steeds de basis van de manier waarop ik naar werk, verantwoordelijkheid en ambitie kijk.

Alberto, María Rosa y Alberto. Hace tiempo que me habéis hecho sentir parte de la familia. Durante estos años habéis vivido este proceso muy de cerca: los momentos de presión, las dudas inevitables, pero también las celebraciones cuando finalmente las cosas salían bien.

Sé que habéis seguido este camino con orgullo y con interés genuino por el trabajo que estaba realizando. Ese apoyo, incluso desde la distancia, siempre se ha sentido. Gracias también por haberme acogido desde el principio con tanta naturalidad y generosidad.

Claudia, hace cuatro años tomaste una decisión que decía mucho sobre quién eres. Este doctorado significaba posponer nuestros planes de mudarnos a España durante varios años más. Aun así, cuando recibí la noticia de que había sido aceptado, estabas incluso más emocionada que yo.

Ese momento refleja perfectamente el tipo de persona que eres. Durante estos años hemos vivido en países distintos, organizando nuestra vida entre viajes, llamadas y fines de semana que siempre terminaban demasiado rápido. A pesar de todo, nunca dudaste de que valía la pena seguir adelante.

Este libro lleva mi nombre, pero tu apoyo está presente en cada una de estas páginas. Cerrar este capítulo significa finalmente poder empezar el siguiente.

Esta vez juntos, en nuestro nuevo hogar en Madrid.

Abstract

Artificial Intelligence (AI) is revolutionizing healthcare by enabling faster diagnoses, predicting patient outcomes, and supporting more personalized treatments. From detecting early signs of cancer in imaging to forecasting life-threatening complications in intensive care units, machine learning has shown promise to enhance clinical decision-making and improve patient outcomes. However, these advancements depend on access to large and diverse datasets, something that is difficult to achieve in the healthcare domain.

Due to strict privacy regulations such as the General Data Protection Regulation (GDPR) in the European Union and the Health Insurance Portability and Accountability Act (HIPAA) in the United States, aggregating sensitive patient data across institutions is often infeasible. Hospitals and research centers often operate within data silos, unable to share information due to ethical concerns, legal constraints, and institutional policies. As a result, AI systems are frequently trained on limited datasets that do not capture the full variability of patient populations, leading to biased models with limited generalizability.

To address these challenges, this work explores Federated Learning (FL) as a privacy-enhancing alternative to centralized machine learning. FL enables multiple institutions to collaboratively train AI models without transferring patient data. Each participating hospital trains a local model on its own data and shares only the learned model updates. These are then aggregated into a global model in a way that ensures no raw data ever leaves its source, preserving privacy while facilitating learning at scale.

Yet, deploying FL in real-world healthcare settings remains an open challenge. Medical data is highly heterogeneous due to variations in patient demographics, clinical workflows, and diagnostic practices. Moreover, institutions differ in their computational resources and availability, and transmitting updates over secure channels introduces latency and potential attack surfaces. FL also faces algorithmic challenges such as slower convergence, non-IID data

distributions, and fairness concerns, particularly when certain institutions dominate participation or data representation.

This work aims to provide (partial) solutions to several of these challenges in order to facilitate further adoption of FL in healthcare. First, we propose a privacy-enhancing client recruitment strategy that leverages global task-specific statistics to improve learning efficiency and reduce computational complexity. Second, we extend this approach by incorporating local resource constraints and defining a client-level utility function that balances cost and informativeness. Third, we introduce a heuristic selection mechanism designed to better accommodate the non-IID nature of medical data, improving robustness and convergence. Finally, we apply FL to complex learning pipelines, including medical image captioning, demonstrating how interpretability and modularity can be preserved even in decentralized training scenarios.

Each of these methods is evaluated using real-world clinical datasets, including intensive care unit (ICU) records and medical imaging data. Across these use cases, our proposed approaches outperform state-of-the-art baselines, or at the very least perform on par, showcasing improvements in model performance, scalability, interpretability, and privacy enhancement.

Ultimately, this work presents a roadmap for practical and ethical AI deployment in healthcare. By bridging the gap between technical feasibility and regulatory constraints, it offers hospitals and researchers the tools to use AI without compromising patient trust. Through Federated Learning, we can enable data-driven healthcare innovation and decision making while alleviating typical concerns related to privacy, generalizability and utility.

Beknopte samenvatting

Artificiële Intelligentie (AI) zorgt voor een revolutie in de gezondheidszorg door snellere en betere diagnoses te faciliteren, de resultaten van patiënten te voorspellen en gepersonaliseerde behandelingen te ondersteunen. Van het detecteren van kanker in vroege fases via medische beeldvorming tot het voorspellen van levensbedreigende complicaties op intensive zorgen, machine learning heeft aangetoond veelbelovend te zijn voor het verbeteren van klinische besluitvorming. Deze vooruitgang is echter afhankelijk van toegang tot grote en diverse datasets, iets wat moeilijk te bereiken is in de gezondheidszorg.

Vanwege strenge privacyregels, zoals de General Data Protection Regulation (GDPR) in de Europese Unie en de Health Insurance Portability and Accountability Act (HIPAA) in de Verenigde Staten, is het samenvoegen van gevoelige patiëntgegevens tussen instellingen vaak niet haalbaar. Ziekenhuizen en onderzoekscentra werken vaak in silo's en kunnen geen, of moeilijk, informatie delen vanwege ethische bezwaren, wettelijke beperkingen en institutioneel beleid. Als gevolg hiervan worden AI-systemen vaak getraind op beperkte datasets die niet de volledige variabiliteit van patiëntenpopulaties vastleggen, wat leidt tot modellen met bias en beperkte generaliseerbaarheid.

Om deze uitdagingen aan te pakken, onderzoekt dit werk Federated Learning (FL) als een privacyverhogend alternatief voor gecentraliseerde machine learning. FL stelt meerdere instellingen in staat om samen AI-modellen te trainen zonder patiëntgegevens over te dragen en met elkaar te delen. Elke deelnemende instelling traint een lokaal model op zijn eigen gegevens en deelt alleen de modelupdates. Deze worden vervolgens samengevoegd tot een globaal model op een manier die ervoor zorgt dat geen enkele ruwe data de bron verlaat, waardoor privacy behouden blijft terwijl leren op schaal mogelijk wordt.

Toch blijft het inzetten van FL in de echte gezondheidszorg een uitdaging. Medische gegevens zijn zeer heterogeen door variaties in de demografie van patiënten, klinische workflows en diagnostische praktijken. Bovendien verschillen

instellingen in hun middelen en beschikbaarheid, en het verzenden van updates via beveiligde kanalen introduceert vertraging en overhead. FL heeft ook te maken met algoritmische uitdagingen zoals langzamere convergentie en non-IID data.

Dit werk heeft als doel (gedeeltelijke) oplossingen te bieden voor verschillende van deze uitdagingen om verdere toepassing van FL in de gezondheidszorg te vergemakkelijken. Ten eerste stellen we een privacyverhogende wervingsstrategie voor die gebruik maakt van globale taakspecifieke statistieken om de leerefficiëntie te verbeteren en de computationele complexiteit te verminderen. Ten tweede breiden we deze aanpak uit door lokale resourcebeperkingen op te nemen en een nutsfunctie op cliëntniveau te definiëren die kosten en informativiteit in balans brengt. Ten derde introduceren we een heuristisch selectiemechanisme dat ontworpen is om beter rekening te houden met de non-IID aard van medische gegevens, wat de robuustheid en convergentie verbetert. Tot slot passen we FL toe op complexe use-cases, waaronder het schrijven van radiologische verslagen, waarbij we laten zien hoe de interpreteerbaarheid en modulariteit behouden kunnen blijven, zelfs in gedecentraliseerde trainingsscenario's.

Elk van deze methoden wordt geëvalueerd met behulp van echte klinische datasets, waaronder ICU-dossiers (Intensive Care Unit) en medische beeldgegevens. In al deze use cases presteren de door ons voorgestelde benaderingen beter dan de state-of-the-art baselines, of op zijn minst gelijkwaardig, met verbeteringen in modelprestaties, schaalbaarheid, interpreteerbaarheid en privacyverbetering.

Uiteindelijk presenteert dit werk een roadmap voor een praktische en ethische toepassing van AI in de gezondheidszorg. Door de kloof tussen technische haalbaarheid en wettelijke beperkingen te overbruggen, biedt het ziekenhuizen en onderzoekers de middelen om AI in te zetten zonder het vertrouwen van de patiënt te schaden. Door middel van Federated Learning kunnen we datagestuurde innovatie en besluitvorming in de gezondheidszorg mogelijk maken en tegelijkertijd de gekende zorgen over privacy, generaliseerbaarheid en bruikbaarheid verhelpen.

List of Abbreviations

AI Artificial Intelligence.

AMICCAM Augmented Medical Image Captioning with Class Activation Maps.

AUC Area Under the Receiver Operating Characteristic Curve.

BCE Binary Cross Entropy.

BERT Bidirectional Encoder Representations from Transformers.

BERTSc BERTScore.

BLEU Bilingual Evaluation Understudy.

BP Brevity Penalty.

CDSS Clinical Decision Support System.

CIDEr Consensus-based Image Description Evaluation.

CNN Convolutional Neural Network.

CvT Convolutional Vision Transformer.

CXR Chest X-ray.

CXR-BERT Chest X-ray BERT.

DL Deep Learning.

DP Differential Privacy.

EF Early Fusion.

ER Entity Recognition.

F1 F1-score.

FedSim Federated Simulation.

FedSurg Federated Surgery.

FID Fréchet Inception Distance.

FL Federated Learning.

FPR False Positive Rate.

GDPR General Data Protection Regulation.

GPT Generative Pre-trained Transformer.

GRU Gated Recurrent Unit.

HE Homomorphic Encryption.

HIPAA Health Insurance Portability and Accountability Act.

ICU Intensive Care Unit.

LCS Longest Common Subsequence.

LLM Large Language Model.

LoS Length of Stay.

LSTM Long Short-Term Memory.

MAE Mean Absolute Error.

MAPE Mean Absolute Percentage Error.

MF Mid Fusion.

ML Machine Learning.

MPC Multi-Party Computation.

MSE Mean Squared Error.

MSLE Mean Squared Logarithmic Error.

NLP Natural Language Processing.

Non-IID Non Independent and Identically Distributed.

P Precision.

PACS Picture Archiving and Communication System.

RAFA Robust Adaptive Federated Aggregation.

ReLU Rectified Linear Unit.

RNN Recurrent Neural Network.

ROUGE Recall-Oriented Understudy for Gisting Evaluation.

SOTA State-of-the-Art.

SSIM Structural Similarity Index Measure.

TPR True Positive Rate.

List of Symbols

α, β, γ	Adaptive weight parameters
$\Delta \mathbf{w}$	Aggregated global model update
$\Delta \mathbf{w}_i$	Model update from client i
Δw_l	Aggregated update for layer l
$\Delta w_{i,l}$	Update from client i for layer l
ϵ	Privacy budget in Differential Privacy
η	Learning rate
$\mathbf{w}_0$	Initial global model parameters
$\mathbf{w}_t$	Global model parameters at round t
$\mathbf{w}_g$	Global model parameters
$\mathbf{z}$	Latent representation vector
$\mathcal{C}_{\text{random}}$	Randomly selected client subset
$\mathcal{C}_{\text{selected}}$	Set of selected clients
$\mathcal{C}_{\text{top}}$	Clients with top contribution scores
$\mathcal{D}_i$	Dataset of client i
$\mathcal{F}$	Fraction of selected clients per round
$\mathcal{L}$	Loss function
$\mathcal{X}_i$	Feature space of client i
$\mathcal{Y}_i$	Label space of client i

μ	Proximal regularization parameter in FedProx
$\nabla f_i(w)$	Gradient of local loss for client i
σ	Activation or noise parameter
θ	Model parameter vector
E	Number of local steps per communication round
$f_i(w)$	Local objective function of client i
g	Gradient vector
H_i	Heterogeneity score for client i
I_i	Informativeness score for client i
J_i	Client i's contribution score
J_{total}	Sum of all contribution scores
l	Index over model layers
N	Total number of clients
S_i	Statistical divergence score for client i
T	Number of communication rounds
w	Model parameters (general)
x_i	Sample data point for client i
y_i	Label corresponding to x_i

Contents

Abstract v

Beknopte samenvatting vii

List of Abbreviations xiii

List of Symbols xvi

Contents xvii

List of Figures xxi

List of Tables xxv

1 Introduction and Background 1
- 1.1 Motivation . . . 1
- 1.2 Research problem . . . 2
- 1.3 Structure and contributions of this work . . . 3
- 1.4 Artificial intelligence in healthcare . . . 6
- 1.5 History of federated learning . . . 7
- 1.6 Federated learning paradigm . . . 9
- 1.7 Data privacy & security in FL . . . 12
- 1.8 Challenges in federated learning . . . 14
  - 1.8.1 Data heterogeneity . . . 14
  - 1.8.2 Communication efficiency . . . 16
  - 1.8.3 Client selection and participation . . . 18
  - 1.8.4 Security threats . . . 20
  - 1.8.5 Existing FL deployments . . . 22
- 1.9 Summary and research directions . . . 23

2 Client Recruitment for Federated Learning in ICU Length of Stay Prediction **25**
2.1 Introduction . . . 26
2.2 Related work . . . 28
2.3 Federated learning basics . . . 29
2.4 Methods . . . 30
2.4.1 Data . . . 30
2.4.2 Model architecture . . . 31
2.4.3 Client recruitment . . . 33
2.4.4 Experimental settings . . . 36
2.5 Results and discussion . . . 38
2.5.1 Client recruitment results . . . 39
2.5.2 Recruitment parameter effects . . . 41
2.6 Conclusion, Limitations & Future Work . . . 43
2.6.1 Conclusion . . . 43
2.6.2 Limitations & future work . . . 43
2.7 Supplementary context . . . 43
2.7.1 Situating this work . . . 44
2.7.2 Recruiting on the outcome distribution . . . 44
2.7.3 Sensitivity to binning . . . 45
2.7.4 The γ_{th} parameter . . . 45
2.7.5 GRU and LSTM . . . 46
2.7.6 Use of ReLU . . . 47
2.7.7 All clients versus recruited clients . . . 47
2.7.8 Selection paradox . . . 48

3 Target Informed Client Recruitment for Efficient Federated Learning in Healthcare **49**
3.1 Background . . . 50
3.2 Methods . . . 56
3.2.1 Client recruitment . . . 56
3.2.2 Experimental setup . . . 59
3.2.3 Regression . . . 59
3.2.4 Classification . . . 70
3.3 Results . . . 75
3.3.1 Regression . . . 75
3.3.2 Classification . . . 78
3.4 Discussion . . . 82
3.4.1 Regression . . . 82
3.4.2 Classification . . . 86
3.5 Conclusions . . . 87
3.6 Supplementary context . . . 88
3.6.1 Hardware-aware recruitment . . . 88

3.6.2 Batch size and convergence . . . 89
3.6.3 Hardware-aware selection . . . 89

4 RAFA: Robust Adaptive Federated Aggregation for 3D MRI Segmentation 91
4.1 Introduction . . . 92
4.2 Related work . . . 93
4.3 Methods . . . 95
4.3.1 RAFA: Robust adaptive federated averaging . . . 95
4.3.2 Client selection . . . 96
4.3.3 Weight aggregation . . . 98
4.3.4 Adaptive weighting . . . 99
4.3.5 The RAFA pipeline . . . 101
4.4 Experimental setup . . . 101
4.4.1 Data . . . 102
4.4.2 Data distributions . . . 105
4.4.3 Models . . . 108
4.4.4 Metrics . . . 110
4.4.5 Federated learning . . . 113
4.5 Results . . . 114
4.6 Discussion . . . 117
4.7 Conclusion . . . 118
4.8 Supplementary context . . . 118
4.8.1 Design choices . . . 119
4.8.2 Interaction between α, β, and γ . . . 119
4.8.3 Stability versus reactivity . . . 120
4.8.4 Significance versus informativeness . . . 121
4.8.5 Relationship to reinforcement learning . . . 121
4.8.6 Handling new clients . . . 121
4.8.7 Exponential smoothing . . . 122

5 From Pixels to Insights: Leveraging Class-Dependent Activation Maps for Augmented Medical Image Captioning 123
5.1 Introduction . . . 124
5.2 Related work . . . 126
5.2.1 CXR classification . . . 127
5.2.2 Vision encoder-decoder architectures . . . 127
5.2.3 Explainability in AI . . . 128
5.2.4 Federated learning for medical image captioning . . . 129
5.3 Methods . . . 129
5.3.1 Data . . . 130
5.3.2 Classification . . . 130
5.3.3 Report generation: vision encoder-decoder . . . 133

5.4 Two-stage federated learning . . . 137
5.4.1 Evaluation metrics . . . 138
5.5 Results . . . 139
5.5.1 Classification performance . . . 142
5.5.2 Radiology report generation . . . 142
5.6 Discussion . . . 144
5.7 Conclusion and limitations . . . 145
5.8 Supplementary context . . . 145
5.8.1 Validation of generated captions . . . 146
5.8.2 Saliency is not causality . . . 146
5.8.3 Error propagation . . . 147
5.8.4 Separating along classes . . . 147
5.8.5 Improvement over single-center learning . . . 148
5.8.6 Non-IID bottleneck . . . 149

6 Conclusions and Future Perspectives 151
6.1 Contributions . . . 152
6.2 Limitations and future research . . . 155
6.3 Future perspective: towards integrated federated decision support systems . . . 157

A Appendix 161
A.1 Features . . . 161
A.2 Federated Algorithms . . . 161

Bibliography 165

List of Figures

1.1 Visual overview of the chapters contained in this work, where they should be situated within the scope of Artificial Intelligence, Machine Learning and Federated Learning. The figure shows how the chapters are related and which core topics of the domain they discuss and address. 4
1.2 Types of Federated Learning: (a) Federated Transfer Learning, (b) Vertical Federated Learning, and (c) Horizontal Federated Learning 11

2.1 The distribution of the target (LoS) in the global training data with a cutoff at 25 days on the x-axis. The y-axis represents the number of patients (frequency) with their corresponding LoS in days as indicated on the x-axis. 34
2.2 Subset of the, in total 189, local target distributions in terms of Length of Stay in days and the corresponding patient counts. The distributions shown correspond to, from left to right, the hospitals with the identifiers; 24, 74, 100 and 143. 35
2.3 General overview of the client recruitment and federated training procedure with M the central server and $c_1, ..., c_n$ the potential clients. Each of the clients is depicted with their corresponding target distributions for which the red cross indicates clients that are discarded to participate in the federation based on the recruitment procedure. M is initially in charge of recruiting clients based on the representativeness of the locally hosted data and will subsequently orchestrate the federated training procedure for either the GRU or LSTM models. The GRU and LSTM cells included are a visual representation of the governing equations in (3.5) and (3.6). 37

2.4 Runtime versus MSLE (left) and runtime versus MAE (right) in function of gradually increasing values for γ_{th}. Corresponding to a gradually increasing number of recruited clients, denoted as N_{rc}, for each of the training iterations. 42

3.1 Overview of the experimental setup of the client recruitment approach for federated learning in the regression and classification settings. 60
3.2 The global target distribution calculated as P_g, (a), and a subset of six local target distributions of the, in total, 189 target distributions, (b). Each of the distributions shows the frequency in terms of patients with an observed LoS, restricted until 25 days, locally and globally. 63
3.3 Grad-CAM visualizations of the pixel importance for each of the 3 labels corresponding to the multi-view CXR selected from the test set. Grad-CAM visualizations are shown on the right and, the raw postero anterior and lateral views are shown on the left. 84

4.1 Five randomly selected images from MNIST (top row) as well as five randomly drawn images from CIFAR-10 (bottom row) . . . 103
4.2 A single slice from the FeTS dataset with the 4 main modalities, the corresponding multi-label maks and the representative overlay of the mask on top of the ground truth T1 image. 104
4.3 Raw T1 MRI scan with the segmentation mask overlay for 100 subsequent slices (slice 25 - 125). 105
4.4 Overview of the data distributions used in this work for the MNIST and CIFAR-10 datasets. The plots show here reflect the MNIST distributions and absolute values. The same characteristics, however, hold for the CIFAR-10 distributions. . 107
4.5 Overview of the institutional and artificial data partitioning as per the FeTS meta data. 108
4.6 Results for the three FL strategies after 30 rounds of training on the exponential datasplit over 100 clients for the CIFAR-10 dataset. The image shows the training curves in terms of accuracies on the left whereas the resulting performance after 30 rounds is shown on the right along with the achieved standard deviation. 116

5.1 Example of an original, raw CXR image in the dataset with the corresponding findings and impression sections from the radiology report. 130

5.2 The figure represents the DenseNet-121 architecture as applied to CXR classification. The network starts with a convolution followed by a series of four Dense Blocks (DB1–DB4) containing 6, 12, 24, and 16 convolutional layers, respectively. Each Dense Block is preceded by a transition layer consisting of a 1×1 convolution followed by 2×2 average pooling. The model concludes with global average pooling and a fully connected output layer producing multi-label predictions for 14 chest related pathologies. 131
5.3 Histogram showing the distribution of the number of observed classes per instance (left) and the number of observed topics in the corresponding reports (right). 134
5.4 The proposed two-stage pipeline. A DenseNet-121 classifier produces class-specific Grad-CAMs that are concatenated with the raw image to form a 7-channel tensor. A ViT encoder converts this tensor into visual embeddings, which interact with text embeddings through a cross-modal attention block and a GPT-2 decoder to generate the structured radiology report. 136
5.5 Example of the augmented input, the raw CXR is enhanced with the top 4 saliency maps serving as a 7-channel input to the Vision Encoder-Decoder model. 139
5.6 Grad-CAM activation maps for three pathology classes, overlaid on chest X-rays. These saliency maps serve as additional input channels to the captioning model, encoding pathology-specific visual semantics. 142

List of Tables

2.1 The extracted and preprocessed data cohort showing the train, validation and test splits used throughout as well as the descriptive statistics, mean and median Length of Stay (LoS), number of hospitals and features in the cohort. 31
2.2 Model hyperparameter settings used for both central and federated training. 36
2.3 Settings specific to the federated strategies with ϵ the total number of clients in the federation, δ the number of clients randomly sampled from ϵ in each round of training and (γ_{dv}, γ_{sa}, γ_{th}) the hyperparameters for client recruitment 39
2.4 Model performance for central and federated models with and without client recruitment. Statistical significance among the federated models in comparison to Fed-SC is indicated as * at the 5% significance level and ** at the 1% significance level. . . 40
2.5 Parameter settings for the quality greedy (QG) and quantity greedy (DG) approaches. 41
2.6 Model performance for federated training with data greedy and quality greedy recruitment strategies. 42

3.1 Condensed overview of seminal work related to federated learning approaches, client selection and aggregation. 54
3.2 Overview of the extracted and preprocessed data cohort for the eICU dataset. 61
3.3 Extracted temporal and static features from eICU dataset . . . 62
3.4 Hyperparameter settings used for both central and federated training. 66
3.5 CXR data set observation count and corresponding splits . . . 71
3.6 Overview of the positive, negative and uncertain label observations for each of the 14 classes in the MIMIC CXR data set. . . 71
3.7 DenseNet-121 model architecture. 73

3.8 Paramater settings for the client recruitment approach in both the regression and classification settings, with ϵ the total number of clients in the federation, δ the number of clients randomly sampled from ϵ in each round of training and (γ_{dv}, γ_{sa}, γ_{tr}, γ_{th}) the hyperparameters for client recruitment. 76
3.9 Model performance for central and federated models with and without client recruitment. Statistical significance among the federated models in comparison to Fed-SC is indicated as * at the 5% significance level and ** at the 1% significance level. . . 77
3.10 Model performance for the Fed-SRC approach retrained with a subset of recruited clients that either constitute the least representative clients according the the recruitment approach or a set of random clients. 79
3.11 Model performance for Fed-SRC tested against data from non-recruited clients only with and without a single round of domain adaptation (DA). 80
3.12 Model performance expressed in terms of AUC scores for each of the classes of interest for central and federated models with and without client recruitment. 81
3.13 Classification training time and AUC over the classes of interest 82
3.14 Similarity in terms of FID (0 would signify perfect similarity) and SSIM (1 would signify perfect similarity) for the Grad-CAM activation maps corresponding to 2000 images in the test set and the Fed-SC and Fed-SRC models respectively with activation maps from the central model serving as the ground truth. . . . 83

4.1 CNN Architecture for MNIST and CIFAR-10 Classification . . 109
4.2 Attention U-Net 3D Architecture Overview 111
4.3 Experimental Setup Overview 114
4.4 Performance Metrics for Federated Learning Strategies on MNIST and CIFAR-10 Datasets 115
4.5 Performance Metrics for 3D Segmentation on Test Set with Institutional and Artificial Data Distributions 117

5.1 ViT–GPT2 Training Hyperparameters 136
5.2 Federated learning hyperparameter settings used in both training stages. 139
5.3 Summarized representation of the different metrics used to evaluate clinical and linguistic accuracy of the two-stage federated approach. 140
5.4 Classifier performance (DenseNet-121) measured in terms of AUC scores for selected pathology classes. 143

5.5 Evaluation metrics comparing the baseline ViT-GPT2 model against AMICCAM and AMICCAM-FL on the full MIMIC-CXR dataset. 144

A.1 Extracted temporal and static features 163

Chapter 1

Introduction and Background

1.1 Motivation

Artificial Intelligence (AI) is transforming healthcare by enabling data-driven decision-making, enhancing diagnostic workflows, and personalizing patient care. From radiological interpretation and pathology classification to intensive care monitoring and predictive analytics, machine learning (ML) and deep learning (DL) models have achieved near-human performance in clinical settings [27, 70].

This transformation is driven by vast amount of available health data and increasingly powerful computational techniques and resources. Convolutional neural networks (CNNs) have resulted in novel approaches for medical image analysis [62], recurrent neural networks (RNNs) and transformers are improving the predictive modeling from electronic health records (EHRs), and large language models (LLMs) are now being explored for clinical reasoning and report generation [84]. These innovations come with several benefits, including improved diagnostic accuracy, reduced diagnostic delay, optimized hospital operations, and so on.

However, utilizing the full potential of AI in healthcare remains an open challenge. The development of performant and generalizable AI models typically requires access to diverse, large-scale, and representative datasets. In clinical settings, such data are often distributed across institutions. Privacy regulations such as the General Data Protection Regulation (GDPR) [28] in Europe and the Health Insurance Portability and Accountability Act (HIPAA) [14] in the United States impose strict constraints on data access, usage, and transfer [80], making centralized data collection a realistic challenge and more often than not, legally

infeasible.

Apart from the regulatory limitations, centralized training pipelines come with ethical and operational risks. Patient consent, data sovereignty, and institutional trust are realistic concerns, especially when handling privacy-sensitive health data. Moreover, technical limitations such as data heterogeneity, inconsistent labeling standards, and variation in diagnostic equipment pose additional challenges.

These challenges have sparked growing interest in privacy-preserving and decentralized learning paradigms. Among these, Federated Learning (FL) has emerged as a promising solution [68]. FL enables multiple institutions or devices to collaboratively train a shared global model while keeping data localized. Only model updates or gradients are exchanged, preserving data sovereignty and reducing privacy risk. This paradigm is well suited to alleviate the legal, and ethical requirements of operating with data in the healthcare domain.

Nevertheless, vanilla FL comes with several challenges. This thesis is motivated by the need to address the practical, technical, and theoretical limitations of FL in real-world healthcare deployments. As outlined in Chapter 1, current FL frameworks face several challenges related to client selection, aggregation algorithms, efficiency and data heterogeneity.

This work aims to address these core challenges by introducing approaches that further optimize the FL paradigm, building towards a set of tools that will facilitate the adoption of FL in practice and spark further research in similar avenues.

1.2 Research problem

As indicated, FL is a promising paradigm that can alleviate the regulatory and ethical challenges related to working with distributed medical data. By allowing institutions to collaboratively train models without transferring raw data, FL enhances local data sovereignty while enabling knowledge and intelligence sharing at scale. However, FL faces multiple unresolved challenges that hinder its adoption in healthcare. Clinical datasets are often not identically and independently distributed: patient demographics, prevalence rates, and diagnostic criteria vary widely across institutions. Such heterogeneity leads to unstable optimization, biased global models, and fairness concerns. Moreover, the iterative nature of FL training creates substantial communication overhead.

Another layer of complexity arises from the reality that not all institutions can participate equally in every training round. Availability, computational power, and administrative constraints vary across sites, leading to irregular client participation. In addition, when working with multiple institutions, data quality and availability becomes an issue. At the same time, FL systems remain vulnerable to a range of privacy and integrity attacks. Even when raw data are not shared, adversaries can exploit model updates to infer sensitive information.

Based on these open challenges the following main research question is formulated; Can we formulate improvements to the vanilla FL paradigm such that we can (i) improve the privacy preservation aspect, (ii) improve efficiency by addressing challenges related to communication overhead and convergence, (iii) work towards faster and more reliable convergence in non-iid settings, and (iv) demonstrate feasibility of FL in the more complex, practical, medical deep-learning pipelines?

The different chapters in this work aim to formulate an answer to this main question and the corresponding sub-research questions. The overall goal and corresponding contribution of this work is to provide valuable insights into what optimizations to FL can be made such that the current challenges are eliminated or reduced to foster further adoption in privacy-sensitive domains. This in turn will result in enhanced cross-institutional collaboration, improved intelligence and performance of the resulting solutions and ultimately reduced workload for medical practitioners. Indirectly, this will result in improved patient satisfaction, one of the major key performance indicators for medical institutions globally.

1.3 Structure and contributions of this work

This work is organized to first provide extensive background on FL as a paradigm, its origins, limitations and current well-known deployments. Each of the subsequent chapters then formulates an improvement to vanilla FL in relation to our main research question. More specifically, the central idea and contributions for each of the chapters in this work are outlined in the following paragraphs. Figure 1.1, shows where each of the chapters can be situated and how they are related to each other.

Chapter 1 introduces the foundations of federated learning, situating it within the broader context of machine learning and privacy-preserving computation. It outlines the historical evolution of FL, its key design paradigms, and discusses the core challenges that come with widespread deployment in healthcare, including data heterogeneity, communication bottlenecks, and privacy challenges.

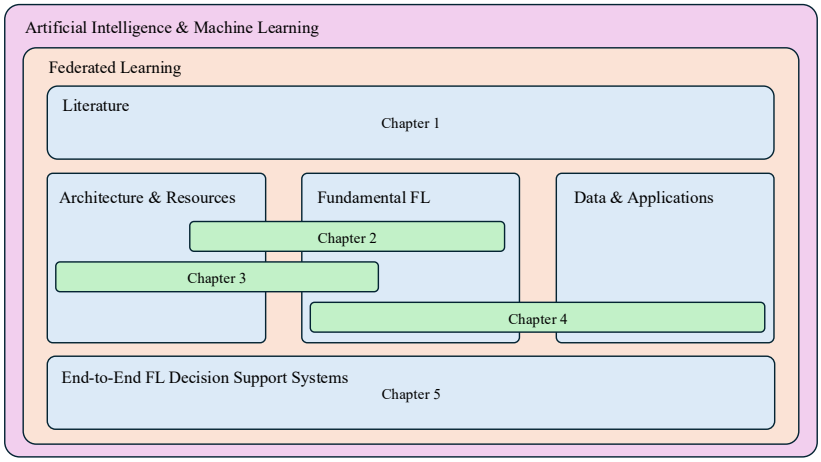


Figure 1.1: Visual overview of the chapters contained in this work, where they should be situated within the scope of Artificial Intelligence, Machine Learning and Federated Learning. The figure shows how the chapters are related and which core topics of the domain they discuss and address.

Chapter 2 investigates the problem of client recruitment as a pre-training step in Federated Learning networks for medical applications. Current FL implementations often assume that all clients participate in model training. This assumption leads to unnecessary communication overhead, slow convergence, and uneven contributions, issues that become more pronounced in real-world non-IID healthcare settings. To address this, we propose a lightweight client recruitment strategy that leverages only local output distributions and sample sizes to estimate the utility of prospective participants before training. Applied to the task of predicting patient Length of Stay using data from 189 intensive care units, the approach demonstrates that selectively recruiting clients significantly reduces computational costs and training time without sacrificing predictive performance. By strategically limiting participation to the most informative and cost-effective clients, this work contributes to the overarching goal of enhancing the efficiency, scalability, and privacy-preserving capabilities of the vanilla FL paradigm.

Chapter 3 proposed an improvement to the FL paradigm in healthcare by refining the process of client recruitment through a representativeness-driven selection strategy. Unlike previous approaches that relied solely on statistical distributions or uniform participation, this method integrates knowledge of local

hardware capabilities alongside data characteristics, specifically, the divergence of local target distributions and sample sizes, to estimate a client's expected utility. Evaluated across representative medical regression and classification tasks, the proposed strategy achieves model performance on par with, or exceeding, both standard federated and centralized training approaches, while reducing training time by a factor of 3–4. Crucially, the approach achieves these gains while significantly lowering communication costs. Moreover, excluded clients remain able to personalize global models through local fine-tuning, confirming the inclusive potential of the framework. By explicitly modeling representativeness as a function of target distribution, data volume, and hardware constraints, this is a practical step forward in addressing real-world challenges related to FL.

Chapter 4 introduces Robust Adaptive Federated Aggregation (RAFA), an aggregation framework designed to address the challenges of data heterogeneity in federated learning. Unlike prior approaches that either regularize local updates (e.g., FedProx) or apply fixed momentum strategies (e.g., FedAvgM), RAFA dynamically adjusts client influence based on a recency-weighted history of their contributions, ensuring robustness to skewed and non-IID data distributions. The method is validated on standard benchmarks such as CIFAR-10 and MNIST, real-world benefits are shown through an implementation for the real-world FeTS 3D MRI tumor segmentation dataset. This proposed approach is relevant in medical federations, where data silos not only differ in volume but also in data type and clinical protocols. By enabling reliable aggregation under severe statistical divergence.

Chapter 5 presents a interpretable, and federated two-stage learning pipeline for medical image captioning, addressing the increasing demand for automated support in radiological reporting, particularly in high-volume modalities such as chest X-rays. Traditional captioning systems often operate as black boxes, relying solely on learned latent features, which limits their clinical acceptability due to lack of transparency. To overcome this, the proposed method introduces an augmented image representation that combines original CXR pixel data with class-specific saliency information derived from Gradient-weighted Class Activation Maps (Grad-CAM). This is achieved by first training a multi-label classifier to identify the most probable pathologies and subsequently using these predictions to extract top-k class activation maps. These saliency maps are then fused with the original image to form a semantically enriched (3+k)-channel input for the captioning model. The resulting pipeline is embedded within a federated learning framework, demonstrating the feasibility of deploying complex multi-stage deep learning architectures in decentralized healthcare environments without compromising patient data privacy. Evaluated on the MIMIC-CXR dataset, this approach matches state-of-the-art performance across multiple

language generation metrics while significantly improving interpretability.

Chapter 6 concludes the thesis by summarizing the findings and contributions, and reflecting on the broader implications for privacy-enhancing AI in healthcare.

1.4 Artificial intelligence in healthcare

The recent advancements in Artificial Intelligence (AI) and Machine Learning (ML) have sparked a shift in the medical domain, transitioning from rule-based expert systems to modern, data-driven approaches powered by statistical learning and deep neural networks. In healthcare, these modern approaches are often adopted in the form of Clinical Decision Support Systems (CDSS) where advanced methods provide practitioners with relevant inights, analyses and information in support of the (clinical) decision making process [26, 70].

In particular, the introduction of deep learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), improved performance in medical imaging interpretation, temporal signal analysis, and natural language understanding. These technologies have enabled AI systems perform on par, or even surpass human-level accuracy in tasks like pneumonia detection on chest radiographs [85], skin cancer classification [26], diabetic retinopathy screening [35] and many more.

In real-world healthcare environments, AI-driven systems support clinicians in a number of use cases[1]:

- **Medical Imaging:** Deep learning models assist radiologists in detecting tumors, segmenting lesions, and prioritizing urgent cases in modalities like CT, MRI, and histopathology [62].
- **Patient Monitoring:** Time-series models analyze ICU telemetry data to anticipate cardiac arrest, sepsis onset, respiratory complications, readmission risk or predict length of stay [79, 107].
- **Clinical Decision Support:** ML models support physicians by providing treatment recommendations, risk scores, and drug dosage predictions [5].

Despite these advances, the adoption of ML in healthcare faces several challenges, mainly related to the nature of medical data. Some of the major challenges in this setting relate to:

[1] This list serves an exemplary purpose and is by no means exhaustive.

- **Patient privacy:** Patient records are protected by stringent privacy regulations (e.g., GDPR in Europe [28], HIPAA [14] in the US).
- **Distributed and siloed:** Data is generated and stored across different hospitals, labs, and electronic health record systems, often in non-standardized formats.
- **Heterogeneity:** Data varies in type (images, notes, time-series), quality, and completeness across institutions.
- **Non-IID:** Disease prevalence, diagnostic protocols, and treatment outcomes differ between regions and patient populations.

Standard machine learning, and especially deep learning, workflows require a large amount of data which is often achieved by pooling, i.e. aggregating, data somewhere centrally in support of the learning process. This practice in itself introduces significant privacy, legal, and infrastructural risks. Transferring data across institutions is not only costly and logistically difficult, but also often and ethical and legal challenge, resulting in delayed or cancelled projects that could otherwise yield relevant insights that would benefit all stakeholders. These limitations have sparked research into privacy-enhancing, decentralized learning approaches. Intuitively, the open research question that most research in this field aims to answer can be summarized as follows; In what ways can models and insights be derived from distributed data that would benefit all stakeholders without violating patient privacy or the need to centrally aggregate data? As will be discussed in subsequent chapters, this general question comes with many secondary research questions that dive into efficiency, privacy and complexity of the proposed approaches.

One of the frameworks that is in the early stages of adoption in healthcare is Federated Learning [68], which enables collaborative model training across institutions while ensuring that data remains local. FL comes with interesting characteristics that by default enhance privacy and reduce the aforementioned risks. Therefore, FL as a paradigm is (i) widely studied in relation to the potential applications in healthcare and (ii) in the early stages of adoption for real-world, medical applications.

1.5 History of federated learning

Federated Learning was formally introduced by McMahan et al. in 2017 [67] as a privacy-preserving approach to decentralized machine learning. Along with the framework, McMahan et al. proposed the Federated Averaging

(FedAvg) algorithm [67]. This framework enables local devices or institutions to collaboratively train a shared global model without the need to centrally aggregate data. Initially developed for edge and mobile computing, such as personalized keyboard prediction [37], FL significantly reduces privacy risks by keeping data localized while only model updates are aggregated centrally [12].

Even though FL was coined as a framework in 2017, the foundation for decentralized, privacy-enhancing machine learning predates FL. In healthcare in particular, concerns over sensitive patient data, governance, and legal boundaries have long driven interest in models that "move the algorithm to the data" instead of centrally aggregating data. One example in the healthcare space is the Personal Health Train (PHT) initiative in the Netherlands, which fully predates the introduction of FL as it is known to this day. The PHT architecture allows algorithms (or "trains") to visit distributed data silos (e.g., hospitals, biobanks), where they are executed locally under strict data governance policies, and only aggregated results or model updates are returned. This approach fully aligns with the principles underlying federated learning [20, 106]. It is worth noting that federated and distributed learning concepts have also been formalized in the patent literature. The STADIUS patent on federated learning (US20190102670A1)[2] [15] describes mechanisms for secure, distributed model training in privacy-sensitive environments, predating much of the current academic FL literature.

Over recent years, research into medical FL gained traction which in turn is the driving factor behind large scale, entreprise-ready initiatives such as NVIDIA Clara that introduced federated workflows to facilitate training across multiple hospitals for medical imaging tasks without data pooling [89]. Similarly, the Federated Tumor Segmentation (FeTS) challenge demonstrated the power of FL for collaborative brain tumor segmentation across international institutions [4, 77, 87]. In the pharmaceutical sector, projects like MELLODDY enabled federated learning across multiple pharma companies for drug discovery, preserving proprietary datasets while enabling cooperative model training [40].

Thus, while FL was only coined FL following the seminal work by McMahan et al. [67], the foundations for have been established by earlier approaches related to distributed learning. In healthcare, such an example is the Dutch Personal Health Train. The convergence of these ideas with FL frameworks has since accelerated innovation in sensitive domains such as healthcare, where regulatory and ethical constraints on data sharing pose significant limitations.

[2] https://patents.google.com/patent/US20190102670A1/en

1.6 Federated learning paradigm

Traditional centralized machine learning requires aggregating data from multiple sources for training. This approach raises concerns in domains such as healthcare, where privacy regulations (e.g., GDPR, HIPAA), data ownership, and technical constraints often limit data sharing. Centralized data aggregation not only increases the risk of privacy breaches but also poses logistical and legal challenges [57].

Federated Learning offers a partial solution by enabling decentralized model training across multiple clients or institutions, where data remains local and only model updates are shared. Instead of sending raw data to a central server, each client trains a model on-site and shares model parameters or gradients with a coordinating server. These updates are then aggregated, using an aggregation algorithm such as Federated Averaging (FedAvg), into a new global model, which is redistributed to clients for the next training round [68]. This iterative protocol continues until a convergence criterion is met.

Despite the privacy-preserving advantages, FL faces several technical challenges. One of the major issues is the heterogeneity of data across participating clients. In practice, medical datasets are often non-independent and identically distributed (non-IID), leading to local model drift and slower or unstable convergence of the global model [57]. Addressing this requires robust optimization strategies and aggregation techniques that account for client imbalance, data skew, and system-level heterogeneity.

The term non-IID appears frequently in the FL literature but can refer to several distinct statistical phenomena. For clarity, this thesis adopts the following working definition. Data across clients is considered non-IID when the joint distribution $P(X, Y)$ varies across clients, i.e., for clients i and j, $P_i(X, Y) \neq P_j(X, Y)$. This condition encompasses several sub-types relevant to the healthcare applications studied in this work: (i) label distribution skew, where $P(Y)$ differs across clients, for example, hospitals observing different LoS distributions due to patient acuity, admission policies, or discharge practices; (ii) feature distribution skew, where $P(X)$ differs due to measurement equipment, clinical protocols, or patient demographics; (iii) concept shift, where $P(Y|X)$ varies, meaning the same patient profile may yield different outcomes at different institutions due to differing treatment protocols; and (iv) quantity skew, where clients differ in sample size, which interacts with non-IID challenges by amplifying the influence of data-rich clients during aggregation. It is also important to note that the IID assumption can be violated within a single client, for instance, temporal dependencies among patients admitted during the same epidemic violate the independence assumption locally. These distinctions

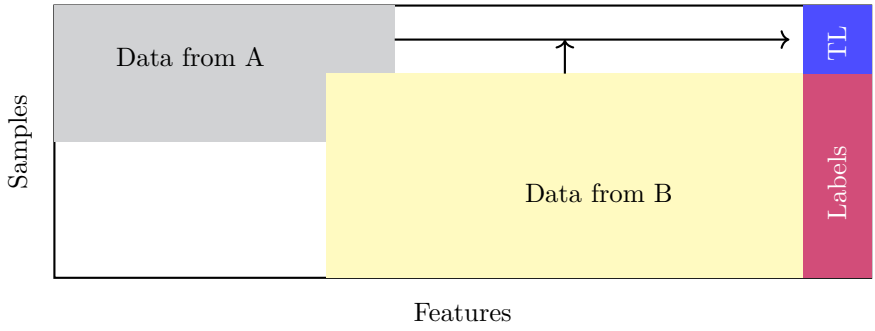
are important because different sub-types of non-IID data call for different mitigation strategies.

To better tailor FL to different collaboration scenarios, three main categories have been defined and are illustrated in Figure 1.2:

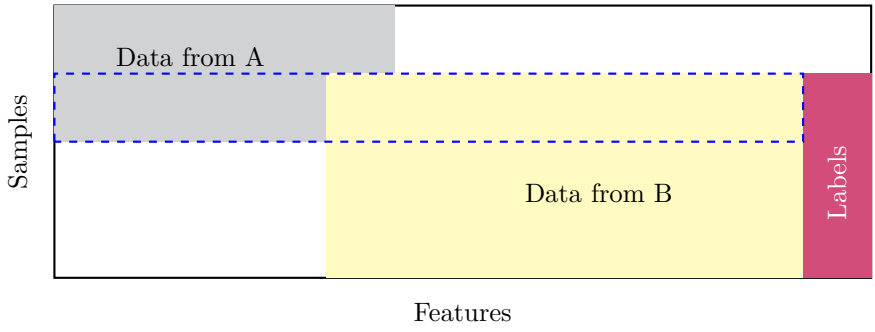
- **Horizontal Federated Learning (HFL):** Applicable when clients (e.g., hospitals) share the same feature space but have data on different patients.
- **Vertical Federated Learning (VFL):** Used when clients have different features for the same set of entities. For instance, genomic research centers and hospitals may jointly analyze patient-level data where one party holds genomic sequences and another holds clinical records.
- **Federated Transfer Learning (FTL):** Relevant when both feature and sample spaces have limited overlap. FTL enables knowledge transfer across domains or populations where data alignment is sparse or indirect, leveraging shared model representations to bridge the gap.

This work focuses mainly on optimizations that can be made for the horizontal federated learning framework, thus, identical feature space is assumed or enforced across clients.

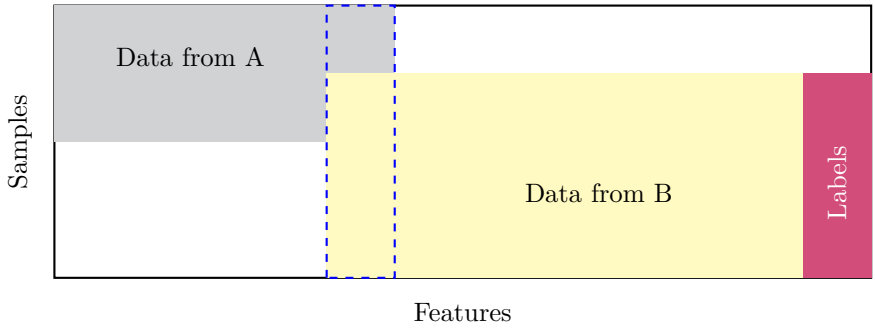
Before proceeding, it is important to clarify the target estimand that underlies the optimization objective throughout this thesis. When we refer to improving the "global model," we mean a single set of parameters optimized to minimize the weighted empirical risk across the recruited client population, where each client's contribution is proportional to its sample size. This corresponds to a pooled-population estimand. Alternative targets exist and are legitimate: one could optimize for average per-hospital performance (treating each hospital equally regardless of size), worst-site performance (a minimax objective), or a specific deployment site (personalization). Each choice implies different aggregation weights, recruitment criteria, and evaluation protocols. The pooled-population target was chosen because it aligns with the standard FedAvg aggregation and because, in the clinical tasks studied here, the primary interest lies in a model that is reliable across a broad patient population. Where relevant, individual chapters also report per-site and worst-site metrics to assess robustness under this choice.



(a) Federated Transfer Learning



(b) Vertical Federated Learning



(c) Horizontal Federated Learning

Figure 1.2: Types of Federated Learning: (a) Federated Transfer Learning, (b) Vertical Federated Learning, and (c) Horizontal Federated Learning

1.7 Data privacy & security in FL

Privacy and data security related challenges have been the main driving factor in the adoption and further development of FL, particularly within sensitive domains like healthcare. Regulatory frameworks such as the General Data Protection Regulation (GDPR) in Europe, the Health Insurance Portability and Accountability Act (HIPAA) in the United States, and the emerging European AI Act impose strict legal obligations on the processing of personal data. These regulations emphasize principles including data minimization, purpose limitation, transparency, and accountability, all of which influence how machine learning systems are designed and operated.

In FL, although raw data remains local, the exchange of model updates or gradients can still leak sensitive information. This has been demonstrated through various forms of reconstruction or inference attacks, where adversaries exploit intermediate training signals to infer attributes or even reconstruct original inputs. Consequently, FL alone is not inherently private, and additional safeguards must be introduced to protect participants from unintended data leakage.

A variety of privacy-enhancing technologies have been proposed and integrated with FL architectures to mitigate these risks:

Differential privacy introduces statistical noise to the data or model updates in a mathematically controlled manner, limiting the contribution of any individual data point. When applied in FL, DP mechanisms typically perturb gradient updates or model weights before transmission to the server, thereby ensuring that any single client's data cannot be confidently inferred from the final model [25]. The strength of DP lies in its formal, quantifiable privacy guarantees, often expressed through ε-differential privacy bounds. However, this added privacy comes at a cost: the noise can degrade model utility, particularly in small data settings or complex tasks where signal-to-noise ratios are low.

Homomorphic encryption allows mathematical operations to be performed directly on encrypted data, producing encrypted results that, when decrypted, match those obtained from the corresponding operations on the raw data. In federated learning, clients can encrypt their model updates before transmission, allowing the server to perform secure aggregation without ever accessing the raw updates. While HE provides strong confidentiality guarantees, it is computationally expensive and often requires specialized cryptographic libraries or hardware support, which may limit scalability in real-world deployments.

Secure multi-party computation (MPC) protocols enable multiple parties to jointly compute a function over their inputs while keeping those inputs

private from one another. Applied to FL, MPC allows for secure aggregation of model updates such that no single party, including the server, can learn any individual client's update. MPC is attractive for its strong security model and compatibility with legal definitions of data non-disclosure. However, it also introduces significant communication overhead and synchronization complexity, making it more suitable for static or small-scale environments.

Trusted execution environments (TEEs) such as Intel SGX, offer a hardware-based approach to privacy by creating isolated regions of memory for secure computation. In FL, TEEs can be used on the server side to securely aggregate updates in an encrypted enclave. While TEEs reduce the need for complex cryptographic protocols, their security depends on hardware integrity and resistance to side-channel attacks. In addition, the entire solution is fully hardware dependent.

Hybrid approaches are often employed to balance security, utility, and efficiency, many recent FL frameworks adopt hybrid strategies that combine multiple privacy-preserving techniques. For example, systems may use DP at the client level and MPC or HE at the server aggregation stage. These combinations aim to provide end-to-end security guarantees while mitigating the performance trade-offs associated with any single technique.

Overall, while FL enhances privacy preservation, it does not eliminate the risks entirely. Proposed solutions come at a cost in terms of performance or efficiency. Therefore, this field is widely studied as there are still significant improvements to be made to further facilitate adoption.

For the purposes of this thesis, the implicit threat model adopted across all chapters is the honest-but-curious server: the central orchestrating server follows the prescribed FL protocol faithfully but may attempt to infer sensitive information from the model updates it receives. Clients are assumed to be honest participants who do not intentionally corrupt their updates; Byzantine attacks from malicious clients are out of scope. This is a reasonable assumption in the cross-silo healthcare setting where participating institutions have reputational and contractual incentives to cooperate genuinely. Under this threat model, the primary privacy risks are reconstruction attacks (recovering training data from shared gradients) and membership inference attacks (determining whether a specific record was used in training). The methods proposed in subsequent chapters do not incorporate formal differential privacy guarantees, but are privacy-enhancing in practice.

Beyond privacy, FL comes with a distinct strategic advantage: it allows participating institutions to retain full control over their data. Unlike centralized data pooling, where an institution concedes data to a third party, FL ensures

that each participant can unilaterally withdraw from the federation at any time without needing to request data deletion from a central repository. This property is particularly relevant in healthcare consortia where institutions may be competitors as well as collaborators. The distinction between privacy (protection against inference) and control (sovereignty over participation) is important and complementary; both motivate the adoption of FL in the healthcare domain.

1.8 Challenges in federated learning

As indicated, FL introduces a decentralized paradigm for training machine learning models, offering privacy benefits. However, this shift in architecture brings several unique challenges that impact the scalability, efficiency, and robustness of FL systems. This section explores key technical and practical challenges in FL.

1.8.1 Data heterogeneity

Among the central challenges in FL, data heterogeneity, or statistical heterogeneity, is a prominent challenge. Unlike traditional centralized machine learning where data can be pooled and shuffled to approximate an independent and identically distributed (IID) setting, FL must operate under the constraint that each client's data remains local, private, and potentially very different from that of others.

In practice, clients in FL, e.g., mobile devices, hospitals, or research centers, collect data from unique contexts, environments, or populations. This leads to non-IID data distributions, which introduce a mismatch between the aggregated global model and the data distributions observed locally by clients. This mismatch is present in two forms:

- **Label distribution skew:** Clients may not observe all classes uniformly. For example, a pediatric clinic may rarely see elderly patients, while a cardiology center may see a high frequency of heart-related diagnoses.
- **Feature distribution skew:** Input features such as imaging resolution, sensor types, or patient demographics may differ across institutions.

In healthcare, this issue is amplified by population stratification. A hospital in a metropolitan area might serve a diverse ethnic population with varied disease

presentations, whereas a rural clinic might serve a more homogeneous group. Furthermore, differences in clinical practices, diagnostic equipment, and data annotation standards further enforce data inconsistency.

Heterogeneous data leads to what is commonly referred to as client drift, a phenomenon where local models trained on divergent data yield updates that are inconsistent or even contradictory. When aggregated, these conflicting updates can cause convergence to suboptimal global models. Moreover, standard optimization algorithms like FedAvg are not robust to such variability, resulting in slow or unstable training progress.

To address these issues, a wide range of algorithmic strategies have been proposed, focusing on constraining local divergence, improving aggregation robustness, or rethinking the notion of a global model altogether.

FedProx [93] modifies the local objective by adding a proximal term that penalizes deviation from the current global model:

$$\min_w f_i(w) + \frac{\mu}{2}\|w - w_g\|^2 \tag{1.1}$$

Here, $f_i(w)$ is the local objective function for client i, w_g is the global model, and μ is a tunable parameter controlling the regularization strength. This term effectively reduces client drift by tethering local updates closer to the global model, particularly beneficial when client datasets are small or imbalanced.

FedNova [112] addresses heterogeneity by normalizing client updates based on the number of local steps and data volume. It mitigates the dominance of clients with more compute or larger datasets and prevents aggregation bias that could arise from uneven training lengths. FedNova ensures update alignment in terms of contribution scale and is particularly effective in cross-silo healthcare collaborations where institutional capacities vary.

Rather than enforcing a single global model, clustered FL approaches such as FedCluster or FedGroup dynamically partition clients into subsets based on similarities in data distribution or model gradients [94]. Each cluster learns its own global model, allowing for better local adaptation without entirely forgoing generalization. This is especially useful in healthcare networks with diverse institutional profiles, e.g., grouping teaching hospitals, community clinics, and specialized centers separately.

Another promising direction in federated learning is personalized FL [104], which aims to tailor models to the unique data distributions of individual clients while still leveraging the benefits of shared learning. One common approach involves fine-tuning, where each client locally adapts the globally trained model to its specific data after convergence. Alternatively, meta-learning techniques, such

as Per-FedAvg [67], optimize the global model for rapid adaptation during local updates. Another strategy is multi-task learning, where each client's learning objective is treated as a related but distinct task, enabling the system to jointly optimize a shared representation alongside personalized outputs.

Such approaches acknowledge that no single model may optimally serve all participants, especially in highly personalized domains like patient-specific prognosis or treatment response prediction.

While techniques such as FedProx, clustered federated learning, and personalization strategies have proven effective in mitigating the challenges posed by data heterogeneity, they are not without trade-offs. For instance, the inclusion of a proximal term, as in FedProx, can improve stability during aggregation but may also slow down local convergence and necessitate additional hyperparameter tuning. Clustered approaches that group clients with similar data distributions help balance local specialization with global learning, though they require extra computational overhead and coordination to maintain group consistency. Personalization methods can lead to improved performance on non-IID data but may limit the generalizability of the shared model.

These trade-offs are of particular importance in sensitive domains like healthcare, where model consistency, fairness, and clinical reliability are critical. Striking the right balance between personalization and generalization remains an open challenge. Hybrid strategies that integrate personalization within clustered or fairness-aware frameworks are being actively explored to address these concerns. Most existing studies rely on overly simplistic or synthetic non-IID benchmarks, such as class imbalance in MNIST, which do not adequately capture the complexity of real-world medical datasets. There is a growing need for more representative federated healthcare benchmarks and shared datasets to enable robust, reproducible evaluation of these strategies in practical settings.

1.8.2 Communication efficiency

Communication overhead is a bottleneck in FL, and is directly defined by the number of clients present in a certain federation. Each round of training typically involves transmitting model parameters or gradients between clients and a central server. This challenge becomes an even more prominent issue in resource-constrained environments, where upload and download speeds are limited and latency is unpredictable.

In traditional FL settings, all selected clients are required to send their model updates at the end of each local training epoch. These updates, often consisting of millions of 32-bit floating-point parameters in deep learning models, must

be transmitted multiple times over the course of training. This can quickly become infeasible in real-world settings, particularly for high-resolution medical imaging models or sequence-based clinical prediction tasks.

To address this, several communication-efficient techniques have been developed. A key approach involves reducing the size of each shared update without reducing model accuracy. We distinguish two main approaches:

- **Quantization:** This method reduces the number of bits used to represent each weight or gradient value. For instance, 32-bit floating-point numbers can be quantized to 8-bit or even binary representations. Popular methods include stochastic quantization, ternary encoding, and dynamic precision adjustment [36]. These reduce the size of transmitted updates and can be combined with entropy coding for further gains.
- **Sparsification:** Rather than transmitting the full gradient or weight update vector, clients transmit only the top-k largest-magnitude elements, and set the rest to zero. The server then accumulates the sparse updates over time. Accumulated sparsity ensures that the dropped information is not permanently lost.

One of the most effective frameworks in this space is Deep Gradient Compression (DGC) [61], which combines multiple techniques, such as gradient sparsification, momentum correction, local gradient clipping, and delayed updates, to achieve up to a 200-fold reduction in communication cost, while maintaining comparable model performance to uncompressed baselines.

Another important dimension of communication efficiency lies in reducing the frequency of communication. In FedAvg, a tunable parameter E determines how many local epochs each client trains before sending an update. Increasing E reduces the number of communication rounds needed but may lead to local overfitting or divergence in non-IID settings.

This trade-off between computation and communication can be optimized depending on the device's capacity. For example, edge nodes in hospitals with high computational resources but low network throughput can be configured to train longer locally. More advanced strategies like adaptive local update scheduling and loss-aware local training dynamically adjust the value of E based on convergence metrics or update stability.

Synchronous training, where the server waits for all selected clients to return their updates, can be highly inefficient when some clients are slower (stragglers). Asynchronous FL protocols [54, 118] allow the server to incorporate updates as

they arrive, with staleness-aware weighting. This not only reduces idle time but improves tolerance to client dropout or unstable connections.

Such protocols are particularly useful in clinical settings where different institutions may experience different loads or network conditions. For example, a regional hospital may return updates within minutes, while a small clinic with limited bandwidth may take hours or intermittently disconnect.

In practical FL deployments, a hybrid approach to communication optimization is often adopted. For instance, an FL pipeline might begin training with high-frequency updates to ensure quick model stabilization and later shift to more infrequent and compressed updates as the model converges. This phased strategy aligns with operational constraints in medical environments.

Hospitals participating in federated imaging projects, such as those using NVIDIA Clara or Intel OpenFL, frequently use such hybrid communication strategies. Some deployments even layer communication optimization with privacy-preserving mechanisms such as secure aggregation or differential privacy, further complicating the bandwidth equation but ensuring compliance with data protection standards.

It is also worth considering the time scale of model training in the healthcare context. Unlike consumer applications where models may need near-real-time updates, a model for ICU Length-of-Stay prediction may only need retraining quarterly or biannually as patient populations and clinical protocols evolve. This relaxed latency requirement means that the primary efficiency gains from the methods proposed in this thesis, client recruitment and adaptive aggregation, lie not only in faster convergence but also in reduced organizational burden: fewer participating institutions means fewer coordination points, fewer institutional review board approvals for each training cycle, and a simpler operational footprint. Communication efficiency therefore has both a technical dimension (fewer bytes transmitted) and an operational dimension (fewer institutional handshakes required).

1.8.3 Client selection and participation

Client selection in FL is a component that significantly affects model convergence, computational efficiency, and fairness. Unlike centralized settings where all data is readily accessible, FL requires choosing a subset of clients to participate in each communication round. While random sampling is the default strategy due to its simplicity and probabilistic fairness, it does not account for the varying availability, quality, or informativeness of clients' data.

In real-world deployments, particularly in mobile or resource-constrained environments, many clients may be unavailable during scheduled training rounds (i.e., stragglers). This can be due to battery constraints, low network bandwidth, local policy restrictions, or maintenance schedules. In healthcare settings, availability may depend on hospital IT infrastructure, data governance approvals, or institutional workloads. Consequently, random sampling often results in dropped rounds or stale updates, reducing training efficiency and slowing convergence.

To mitigate this, participation-aware sampling has been proposed, where clients are selected based not only on randomness but also on their recent availability history, expected uptime, or system health scores [72].

Beyond availability, the utility of client data plays a role in effective client selection for federated learning. In practice, not all client updates are equally valuable to the global model. Active client selection methods seek to maximize the impact of each training round by choosing clients whose contributions are expected to yield the highest information gain or reduce model uncertainty. These strategies assess the informativeness of local updates and prioritize clients accordingly, especially in scenarios where communication or computation budgets are limited.

One common approach involves uncertainty sampling, where clients are selected based on the level of uncertainty their local models exhibit when predicting on a shared validation set. Clients producing predictions with higher entropy or lower confidence are deemed more informative, as updating from these distributions can improve global model calibration. Another approach emphasizes gradient diversity by favoring clients whose model updates deviate most from the current global gradient direction [111]. By selecting these divergent clients, the algorithm encourages exploration of underrepresented data directions and avoids overfitting to more homogeneous client distributions, ultimately supporting broader generalization across the federation.

Unequal client participation presents a significant challenge to fairness in federated learning, particularly in healthcare networks where disparities in institutional resources exist. Larger hospitals with robust computational infrastructure and stable connectivity are more likely to participate consistently in federated rounds, while smaller or rural clinics may be intermittently involved or entirely excluded due to technical constraints. This imbalance in participation can lead to models that are optimized for dominant data distributions, thereby failing to generalize to underrepresented patient populations.

To mitigate this issue, fairness-aware approaches to federated learning have been proposed. One strategy involves stratified client sampling, where the selection

process is guided by metadata such as geographic location, institutional size, or patient demographics to ensure balanced representation across subgroups. Another method applies weighted aggregation, increasing the influence of updates from clients that serve minority or low-resource populations, thus amplifying their contribution to the global model [58].

These strategies are particularly relevant in clinical scenarios involving health equity, such as predicting outcomes for minority ethnic groups, rural patients, or those with rare diseases

An assumption underlying several of the methods in this thesis is that of irregular client participation, where not all institutions are available in every training round. It has been noted that healthcare applications tend to run in fairly stable environments, which may seem at odds with this assumption. However, irregular participation is realistic in several healthcare-adjacent scenarios: (i) multi-national federations spanning different regulatory jurisdictions, where periodic legal reviews or policy changes may temporarily prohibit a site from participating; (ii) resource-constrained settings where computational infrastructure is shared with clinical systems, creating intermittent availability; (iii) longitudinal studies where new clinical sites join or existing ones leave over the course of months or years; and (iv) outside the hospital context, mobile health applications, wearable devices, and edge computing in telemedicine all exhibit highly variable device availability. The methods in this thesis are designed to be robust to these scenarios while also functioning well in the more stable cross-silo setting that characterizes most hospital-to-hospital federations.

1.8.4 Security threats

While FL reduces the risks associated with centralized data aggregation, it is still susceptible to privacy breaches and adversarial threats. Because model updates or gradients are shared with a central server or coordinator, sensitive information can still be inferred indirectly.

Several types of attacks have been identified in the literature that compromise either the privacy of client data or the integrity of the global model:

- **Membership Inference Attacks:** These attacks attempt to determine whether a specific data point was part of a client's local training dataset. This is especially concerning in healthcare where confirming a patient's presence in a dataset could reveal sensitive diagnoses or treatments [102].
- **Model Inversion Attacks:** In these attacks, adversaries attempt to reconstruct input features, such as images or clinical attributes, from

shared gradients or model outputs [31]. In FL, the gradients exchanged during training can contain enough signal to partially recover medical images or infer demographic features.

- **Backdoor and Poisoning Attacks:** A malicious client can insert a hidden trigger into the local training data (e.g., a pixel pattern in X-ray images) and assign it a misleading label. If such updates are aggregated, the global model may learn this backdoor, misclassifying any input containing the trigger [3]. This poses safety risks in clinical diagnostics.

For instance, consider a hospital participating in FL for training a model to detect pneumonia in chest X-rays. A malicious agent in this hospital could insert a small, barely noticeable watermark in some training images and label all such cases as "no pneumonia," regardless of actual diagnosis. If this poisoned update dominates the aggregation (due to having a large dataset or weight), the global model will learn this association. Later, an attacker can add the watermark to any new X-ray to force the model to output "healthy," potentially leading to misdiagnosis.

As discussed in section 1.7 on privacy considerations, methods such as Secure Multi-Party Computation and Homomorphic Encryption play a role in enhancing security during federated training. Building on these principles, several additional defense strategies have been developed to mitigate the specific vulnerabilities of federated learning, including inference, inversion, and poisoning attacks.

One widely adopted approach is secure aggregation, which ensures that the central server can only access the aggregated sum of client model updates, without visibility into any individual contribution. Cryptographic schemes such as the protocol introduced by Bonawitz et al. [11] facilitate this by enabling encrypted summation of model parameters. While effective at concealing client-specific data, secure aggregation introduces computational and communication overhead, particularly in large-scale implementations.

Another approach levreages differential privacy to probabilistically obscure the influence of any single data point. For instance, DP-FedAvg [33] applies gradient clipping and adds calibrated Gaussian noise to local updates before transmission. Though DP offers strong formal privacy guarantees, it reduces model accuracy.

A complementary class of techniques focuses on robust aggregation to guard against malicious or anomalous clients. Algorithms like Krum [10], Bulyan [71], and the coordinate-wise median seek to suppress the influence of outliers by comparing updates across clients and filtering suspicious inputs. These methods

are particularly relevant in defending against model poisoning or backdoor attacks. However, their reliance on statistical similarity may result in the exclusion of legitimate but non-conforming updates, which is problematic in heterogeneous healthcare environments.

Collectively, these strategies offer a layered defense model, each addressing a distinct threat surface in federated systems. Yet, they also highlight the fundamental trade-offs between privacy, robustness, and utility.

Each defense mechanism provides a partial solution and often needs to be combined with others to achieve robust protection. Secure aggregation enhances confidentiality but does not protect against poisoning. Differential privacy guards against inference attacks but at the cost of model utility. Robust aggregation helps with integrity but assumes a certain level of client honesty and consistency in data distributions.

Moreover, the integration of defense mechanisms must consider usability and scalability. Protocols that require extensive coordination or cryptographic infrastructure may be impractical in resource-constrained environments. Future research continues to explore lightweight, adaptive security solutions that align with the operational needs of healthcare systems.

1.8.5 Existing FL deployments

While FL is a rapidly evolving field, a number of real-world deployments have demonstrated its practical feasibility and value, particularly in domains like healthcare and pharmaceuticals. These deployments not only validate the theoretical strengths of FL but also shed light on the infrastructure, interoperability, and governance challenges inherent in large-scale federated systems.

The Federated Tumor Segmentation (FeTS) Challenge [4, 77, 87], is one of the most prominent demonstrations of FL in a healthcare context. Coordinated by institutions including the University of Pennsylvania and Intel, the initiative brought together over 20 institutions across the globe to collaboratively train brain tumor segmentation models using multi-parametric MRI data.

FeTS tackles one of the most difficult problems in neuro-oncology, glioblastoma segmentation, without requiring centralized data aggregation. Each institution deployed a secure FL client, trained local deep learning models (3D U-Nets [115]), and communicated encrypted updates to a central aggregation server.

The challenge demonstrated that models trained via FL achieved comparable Dice scores and generalization to those trained on centralized data, while fully

adhering to institutional privacy protocols. It set a benchmark for clinical-grade FL implementations and inspired follow-up projects in segmentation and radiogenomics.

The MELLODDY (Machine Learning Ledger Orchestration for Drug Discovery) project [40] showcased FL feasibility at scale. Coordinated by Owkin and supported by the IMI (Innovative Medicines Initiative), MELLODDY united ten of the largest pharmaceutical companies to train machine learning models on molecular property prediction and compound screening, without exchanging proprietary data.

The resulting federated models outperformed those trained on any individual partner's dataset in tasks such as predicting ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties. Importantly, MELLODDY offered a proof-of-concept that FL can enable meaningful precompetitive collaboration between competitive actors in the pharmaceutical industry.

Despite these challenges, these deployments validate FL's transformative potential for enabling privacy-preserving, cross-institutional AI in both healthcare and life sciences.

1.9 Summary and research directions

FL presents a powerful framework for enabling privacy-enhancing machine learning across decentralized data sources. Throughout this chapter, we have discussed the FL paradigm along with its corresponding pitfalls and benefits, the motivations for its adoption in medical AI, and the trade-offs that arise when aiming to protect privacy without compromising performance or fairness.

Despite its promise, FL continues to face several open challenges that slow down its full-scale adoption in privacy-sensitive domains such as healthcare:

- **Data heterogeneity:** Healthcare data is typically non-IID, leading to divergence in local updates and difficulties during model aggregation.
- **Communication inefficiency:** Realistic FL deployments must account for limited connectivity, especially in low-resource or rural clinical settings.
- **Security threats:** Attacks such as model inversion, membership inference, and gradient leakage expose vulnerabilities in naïve FL protocols.
- **Client variability:** Hospitals and devices differ drastically in compute, bandwidth, and availability, necessitating adaptive and fault-tolerant systems.

- **Model personalization:** A global model may underperform on institutions with rare disease patterns or atypical patient populations.

To address some of these challenges, this thesis contributes four approaches designed to extend the robustness, and deployability of FL in healthcare settings. Each contribution is motivated by one of the several bottlenecks discussed in this section and is validated through empirical studies across (multi-modal) clinical datasets.

Together, these contributions address major limitations in current FL workflows and bridge the gap between theoretical FL constructs and practical deployments in healthcare. The next four chapters elaborate on each contribution in detail, presenting both formal analysis and real-world validation. Together, they form a coherent research trajectory that lays the groundwork for scalable federated learning in healthcare applications.

A fundamental question is whether a single global model is justified when there is significant variation across clients. This thesis operates from the premise that a well-trained global model serves as a strong initialization point from which local fine-tuning can yield high-performing site-specific models. This view is supported empirically in Chapter 3, where excluded clients that receive the global model and perform a single round of local fine-tuning achieve substantially better performance than the global model applied directly. The global model captures shared structure across the federation, while local adaptation addresses site-specific patterns.

Chapter 2

Client Recruitment for Federated Learning in ICU Length of Stay Prediction

This chapter has been published as:

Scheltjens, V., Wamba Momo, L.N., Verbeke, W., De Moor, B. (2023). **Client Recruitment for Federated Learning in ICU Length of Stay Prediction.** *In: Proceedings of 2023 IEEE International Conference on e-Science.* Presented at the 2023 IEEE International Conference on e-Science, Limassol, Cyprus, 09 Oct 2023-13 Oct 2023. ISBN: 979-8-3503-2223-1.

Paper abstract:

Background: Machine and deep learning methods for medical and healthcare applications have shown significant progress and performance improvement in recent years. These methods require vast amounts of training data which are available in the medical sector, albeit decentralized. Medical institutions generate vast amounts of data for which sharing and centralizing remains a challenge as the result of data and privacy regulations. Federated Learning (FL) is well-suited to tackle these challenges. However, FL comes with a new set of open problems related to communication overhead, efficient parameter aggregation, client selection strategies and more.

Methods: In this work, we address the step prior to the initiation of a federated

network for model training, client recruitment. By intelligently recruiting clients, communication overhead and overall cost of training can be reduced without sacrificing predictive performance. Client recruitment aims at pre-excluding potential clients from partaking in the federation based on a set of criteria indicative of their eventual contributions to the federation. In this work, we propose a client recruitment approach using only the output distribution and sample size at the client site.

Results: We show how a subset of clients can be recruited without sacrificing model performance whilst significantly improving computation time.

Conclusions: By applying the recruitment approach to the training of federated models for accurate patient Length of Stay prediction using data from 189 intensive care units (clients), we show how the models trained in federations made up from only recruited clients significantly outperform federated models trained with the standard procedure in terms of predictive power and training time.

2.1 Introduction

Recent machine learning (ML) and deep learning (DL) techniques have proven to be of significant value for healthcare applications [8,13,97]. An abundance of medical data is continuously generated, especially in Intensive Care Units (ICU) where patients are monitored uninterrupted and vitals are charted continuously. DL techniques are particularly good at extracting underlying complex relations in such large datasets, thus the vast amount of available data bodes well for these data hungry models. One of the main limitations in relation to medical data of any sort relates to data availability. As a result of regulatory restrictions both in Europe through the General Data Protection Regulation (GDPR) [28] and more recently the European Union Artificial Intelligence Act, and the United States with the Health Insurance Portability and Accountability Act (HIPAA) [14] imposing similar restrictions, data sharing poses a significant challenge in the medical sector both for primary and secondary use [19]. In both cases, sharing and centrally accumulating privacy-sensitive data is bounded by a legal framework, necessary to protect the privacy of the individuals behind the data. This, however, poses a challenge for research and applications that require large amounts of medical data.

In recent years, the value of federated learning (FL) has been illustrated for medical applications, especially in the field of computer vision for medical image segmentation and classification [59, 89, 101]. Here, the FL approach allows to

learn complex models over decentralized data without direct access to the data. FL, however, comes with a new set of challenges. In the real world, decentralized data is often not independent and identically distributed *(non-IID)* which harms the training procedure and model performance [129]. The reduced performance can mainly be attributed to the weight divergence in local models as a result of the *non-IID* data [129].

Selection procedures of federated training algorithms may sample those contributors whose data is most informative at training time. This, however, requires all potential contributors to form part of the federation and participate in at least one round of training. In the work by Yichen Ruan et al. [92], the theoretical foundations are provided for a client recruitment process which precedes the initiation of the federation. This aims at building the federation with those clients for which it can, a priori, be said that their contributions to the federation will be valuable. More specifically, clients for which the local data does not form a good representation of the population are considered less representative and therefore, of less value to the federation. Intuitively, when aiming for accurate global predictive performance, less representative clients result in higher weight divergence in the local models at training time [129], which harms the training procedure and predictive power of the resulting global model.

Whilst in [92], a sound optimization framework is provided for client recruitment, in this work, by building on [92], we define a client recruitment approach and the construct of client-level representativeness using only the local target distribution and sample size. Subsequently, the practical relevance of this approach is demonstrated through an application on the real-world eICU dataset [34, 78, 79] for which only 54 out of the 189 potential clients are recruited. With the recruited client subset making up the federation, we are able to learn better performing models, in terms of predictive power and training time, to predict patient Length of Stay (LoS) in ICU compared to the standard FL approach as proposed in [67].

The remainder of this work is structured as follows. Section 2.2 outlines the related work. In Section 2.3 the basics of FL are discussed. In Section 4.3 the client recruitment approach is discussed along with the experimental setup and corresponding description of the data. Section 2.5 outlines the results and provides a discussion of the main findings. In Section 3.5 concluding remarks are presented followed by a discussion of the current limitations and directions for future work.

2.2 Related work

Federated learning was originally proposed by McMahan et al. in 2017 [67] in conjunction with the *FedAvg* algorithm and has since been proven to be a valuable learning framework that can yield accurate models without direct access to the local data.

A significant amount of research efforts have been dedicated to the client selection problem. Client selection in FL deals with client scheduling for each round of training, i.e., for each training round a subset of clients is selected that will contribute in the next training iteration. The standard approach is to randomly select this subset. However, as argued in multiple studies [73,117,121], a better approach is to impose criteria based upon which client selection can be performed. These criteria often relate to how informative updates from certain clients are, or the local computational resources [123,125]. This, however, does not allow for noninformative clients to be pre-excluded from the federation, which is where client recruitment comes in. Client recruitment aims to discard potential clients from the set of available clients for the federation, prior to initiation of the latter. One of the ways to do so is by considering limited statistics on the local dataset that do not contain privacy sensitive information. By pre-excluding potential clients, the foremost benefit is that the overall cost of training a model in the federated setting is reduced [92] without sacrificing predictive performance. In addition, optimally, the least informative clients are pre-excluded which results in increased predictive performance for the resulting global model.

Although FL originates in the large-scale distributed edge computing setting, it has been extensively studied in healthcare [59, 69, 76, 89]. Specifically the work discussed in [69] closely relates to the work proposed here. In [69] the authors extensively assess different parameter settings for federated training on medical ICU data to identify suitable parameters for In Hospital Mortality (IHM) binary classification. In this study, we tackle a different problem, i.e., ICU LoS prediction, using the same eICU dataset. The authors from [69] show that a larger number of local training epochs improves performance whilst reducing the training cost at once as a result of the reduced server-client communication. In this work we extend upon these previous works by proposing a client recruitment scheme using limited statistics on the local data, applied to the critical care setting.

2.3 Federated learning basics

Federated Learning (FL) [67] is a technique that allows for a central model to be trained over distributed data that originate and are stored locally. This approach was originally proposed within the setting of large-scale distributed learning on mobile devices and has been widely used and researched in the medical sector given the privacy enhancing aspect as well as the regulatory restrictions on data sharing within the sector. Each local data source, e.g., a mobile phone or a server hosted at a hospital, is referred to as a *client*. The group of clients that contribute to training a central model, including the central *server* is referred to as the *federation*. The central server orchestrates the learning process over the different clients following the predefined FL algorithm. In that sense, the central server is holistically responsible for (i) initiating the model; (ii) providing a copy of the model to all of the clients; (iii) aggregating over the received model parameters and (iv) send back the updated model to the clients. Each client locally trains the model for a predetermined number of epochs after which only the model parameters are sent back to the central server. In the standard *FedAvg* algorithm [67] considered in this work, local model parameters are averaged into the global model update. In addition, for each communication round, either all or a subset of the clients in the federation are randomly selected for training. We consider this the standard FL setting which comes with most out of the box implementations.

In this work, the clients correspond to 189 hospitals where the main challenge relates to some extent to the concept of *client shift* in conjunction with the understanding that in the real world, data from different hospitals is likely to be *non-IID* distributed. In Fig. 3.2, 4 of the in total 189 local target distributions are shown, illustrating how even when only looking at the target, the data hosted at each of the clients can vastly differ from one client to the next. More specifically, clients host data that originates from different hospitals in different geographic regions, each with potentially different demographic characteristics, etc.. Depending on local sample size and training parameters, models may overfit to the local data and bias the global model when aggregating model parameters. These are common problems to FL training algorithms and have been tackled by introducing more involved aggregation algorithms such as *Weighted FedAvg* or smart client selection [73, 117, 121].

Whilst FL allows to learn a central model from distributed medical data in a privacy enhancing manner, not all data at each of the hospitals are equally valuable. More specifically, smaller local sample sizes are less likely to represent the global distribution which will typically result in a larger empirical training loss, as discussed in [92]. In addition, as a result of different local demographics, even larger local datasets could also diverge in distribution from the population,

which will again reduce predictive performance. Existing client selection and aggregation algorithms provide partial solutions to these issues. These methods, however, require for all clients to form part of the federation and for each client to have participated at least once in a training round. If not, the algorithms do not obtain the required information to guide the further aggregation and selection procedure.

This work builds on the client recruitment work introduced in [92] and the field of FL in healthcare to develop a method for client recruitment by looking only at the local distribution of the target variable and the local sample size. With this, the recruitment process aims to recruit clients for which the local data better represents the population before initiating the federation. This would intuitively lead to better performing models globally and reduced training time.

2.4 Methods

To assess the performance and utility of federated models that are trained with a subset of clients recruited following the approach that will be described in 2.4.3, a single prediction task is defined for both the central, and federated models. The task is for each model to predict the patient LoS in ICU, similar to what has been studied in [1, 90, 91, 107] where LoS is defined as the remaining time in ICU for a given patient. Formally, the task is to yield predictions $\hat{y}$ which approximate y, the true LoS.

2.4.1 Data

For the evaluation of the proposed approach on real-world, multi-center data, the eICU dataset [34, 78, 79] is used, covering over 200,000 eICU admissions to 208 US hospitals. The total admission count covers over 139,000 unique patients registered at one of the US hospitals between 2014 and 2015.

The dataset is made publicly available for research purposes and is particularly interesting for FL as it allows for the data to be mapped to the originating institution. The data is preprocessed in concordance with the preprocessing pipeline proposed in [91]. In summary, the first 24 hours of data post ICU admission, for adult patients are extracted and used to predict LoS. Both temporal and static information are extracted, cleaned, re-sampled, imputed and one-hot encoded. The temporal data is fused with the static patient data which will serve as the input to the models. For patients with multiple recorded stays over the designated time period, only one of the stays is considered to

avoid information leakage when obtaining train, test and validation splits. We refer to the work by Rocheteau et al. [91] for an extensive description of the preprocessing pipeline.

The resulting cohort is comprised of data pertaining to 89,127 patient stays over 189 hospitals, covering a total of 35 features of which 17 are temporal and 18 static. The summary statistics for the resulting data cohort are included in Table 3.2 and a detailed overview is provided in Appendix A.1.

Table 2.1: The extracted and preprocessed data cohort showing the train, validation and test splits used throughout as well as the descriptive statistics, mean and median Length of Stay (LoS), number of hospitals and features in the cohort.

Number of patient stays	89,127
Train	62,375
Validation	13,376
Test	13,376
Mean LoS	3.69
Median LoS	2.27
Number of features	35
Temporal	17
Demographic (static)	18
Number of hospitals (clients)	189

2.4.2 Model architecture

For both the centralized and federated training tasks in this work, the Gated Recurrent Unit (GRU) [17], and Long Short-Term Memory (LSTM) [41] networks are used. Both GRU and LSTM belong to the class of Recurrent Neural Networks (RNN) which are widely used when dealing with sequential data. The GRU architecture is comprised of two sole gates. The reset and update gates, respectively denoted as r_t and z_t in (3.5), resulting in reduced computational complexity. This, in turn, is a desirable characteristic for FL applications where local computational resources and communication overhead pose real challenges [55].

In (3.5), the governing equations for the GRU cell are shown. These outline the computations that occur for every discrete time step t in the input sequence.

$$
\begin{aligned}
r_t &= \sigma(W_r \cdot [h_{t-1}, x_t]) \\
z_t &= \sigma(W_z \cdot [h_{t-1}, x_t]) \\
\tilde{h}_t &= \tanh(W \cdot [r_t * h_{t-1}, x_t]) \\
h_t &= (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t
\end{aligned}
\tag{2.1}
$$

Similarly, in (3.6), the governing equations for the LSTM cell are shown. LSTM is comprised of three gates rather than the two gates present in GRU which are denoted as f_t, i_t and o_t, representing the forget gate, input gate and output gate.

$$
\begin{aligned}
f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\
i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\
\tilde{C}_t &= \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \\
o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\
C_t &= f_t * C_{t-1} + i_t * \tilde{C}_t \\
h_t &= o_t * \tanh(C_t)
\end{aligned}
\tag{2.2}
$$

In both (3.5) and (3.6) σ represents a sigmoid neural network layer whereas tanh represents a tanh neural network layer with the exception of the last step for LSTM where h_t is calculated. Here, tanh corresponds to the tanh activation function such that the values are transformed to fall in $[-1, 1]$. A visual representation of the GRU and LSTM cells is shown in Fig. 2.3.

For both architectures, x_t represents the input at time step t, which corresponds to one hour in this work. In addition, h_t, h_{t-1} denote the hidden states at time $t, t-1$ respectively and $*$ represents the Hadamard product.

The hidden state h_t, i.e., the output of the cell, is provided as the input of a nonlinear Fully Connected Network (FCN), which yields a single output value representing the predicted LoS. The nonlinearity stems from the ReLU activation function leveraged in the FCN as shown in (3.7).

$$
\hat{y}_t = ReLU\left(W_{y_t} h_t + b_{y_t}\right), \tag{2.3}
$$

with $\hat{y}_t$ the predicted value for LoS at time t. Employing $ReLU(x) = max(0, x)$ forces the outcome to be strictly positive. It is impossible for a patient to have a negative LoS, therefore we restrict the model to only yield predictions in the positive domain.

2.4.3 Client recruitment

Consistent with the work discussed in [92], we consider client recruitment to be a mechanism which is to be invoked prior to establishing a federation for training. To this extent, consider the set S of c potential clients with each a local dataset $D_c = \{(x_i, y_i)\}_i$ where x_i denotes the input data and y_i the target.

To facilitate client recruitment, we let each potential client in S report a tuple (P_{co}, n_c) to the central server where P_{co} denotes the local distribution of the target and n_c the local sample size. From here, the global sample size n_g and output distribution P_{go} can be calculated as shown in (3.1).

$$n_g = \sum_c n_c, \quad P_{go} = \sum_c P_{co} \tag{2.4}$$

Using the tuple reported by the potential client, the local representativeness of the client data ν_c in relation to the global dataset is calculated as a function of the output distribution divergence and the local sample size:

$$\nu_c = \gamma_{dv} \underbrace{\left| \frac{P_{go}}{n_g} - \frac{P_{co}}{n_c} \right|}_{\beta} + \gamma_{sa} n_c^{-0.5}, \tag{2.5}$$

where γ_{dv} and γ_{sa} denote weight parameters that influence the importance of the divergence of the output distribution and local sample size respectively. Furthermore, $\tilde{P}_c - P_c$ converges to $\mathcal{N}(\mu, \sigma^2)$ with $\mu = 0$ at the rate of $O(n_c^{-0.5})$, with $\tilde{P}_c$ the empirical local distribution [42, 98]. From this follows that as n_c grows larger, $\tilde{P}_c$ better approximates P_c [92]. Which is a desired characteristic for client recruitment. By inclusion of the term $n_c^{-0.5}$ in (3.2), clients with larger local sample sizes are favored over those with smaller sample sizes.

The local distribution divergence in terms of the target denoted as β in (3.2) is calculated as the difference between the normalized class counts locally and globally. In this work, the target corresponds to the patient LoS in fractional days. To facilitate the computation of the distribution divergence, ten bins are constructed with each bin corresponding to the frequency of target values

(LoS in fractional days) within the range for a given bin. The bins are defined as: $[(0,1),[1,2),[2,3),...,[7,8),[8,14),[14,+\infty)]$. This formulation converts the target from continuous to discrete classes for which the class counts are used to compute β in (3.2).

To select clients for recruitment, the per client representativeness values from (3.2) are sorted and stored in the vector $\boldsymbol{\nu}$. Consider:

$$\nu_g = \sum_c \nu_c, \tag{2.6}$$

where ν_g represents the global representativeness, used to define the recruitment threshold $\iota = \gamma_{th}\nu_g$ with γ_{th} a configurable hyperparameter. Next, by summing over $\boldsymbol{\nu}$, the value ν_c at which the threshold ι is crossed, is identified. All the corresponding clients for values up until that point in $\boldsymbol{\nu}$ are recruited for the federation. This yields a subset of clients which are the most representative in terms of the target distribution divergence and local sample size.

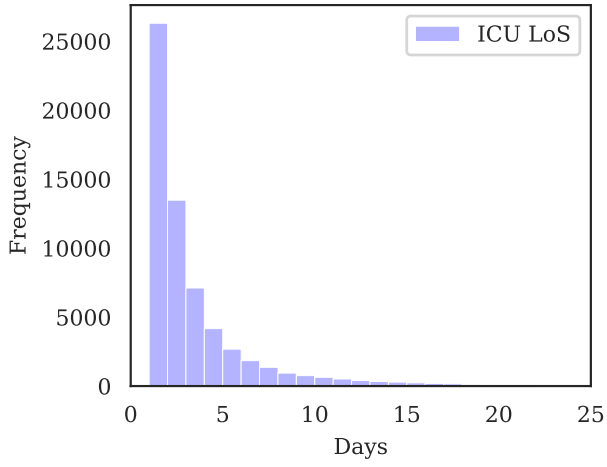


Figure 2.1: The distribution of the target (LoS) in the global training data with a cutoff at 25 days on the x-axis. The y-axis represents the number of patients (frequency) with their corresponding LoS in days as indicated on the x-axis.

Intuitively, when applied to the LoS setting, the central server obtains a tuple (P_{co}, n_c) from each of the potential clients to compute n_g and P_{go} which is visually represented as the global target distribution in Fig. 3.3. This information is then used by the central server to compute for each of the potential clients the local representativeness ν_c as a weighted function in terms

of divergence from the global target distribution and local sample size as shown in (3.2). A small subset of the local target distributions is shown in Fig. 3.2. Visually it is clear how some of the local target distributions diverge to a greater extent from the global target distribution, shown in 3.3, compared to others. In addition, client recruitment requires only a single calculation for each of the clients prior to forming the federation and therefore does not consume any further resources during the remainder of the training procedure.

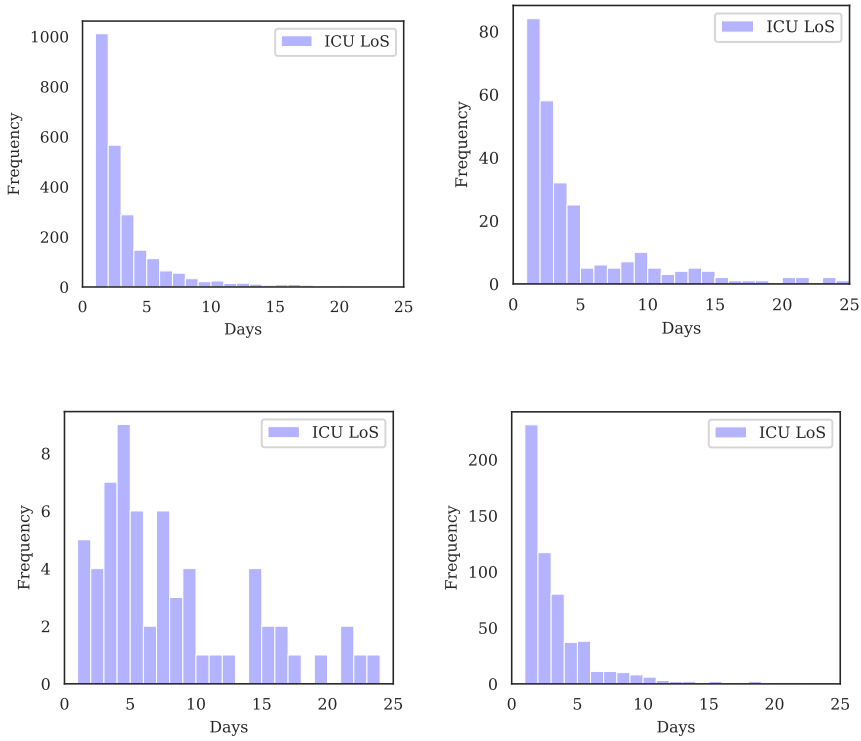


Figure 2.2: Subset of the, in total 189, local target distributions in terms of Length of Stay in days and the corresponding patient counts. The distributions shown correspond to, from left to right, the hospitals with the identifiers; 24, 74, 100 and 143.

2.4.4 Experimental settings

At training time, all training procedures use *AdamW* [63] for optimization and the Mean Squared Logarithmic Error (MSLE) as the loss function, calculated as:

$$MSLE = \frac{1}{n}\sum_{i=1}^{n}(\log(y_i + 1) - \log(\hat{y}_i + 1))^2, \tag{2.7}$$

with y_i the true target value and $\hat{y}_i$ the predicted value. Furthermore, the model hyperparameters are fixed over all training iterations, both central and federated. The exact settings are shown in Table 3.4 with with L the number of layers, N the hidden dimension for each of the layers, η the learning rate, m the batch size, wd the weight decay for the *AdamW* optimizer and r the dropout.

Table 2.2: Model hyperparameter settings used for both central and federated training.

Model	L	N	η	m	wd	r
GRU	2	32	0.005	128	0.005	0.05
LSTM	2	32	0.005	128	0.005	0.05

For evaluation of the performance, all resulting models are evaluated against the hold-out test set containing data from all 189 hospitals. In addition to the MSLE, models are evaluated using the Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and the Mean Squared Error (MSE), shown in (3.9). As an indication of the time complexity, the training time is reported, denoted as τ, in seconds.

$$\begin{aligned} MAE &= \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \\ MAPE &= \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \\ MSE &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \end{aligned} \tag{2.8}$$

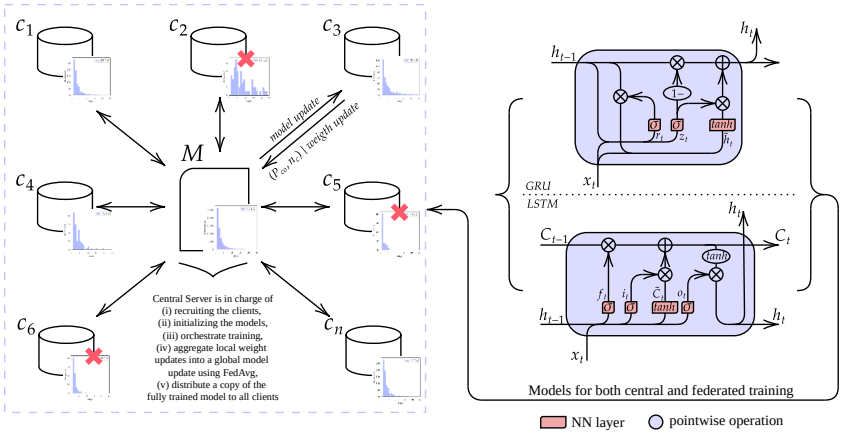


Figure 2.3: General overview of the client recruitment and federated training procedure with M the central server and $c_1, ..., c_n$ the potential clients. Each of the clients is depicted with their corresponding target distributions for which the red cross indicates clients that are discarded to participate in the federation based on the recruitment procedure. M is initially in charge of recruiting clients based on the representativeness of the locally hosted data and will subsequently orchestrate the federated training procedure for either the GRU or LSTM models. The GRU and LSTM cells included are a visual representation of the governing equations in (3.5) and (3.6).

Central training

Central training is performed consistent with the traditional DL procedure in which all data is assumed to be centrally available. The architecture presented in 2.4.2 is trained for a predetermined amount of 15 epochs using the global train and validation sets, which consist of the accumulated data over all potential clients, i.e., the 189 originating hospitals. The resulting central model is subsequently evaluated against the hold-out test set.

Federated training

The FL procedure with and without client recruitment is simulated as a single process using the *FedML* framework as proposed in [38]. In this work, for FL without client recruitment, clients are either all considered in each round

of training or a subset is randomly sampled consistent with the standard client selection implementation described in *FedAvg* [67]. For FL with client recruitment, the client recruitment process described in 2.4.3 is invoked prior to initiating the federation. Following the described implementation, the recruitment of clients is influenced by three user-defined hyperparameters γ_{dv}, γ_{sa} and γ_{th} which respectively define the importance of the divergence in the target distribution, the local sample size and percentage of the global representativeness to be covered by the recruited clients. In Fig. 2.3, a general overview of the federated training procedure with client recruitment is shown with M representing the central server that initially recruits clients for the federation after which the federated training procedure is started for training either the GRU or LSTM network.

Four different federated strategies are implemented and trained with either the GRU or LSTM model. The four approaches differ in terms of the number of clients that partake in the federation, denoted as ϵ, the percentage of clients in the federation that contribute to each training round, denoted as δ, and whether the federation is comprised of recruited clients, or all clients. For each of the federated models, each client trains for four epochs per round of server-client communication for a total of 15 rounds. The resulting model is subsequently evaluated against the hold-out test set.

The specifications for the four different FL strategies are; (i) **Fed-AC:** all clients make up the federation and partake in each training round, (ii) **Fed-SC:** all clients make up the federation, 10% of which are randomly sampled to partake in each training round, (iii) **Fed-ARC:** recruited clients make up the federation and partake in each training round and (iv) **Fed-SRC:** recruited clients make up the federation, 10% of which are randomly sampled to partake in each training round. The settings specific to each of the strategies in the federated setting are summarized in Table 3.8.

2.5 Results and discussion

The results obtained for both the central and federated training procedures as described in 2.4.4 and 2.4.4 are reported in Table 2.4 with τ the training time in seconds and the best performance per metric listed in bold.

Table 2.3: Settings specific to the federated strategies with ϵ the total number of clients in the federation, δ the number of clients randomly sampled from ϵ in each round of training and (γ_{dv}, γ_{sa}, γ_{th}) the hyperparameters for client recruitment

Strategy	ϵ	δ	γ_{dv}	γ_{sa}	γ_{th}
Fed-AC	189	189	-	-	-
Fed-SC	189	19	-	-	-
Fed-ARC	54	54	0.5	0.5	0.1
Fed-SRC	54	5	0.5	0.5	0.1

2.5.1 Client recruitment results

When considering the results reported in Table 2.4, the best performance for MSLE amongst the federated models is obtained by Fed-AC for both GRU and LSTM. Here, all clients make up the federation and partake in each round of training. However, considering all clients at each round of training incurs significant overhead, as illustrated by the high training time for Fed-AC. The second best performing model in terms of MSLE is Fed-SRC, trained with recruited clients, for which a subset was randomly sampled at each round of training. More specifically, as seen in Table 3.8, 54 clients were recruited from which 5 were randomly sampled at each round of training. For both GRU and LSTM, Fed-SRC outperforms Fed-SC and Fed-ARC in terms of MAE, MSE, MSLE and training time. For GRU, Fed-SRC is the best overall performer in terms of MAE and training time, whereas the best MAE for LSTM is obtained by Fed-AC at 2.21 days, nevertheless, Fed-SRC yields an MAE of 2.22 days at a fraction of the training time.

The improved training time is a direct result of the reduced number of clients recruited for the federation. The improved performance can be attributed to the fact that model training is subject to less noise in the data, again as a direct result of the client recruitment procedure. Therefore, more informative model updates are produced at each round of training, resulting in lower empirical loss, which in turn improves performance.

In summary, the client recruitment approach allows for models to be trained that outperform the standard FL approach (Fed-SC) and perform on par or better than the centrally trained model depending on the metric of interest. In addition, the total training time is drastically reduced compared to either central or federated models without client recruitment. In this use case, models with a lower MAE in combination with lower training time are of most value

Table 2.4: Model performance for central and federated models with and without client recruitment. Statistical significance among the federated models in comparison to Fed-SC is indicated as * at the 5% significance level and ** at the 1% significance level.

Model	Strategy	MAE	MAPE	MSE	MSLE	τ(s)
GRU	Central	2.21 ± 0.02	0.57 ± 0.06	21.94 ± 0.63	0.33 ± 0.01	2128 ± 18
	Fed-AC	2.26 ± 0.06	0.63 ± 0.08**	**21.61 ± 0.73****	**0.33 ± 0.02****	5231 ± 29**
	Fed-SC	2.26 ± 0.06	0.46 ± 0.06	23.98 ± 1.26	0.41 ± 0.05	1469 ± 35
	Fed-ARC	2.27 ± 0.12	0.57 ± 0.17*	22.67 ± 1.83**	0.37 ± 0.05*	3359 ± 25**
	Fed-SRC	**2.21 ± 0.03****	**0.46 ± 0.03**	23.49 ± 0.73	0.37 ± 0.03*	**965 ± 24****
LSTM	Central	2.20 ± 0.01	0.52 ± 0.04	22.45 ± 0.45	0.34 ± 0.01	1891 ± 12
	Fed-AC	**2.21 ± 0.04****	0.48 ± 0.06	**23.18 ± 0.52****	**0.37 ± 0.02****	4663 ± 46**
	Fed-SC	2.28 ± 0.08	0.44 ± 0.03	24.36 ± 1.13	0.43 ± 0.06	1301 ± 20
	Fed-ARC	2.23 ± 0.02	**0.43 ± 0.01**	24.05 ± 0.48	0.40 ± 0.02	2866 ± 24**
	Fed-SRC	2.22 ± 0.03**	0.44 ± 0.02	23.96 ± 0.85	0.39 ± 0.03*	**864 ± 13****

to the ICU. For both GRU and LSTM, Fed-SRC outperforms the standard Fed-SC approach, obtaining a better MAE in a fraction of the required training time. Thus, illustrating practical relevance of the client recruitment procedure for larger federations in a real-world, privacy-sensitive, setting.

2.5.2 Recruitment parameter effects

To gain additional insight in the behaviour of the client recruitment procedure proposed in 2.4.3, we assess performance under different settings for the user defined hyperparameters γ_{dv} and γ_{sa}. γ_{dv} influences the importance of the divergence in the distribution between the local target distribution and that of the target in the global data whereas γ_{sa} affects the importance of the local sample size in the client recruitment approach.

We perform this analysis using the GRU model and tune the parameters such that distribution divergence is prioritized over the local sample size and vice versa as follows: **Fed-SRC-QG**: divergence over sample size, and **Fed-SRC-DG**: sample size over divergence. The parameter settings for both the approaches are listed in Table 2.5.

Table 2.5: Parameter settings for the quality greedy (QG) and quantity greedy (DG) approaches.

Strategy	γ_{dv}	γ_{sa}
Fed-SRC-QG	1	0.01
Fed-SRC-DG	0.01	1

The above formulation is intuitively equivalent to the recruitment process being quality greedy (QG) in the former and quantity (data) greedy (DG) in the latter. The quality greedy recruitment strategy allows for clients with smaller sample sizes for which the output does not diverge significantly in distribution to be recruited. The contrary is true for the quantity greedy strategy in which the recruitment process neglects, to some extent, the distribution divergence in the output and favorably ranks clients with large local samples. Essentially, this approach explores performance in the extremes of the weighted function presented in (3.2).

The results in Table 2.6 show that neither the quality greedy approach, nor the quantity greedy approach perform better than the Fed-SRC model shown in Table 2.4. In addition, we note how the data greedy approach results in increased training time due to the larger sample sizes in the local data of the

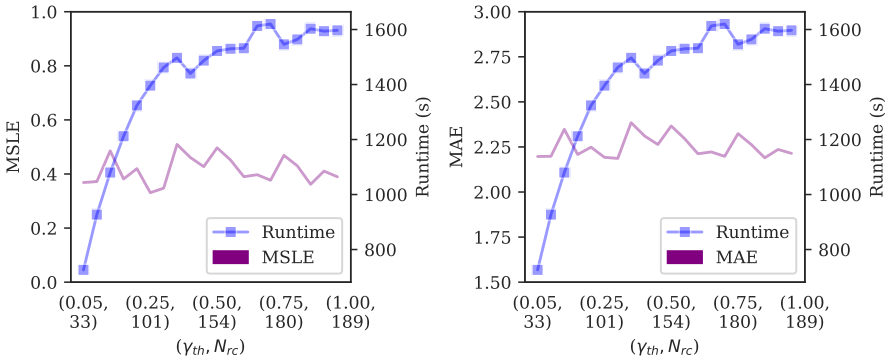


Figure 2.4: Runtime versus MSLE (left) and runtime versus MAE (right) in function of gradually increasing values for γ_{th}. Corresponding to a gradually increasing number of recruited clients, denoted as N_{rc}, for each of the training iterations.

recruited clients. The findings reported in Table 2.6, when compared to the results for Fed-SRC in Table 2.4, show the value of combining both divergence and local sample size in the client recruitment process as neither of the extreme strategies outperform the combined approach.

Table 2.6: Model performance for federated training with data greedy and quality greedy recruitment strategies.

Strategy	MAE	MSLE	τ(s)
Fed-SRC-QG	2.23 $\pm$ 0.03	0.40 $\pm$ 0.03	891 $\pm$ 25
Fed-SRC-DG	2.22 $\pm$ 0.06	0.39 $\pm$ 0.05	1137 $\pm$ 15

Having covered the effects on performance under different settings for γ_{dv} and γ_{sa}, we investigate how different settings for γ_{th} affect the performance and training time. Higher values for γ_{th} directly correspond to more recruited clients for the federation, which in turn corresponds to higher training time. To this extent, the value for γ_{th} is gradually increased in steps of 0.05. For each step we observe the performance on the test set in terms of MSLE, MAE and training time as shown in Fig. 2.4. This shows how there is no direct relation between performance and a higher number of recruited clients, even more so, close to optimal performance can be obtained for low values of γ_{th}, i.e., with few of the most representative clients making up the federation.

2.6 Conclusion, Limitations & Future Work

2.6.1 Conclusion

In this work, we present a client recruitment approach considering only the local output distribution and local sample size. In addition, we show practical relevance of the proposed method in the medical setting. By recruiting clients in function of the herein defined client-level representativeness, those clients with smaller sample size in combination with those for which the output distribution vastly diverges compared to that of the global data are pre-excluded from the set of potential clients for the federation. By applying client recruitment, the predictive performance of the federated models significantly increases compared to the models trained with the standard FL approach. In addition, training time was significantly reduced as a direct result of the reduced number of clients that partook in training.

2.6.2 Limitations & future work

The main limitation of this work stems from the introduction of the recruitment parameter, γ_{th}, which directly affects the number of clients recruited for the federation. In a real-world setting, tuning this parameter is not always feasible. In addition, the present work is executed in a simulated, single process environment for which communication overhead is not a factor.

Future work will evaluate performance in a real-world setting were data is hosted on actual servers corresponding to the hospitals in separate networks. Here, client recruitment is of even greater importance as it can greatly reduce the overall required server-client communication. In addition, a direction for future research is to look at how to, a priori, approximate the optimal setting for γ_{th}. Furthermore, future work will explore alternative recruitment strategies with a focus on sampling from diverse subgroups in the data and assess local performance of the federated models against models trained on the local data only.

2.7 Supplementary context

This section provides additional context for the client recruitment approach applied to ICU Length-of-Stay prediction. It addresses the motivation for key

design choices, discusses limitations of the current evaluation, and clarifies the relationship between this chapter and Chapter 3.

2.7.1 Situating this work

This chapter builds on the theoretical framework for client recruitment proposed by Ruan et al. [92], who provide an optimization-based formulation for selecting clients prior to federation formation. The contributions of the present chapter are threefold: (i) we operationalize client-level representativeness as a function of only the local target distribution and sample size, which avoids the need for any local training prior to recruitment; (ii) we introduce a binning strategy for converting continuous LoS into a discrete distribution suitable for divergence computation; and (iii) we provide empirical validation on a real-world, large-scale ICU dataset comprising 189 clients, which goes substantially beyond the synthetic experiments in [92]. The choice of divergence metric, the construction of the bins, and the threshold mechanism all represent design decisions that required careful calibration to the healthcare domain.

2.7.2 Recruiting on the outcome distribution

A central design choice in this chapter is the use of the target (LoS) distribution as the basis for recruitment. This constitutes selection on Y, which requires careful justification. LoS in ICUs reflects not only patient acuity but also hospital discharge policies, bed pressure, referral pathways, and administrative coding practices. Two hospitals with identical patient populations could exhibit different LoS distributions simply because of differing institutional practices. As such, distributional divergence from the global LoS distribution does not necessarily indicate that a client's data is uninformative; it may instead indicate that the client operates in a clinically or operationally distinct environment.

The justification for this approach rests on pragmatic grounds. First, the target distribution is the only output-side statistic that can be shared without revealing patient-level information and without requiring any local model training. Second, in the specific prediction task studied here, the model is asked to predict LoS given clinical features; a client whose LoS distribution is very different from the global distribution is likely to introduce gradient updates that push the global model in a direction that is less useful for the majority of clients. Third, the empirical results demonstrate that recruitment improves both performance and efficiency, lending post-hoc support to the approach. That said, this strategy can exclude hospitals that are clinically different but whose data would improve transportability. A more robust approach could incorporate

feature-distribution statistics alongside target-distribution statistics, or could employ a two-stage strategy in which initially excluded clients are periodically re-evaluated. This limitation is partially addressed in Chapter 4, where RAFA's adaptive weighting adjusts client influence during training rather than making a binary inclusion/exclusion decision upfront.

2.7.3 Sensitivity to binning

The recruitment score computation requires discretizing continuous LoS values into bins to compute the distribution divergence. The bin structure used, $[(0,1), [1,2), \ldots, [7,8), [8,14), [14,+\infty)]$, was chosen to provide fine granularity in the clinically most common range (0–8 days, covering approximately 85% of patients) while aggregating the long tail into two broader bins. This design reflects the clinical observation that short-stay and medium-stay patients are operationally distinct, while very-long-stay patients (>14 days) represent a heterogeneous group whose specific duration is less predictable and less differentiating across hospitals.

The sensitivity of the recruitment outcome to bin design was examined informally during development. Halving the bin width in the 0–8 day range (i.e., using 16 bins instead of 10) produced nearly identical recruitment outcomes, suggesting that the method is not highly sensitive to granularity in the dense region. However, the definition of the tail bins does influence which hospitals are flagged as divergent: a hospital with a disproportionate number of patients in the 8–14 day range would be treated differently under a finer tail partition. A principled alternative would be to use a continuous divergence measure, such as the Wasserstein distance, which does not require binning. This would come at the cost of requiring more detailed distributional information to be shared. Another failure mode arises when two hospitals differ mainly in rare extreme LoS cases; the coarse tail bins may mask this difference, leading to recruitment of a client whose rare-event behavior is unrepresentative. For LoS prediction, where the primary clinical interest lies in the common-range predictions, this is an acceptable trade-off, but it should be noted for tasks where tail behavior is critical.

2.7.4 The γ_{th} parameter

The threshold parameter γ_{th} directly controls the number of recruited clients and thus the size and composition of the federation. In the experimental setting, γ_{th} was tuned by examining performance across a range of values (Figure 2.4). In a real-world deployment, such tuning may not be feasible due to the cost of

repeated training runs. We propose several principled alternatives for setting γ_{th} without extensive tuning.

First, a budget-based approach: if the deploying consortium has a fixed computational or coordination budget (e.g., a maximum of 50 participating institutions), γ_{th} can be set to recruit exactly that number by choosing the threshold that selects the top-k clients by representativeness score. Second, a diminishing-returns criterion: one can examine the cumulative representativeness curve (ν sorted in descending order) and select the elbow point where marginal representativeness drops below a predefined fraction of the mean. Third, a statistical criterion: since the recruitment score is a sum of divergence and sample-size terms, one can set γ_{th} to exclude clients whose divergence exceeds a specified number of standard deviations from the mean divergence across all clients. In practice, monitoring the validation performance on a held-out global test set during the first few rounds of training can provide an early signal of whether the recruited set is too aggressive (performance degrades on under-represented subgroups) or too conservative (training time is excessive with marginal performance gain).

2.7.5 GRU and LSTM

We selected GRU and LSTM architectures for the LoS prediction task because the input data is inherently sequential: clinical features are recorded at hourly intervals over the first 24 hours of ICU admission, creating a temporal sequence of variable informativeness. Both architectures are well-established for clinical time-series modeling and have demonstrated strong performance on the eICU dataset in prior work [90]. GRU was included because its simpler gating structure (two gates versus LSTM's three) results in fewer parameters and faster training, which is a desirable property in the federated setting where communication cost scales with model size.

Simpler baselines, such as linear regression, ARMA models, or a last-value-carried-forward heuristic, are not included in the comparison. This is a limitation of the experimental design; such models would provide a useful floor against which the added complexity of recurrent architectures could be justified. However, prior work on the eICU dataset has established that recurrent architectures significantly outperform linear models for LoS prediction [90], and the primary contribution of this chapter is the recruitment mechanism rather than the model architecture. Transformer architectures that encode temporal order represent another viable alternative that could offer better parallelization and attention-based interpretability. The choice not to explore transformers in this chapter reflects the chronological development of the research; transformer-

based temporal models for clinical data were less established when this work was conducted. Future work could profitably re-evaluate the recruitment strategy with foundation time-series models as the base architecture.

2.7.6 Use of ReLU

The use of the ReLU activation function in the final fully connected layer (Equation 2.3) may appear unconventional for a regression task. The motivation is straightforward: LoS is a strictly non-negative quantity, and $\text{ReLU}(x) = \max(0, x)$ enforces this constraint directly in the model architecture. Without such a constraint, the model could produce negative predictions, which are physically meaningless. Alternative approaches include using a softplus activation ($\log(1 + e^x)$), which is smooth and strictly positive, or predicting log(LoS) and exponentiating, which additionally addresses the right-skewed nature of the LoS distribution. We chose ReLU for its simplicity and widespread availability, and it empirically performs well on this task. ReLU's non-differentiability at zero and its dead neuron property (gradients are exactly zero for negative pre-activation values) could theoretically cause training difficulties if a large fraction of predictions are near zero. In practice, this was not observed because the LoS distribution has very few values near zero. For future work, the softplus alternative is recommended as a more principled choice for non-negative regression.

2.7.7 All clients versus recruited clients

The main results in Table 2.4 report performance on the global test set, which includes data from all 189 hospitals. This raises a valid concern: by excluding clients whose data diverges most from the global distribution, the recruitment approach may be making the problem easier rather than genuinely improving generalization. To address this, Chapter 3 presents additional experiments (Table 3.11) evaluating the global model exclusively on data from non-recruited clients, both with and without local fine-tuning. These results show that excluded clients do not have access to a model that works well out-of-the-box for their local data, but can achieve strong local performance through a single round of fine-tuning. This two-tier outcome, high global performance for recruited clients, acceptable local performance for excluded clients after fine-tuning, is the expected deployment scenario. We communicate it as such, rather than claiming universal performance improvement.

2.7.8 Selection paradox

An observation that merits explanation is that Fed-AC (all clients, no recruitment) sometimes performs competitively or even best in terms of raw predictive metrics, despite requiring substantially more training time. This is not surprising: including all clients provides the model with the maximum diversity of training data, which can improve generalization. The value proposition of client recruitment is therefore not that it always achieves strictly better prediction performance, but that it achieves comparable or near-comparable performance at a fraction of the cost. In the results presented, Fed-SRC achieves an MAE within 0.01 days of Fed-AC for the GRU model while reducing training time by approximately 50%. In practical healthcare deployments where coordination across institutions carries significant administrative overhead, this trade-off is highly favorable. Whether to prioritize maximum performance (use all clients) or operational efficiency (recruit a subset) is ultimately a deployment decision that depends on the consortium's resources and risk tolerance.

Chapter 3

Target Informed Client Recruitment for Efficient Federated Learning in Healthcare

This chapter has been published as:

Scheltjens, V., Momo, L.N W., Verbeke, W., De Moor, B. (2024). **Target Informed Client Recruitment for Efficient Federated Learning in Healthcare**. *BMC Medical Informatics and Decision Making*, 24 (1), Art.No. ARTN 380.

Paper abstract:

Background: Modern machine learning and deep learning methods have been widely incorporated in decision making processes in healthcare in the form of decision support mechanisms. In healthcare, data are abundant but typically not centrally available and, therefore, require some form of aggregation to facilitate training procedures. Aggregating sensitive data poses a significant privacy risk, which is why, both in Europe and the United States, legal frameworks regulate the treatment of such data. Whilst these measures protect the individual behind the data, they pose a significant challenge that results in extensive legal administration related to data sharing efforts. Federated learning (FL) offers

a way to mitigate these challenges by allowing to learn models in distributed fashion, eliminating the need to aggregate data for the purpose of training. However, FL comes with a new set of challenges related to communication overhead, client selection and efficiency of the FL training procedure, among others.

Methods: In this work, we extend on a previously proposed client recruitment approach by incorporating knowledge on the local hardware such that it becomes possible to recruit a subset of clients for the federation based on the construct of client-level representativeness, which is expressed in terms of the local target distribution divergence, sample size, and the underlying hardware.

Results: We show that, for prominent, medical regression and classification tasks, the recruitment approach yields results that are on par, or better, compared to the central and federated approaches. The proposed approach requires a mere fraction of the data for training and reduces the training time by a factor of 3-4. In addition, we show that excluded clients can still significantly benefit from the resulting federated model through local fine-tuning.

Conclusions: By expressing the representativeness of clients in function of the deviation in the local target distribution, the sample size and efficiency of the underlying hardware, we are able to define a recruitment approach that yields a subset of clients for the federation resulting in significantly reduced training time, without harming predictive performance, whilst improving the privacy preserving characteristics compared to the standard FL and central approaches.

3.1 Background

Machine learning (ML) and deep learning (DL) methods have been widely adopted in the healthcare domain and are accompanied by a significant track record showing their value over a broad spectrum of applications [8, 13, 97]. Medical data is omnipresent in large quantities and different modalities, often generated at different medical facilities. Patients can be admitted to different hospitals, medical care facilities, different wards within hospitals, and so on. Each of these institutions generates data related to the care and stay episodes for the corresponding patients, resulting in large quantities of distributed data in different modalities. For example, patients admitted to the Intensive Care Unit (ICU) are continuously monitored and generate sequences of temporal data. When need be, these patients may be transferred to the radiology ward where medical imaging generates visual representations along with, most often, written radiology reports. As such, the data is abundant, heterogeneous and

distributed. The distributed nature of the data manifests at different levels. At the intra-hospital level, data can reside on different systems pertaining to different wards. At the inter-hospital level, data pertaining to a single patient may partially reside at different institutions as a result of different care episodes or distinct specialties at each of the care facilities. By jointly leveraging all the data as input for ML and DL methods, new insights can be derived and patient care can be improved beyond what was previously possible.

However, the decentralized nature of medical data poses a major challenge for the adoption of ML and DL methods. Especially for DL, large quantities of data are required which often need to be pooled and aggregated centrally to facilitate training. However, legal and regulatory frameworks such as the General Data Protection Regulation (GDPR) [28], the European Union Artificial Intelligence Act, and the United States counterpart, the Health Insurance Portability and Accountability Act (HIPAA) [14], impose significant restrictions on such central aggregation. These frameworks are rightfully instated to warrant privacy for the individuals behind the data, nevertheless, by so doing, inherently and unintentionally, they impose a significant burden for the adoption of advanced DL methods in the healthcare domain. Currently, remediating these challenges requires for de-identification of the data and subsequent agreement upon extensive data sharing agreements.

Federated Learning (FL), as originally proposed by McMahan et al. in 2017 [67], stands as a means to mitigate the restrictions imposed by the legal frameworks and, as showcased in the literature, FL is increasingly studied for applications in healthcare. In recent work, the value of FL has been illustrated for various research and application domains such as for medical imaging, and more specifically classification and segmentation tasks [59, 69, 89, 101, 119]. FL allows to learn complex DL models over decentralized data without the need for direct local access. Therefore, data owners maintain full control over the data, eliminating the need to share or centrally aggregate, which directly translates in a reduction of the administrative and legal efforts that would be required otherwise.

In FL, a central server orchestrates the training procedure of a defined model over multiple local clients. These clients can correspond, in practice, to hospitals, mobile phones, government institutions, pharmaceutic companies etc., essentially any sort of party that locally hosts data and can provide computational resources. The local clients, in combination with the central server, make up the federation which communicate with each other bidirectionally to pass model updates. Learning a model in the federation generally follows the predefined FL algorithm outlined in Algorithm 1, [67]. As illustrated in Algorithm 1, the local clients provide a model update to the central server after a local iteration of training. These model updates are aggregated into a global update according to the

standard *FedAvg* algorithm in which the local weight matrices are averaged into the global update [67].

Algorithm 1 Standard Federated Learning procedure with T, the number of training rounds, C, all the clients in the federation indexed by c and E, the number of epochs for each round of local training.

Server-side execution:
Initialize global model parameters, w
for each training round $t = 1, 2, \ldots, T$ **do**
  Sample a fraction C_t of C clients
  for each client $c \in C_t$ **do**
    $g_c \leftarrow \text{ClientUpdate}(c, w_t)$
  end for
  $\bar{g} \leftarrow \frac{1}{|C_t|} \sum_{c \in C_t} g_c$ ▷ Aggregate local updates
  $w_{t+1} \leftarrow$ Update global model using $\bar{g}$
end for

procedure ClientUpdate(c, w)
  $w_c \leftarrow$ Update local parameters with w
  for local epoch e from 1 to E **do**
    $g_c \leftarrow \nabla f_c(w_c)$ ▷ Local parameter update
  end for
  Send g_c back to the server
end procedure

Although, FL provides a way to mitigate otherwise complex privacy related data challenges, it comes with its own set of challenges. As indicated, FL facilitates learning procedures in distributed settings, which inherently means that some form of communication between the participating instances is required. Depending on the scale and amount of participants, the required communication may result in significant overhead which directly influences the cost and the efficiency of the operation. A vastly different, but major, challenge relates to the nature of distributed data. Real-world, decentralized data is not independent and identically distributed (non-IID). Decentralized sets of data are especially characterized by different underlying distributions driven by demographic and geographic parameters as well as differences in terms of the unit of observation, sampling rate, deviating margins of error on the hardware level and different established processes for monitoring and reporting. This non-identically distributed nature of decentralized data can harm the federated training procedure, or any training procedure for that matter, due to the weight divergence in local model updates driven by the non-IID data [129].

Consequently, this not only harms the training procedure but the eventual outcome in terms of the obtained predictive performance.

In FL, the random client selection procedure for each round of training, as illustrated in Algorithm 1, has been shown to be robust against the non-IID nature of decentralized data [67]. However, a significant amount of research effort is being dedicated to defining more efficient, informed ways of selecting the best clients at each round of training [73,117,121]. These approaches mostly relate to selecting clients in function of previous updates, by prioritizing those that provided a more constructive update [123,125]. For instance, in [18], the authors show how biasing client selection towards selecting clients with larger local loss, results in faster global error convergence. The major downside of these approaches is that the evaluation of the value a certain client brings, is ad hoc. Therefore, all clients need to partake in the federation and participate in at least one round of training, which is computationally expensive and can pose a burden in practice. Similarly, significant research efforts are being dedicated to the aggregation approaches for construction of the global updates. Again to mitigate the downsides of the non-IID data, as well as to allow for more robust and efficient training that reaches convergence at a faster pace compared to the standard FedAvg approach. To that extent, more advanced methods look at weighted averaging depending on the local sample size or class imbalance such as in [66]. Other approaches such as Stochastic Controlled Averaging [50], look to mitigate phenomena such as client-drift resulting from heterogeneous, non-IID data. In Table 3.1 we present a condensed, non-exhaustive, overview of the seminal work related to federated learning, client selection and aggregation.

In this work, we tackle the challenges related to the non-iid nature of decentralized data in FL, for which our proposed approach inherently reduces the communication overhead and improves the privacy preserving aspect. As indicated, real-world, decentralized data is non-iid in nature. Therefore, the utility of the data hosted by each of the participants may vastly differ from one host to another. More specifically, data hosted at a given participant may be limited in terms of the instances, be characterized by underlying distributions that are not representative for the average distributions across all participants, or may simply be hosted on outdated, slow, hardware. In the standard federated learning approach, all participants contribute initially to the federation, and only once the orchestrating party has gained sufficient understanding of how valuable certain participants are, it can make informed decisions about which participants partake henceforth. Such client selection procedures are computationally expensive and require for all participants to contribute at least once to the procedure. Here, we aim to remediate this by relying on client recruitment, the foundations for which are provided in the work by Ruan et al. [92]. In doing so, we try to evaluate, a priori, which participants

Table 3.1: Condensed overview of seminal work related to federated learning approaches, client selection and aggregation.

Approach	**Authors**	**Year**	**Key Features**
FedAvg	McMahan et al.	2017	Averages model updates from multiple clients to create a global model [67]
FedProx	Li et al.	2020	Extends FedAvg by addressing system heterogeneity and varying amounts of client data. [93]
FedCS	Nishio et al.	2019	Optimizes client selection based on resource conditions to enhance training efficiency [74]
SCAFFOLD	Karimireddy et al.	2019	Combines adaptive sampling and stochastic weight averaging to improve training efficiency and accuracy. [50]

can yield valuable contributions. More formally, a subset of participants can be recruited for which it is known, to some extent, that they are representative of the global population and will yield valuable contributions, resulting in faster convergence without having to sacrifice predictive performance. Essentially, participants with non-representative local data can yield harmful model updates at training time. Therefore, these participants are excluded from partaking in the training procedure, consequently, omitting the need for more intelligent selection and aggregation strategies. To do so, a means to evaluate the local representativeness needs to be constructed. Specifically in this work, we extend our proposed client recruitment approach [95] by incorporating knowledge on the local underlying hardware architecture. As such, the resulting recruitment approach depends on the updated construct of client-level representativeness which is expressed in terms of the local divergence in the target distribution, local sample size, and a proxy of the local training time computed based on the hardware information. All of which constitute high-level, non-privacy infringing statistics that are readily available for each of the participants.

The practical validity and utility of the proposed approach is evaluated in two prominent healthcare settings. The first of which entails Length of Stay (LoS) prediction on patients admitted to the ICU, using the real-world eICU dataset [34, 78, 79] for which only 47 out of the 189 potential clients are recruited. The second setting constitutes the multi-label chest radiograph (CXR) classification problem, using the MIMIC CXR data set with structured labels [49], for which 29 out of the 100 potential clients are recruited. These two settings cover both managerial and clinical problems in healthcare that call for regression and classification approaches, and models, respectively. For each of these settings, we show that, with federations constructed of recruited clients only, the resulting models yield predictive performances that either outperform or perform on par with the performances obtained from the central and standard federated models as proposed in [67], albeit, at a fraction of the required training time, whilst providing a full privacy-guarantee for the non-recruited clients. In addition, insights are provided that corroborate the validity of the approach. By evaluating performance of the trained model on clients that did not partake in training (non-contributing clients), we show that with a single round of local fine-tuning, the non-contributing clients can achieve performance on par or better compared to performance for the contributing clients. Furthermore, we visually show that in the classification setting, the model trained with the recruitment approach learns to attribute importance to the exact same sub regions as the other models, whilst having been exposed to a mere fraction of the data at training time.

The remainder of this work is structured as follows: in Section 4.3 the client recruitment problem for federated learning is presented along with our proposed

method and experimental settings. Section 4.5 provides an overview of the obtained results which are further discussed in Section 4.6. At last, concluding remarks are provided along with suggestions for future work in Section 3.5.

3.2 Methods

3.2.1 Client recruitment

In this work, we do not further explore the alternative approaches for client selection and model aggregation. However, focus is shifted towards client recruitment, which we consider the mechanism to be invoked prior to federated learning, consistent with the work in [92, 95]. The client recruitment problem for federated learning involves recruiting a subset of clients from a larger pool to participate in the training process, essentially before any training has occurred. Formally, given a set of clients C each with local data D_c, the problem is to select a subset $C_k \subset C$ for which we can, a priori, say they will have valuable contributions to the federation based on a limited set of statistics.

By considering a set of limited, non privacy sensitive, statistics pertaining to each of the potential clients, we can define, to some extent, how valuable each potential client is to the federation. By restricting the federation to the most valuable clients, the expectation is that the cost of model training significantly reduces without harming predictive performance.

We define client recruitment as the mechanism that operates on a pool of potential clients to establish the federation for training, such that the output constitutes the clients that will partake in the federation. Consistent with the work in [92] and [95] we consider a set C of c potential clients from which a subset of clients will be recruited. Each of the clients in C hosts a local dataset $D_c = \{(x_i, y_i)\}_i$ where x_i and y_i respectively denote the local inputs and corresponding targets.

To initiate the client recruitment procedure, each client in C reports a tuple (P_c, n_c, H_c) to the orchestrating server. The tuple contains three pieces of information; P_c, the local target distribution, n_c, the local sample size and H_c, the underlying hardware that will support the local rounds of training. H_c in itself represents the floating point operations per second (FLOPS) the underlying hardware is capable of computing at a utilization rate that corresponds to the batch size used at runtime. The obtained information allows for n_g and P_g to be calculated as:

$$n_g = \sum_{c=1}^{c} n_c, \quad P_g = \sum_{c=1}^{c} P_c, \tag{3.1}$$

with n_g the global sample size as the sum over all local sample sizes and P_g the global target distribution as the sum over all local target distributions. This global target distribution is what will be used in the subsequent steps to calculate to which extent a particular local target distribution deviates from the global distribution, hence the 'target informed' aspect of the proposed client recruitment approach.

To further facilitate the client recruitment procedure, the local representativeness for each client, ν_c, is calculated as:

$$\nu_c = \gamma_{dv} \underbrace{\left| \frac{P_g}{n_g} - \frac{P_c}{n_c} \right|}_{\beta} + \gamma_{sa} n_c^{-0.5} + \gamma_{tr} \theta_c, \tag{3.2}$$

where ν_c is expressed as the weighted function of the target distribution divergence, local sample size and local computational efficiency with γ_{dv}, γ_{sa} and γ_{tr} the corresponding weight parameters for each of the three components contained in the expression. To illustrate the inner workings of (3.2), we revisit the structure of the tuple c_i, which consists of three components: P (the outcome distribution over local classes), n (the number of local instances), and H (the computational capacity). Consider three clients with the following characteristics:

$$c_1 = ([10, 40], 50, 2.0)$$

$$c_2 = ([9, 21], 30, 1.5)$$

$$c_3 = ([5, 15], 20, 1.0)$$

For these three clients, the global statistics amount to $P_g = [24, 76]$ and $n_g = 100$. Additionally, we will consider a batch size of 4 for this illustration. Thus, the local representativeness for c_1 is calculated as:

$$\nu_1 = \gamma_{dv} \left| \frac{[24, 76]}{100} - \frac{[10, 40]}{50} \right| + \gamma_{sa} \cdot 50^{-0.5} + \gamma_{tr} \cdot \frac{\text{flop}(50/4)}{2.0},$$

where γ_{dv}, γ_{sa}, and γ_{tr} can be set as parameters assigning weight to each of the components in (3.2).

Furthermore, in (3.2), β denotes the divergence of the local target distribution compared to the global target distribution and is calculated as the absolute value of the difference between the normalized global and local target distributions. Furthermore, by including the term $n_c^{-0.5}$, clients with larger local data sets are favored over those with smaller data sets. As discussed in [92], the larger n_c, the better the local empirical distribution, $\tilde{P}_c$, approximates P_c. This observation finds its foundation in the work discussed in [42, 98] in which is shown that $\tilde{P}_c - P_c$ converges to $\mathcal{N}(\mu, \sigma^2)$ with $\mu = 0$ at the rate of $O(n_c^{-0.5})$. The last term contributing to the local representativeness in (3.2), denoted by θ_c, corresponds to the local computational efficiency based on the underlying local hardware details H_c. Each of the components in the ν metric are individually min-max normalized such that the amplitude of the weight parameters affects each of the components in similar fashion. Furthermore, based on H_c, θ is calculated as:

$$\theta_c = \frac{flop(n_c \div bsize)}{H_c}, \tag{3.3}$$

where $flop$ denotes the floating point operations required for a full forward and backward pass of the data contained in a single batch, $bsize$ corresponds to the batch size used for training and H_c denotes the effective FLOPS the underlying accelerator can compute per second. Essentially, θ_c represents local computational efficiency expressed in terms of the approximate time required to process all local data in a single round of training. Intuitively, the inclusion of this term means that, ceteris paribus, clients with more potent local compute are favored over those with less compute at their disposal.

To recruit the final subset of clients for the federation, the local representativeness values, ν_c, are sorted and stored in the vector $\boldsymbol{\nu}$. We define the concept of global representativeness, ν_g, as:

$$\nu_g = \sum_{c=1}^{c} \nu_c, \tag{3.4}$$

which is, in turn, used to define ι, the fraction of ν_g that should be covered by the clients in the resulting federation. To that extent, the fraction is calculated as $\iota = \gamma_{th}\nu_g$ with γ_{th} a user defined parameter. The subset of recruited clients, C_k, is obtained by collecting the first k clients in the sorted vector $\boldsymbol{\nu}$ such that the fraction ν_g they jointly represent satisfies the defined threshold, ι. This subset of k most representative clients constitutes the clients recruited for the federation.

We iterate that the client recruitment procedure is invoked only once during initialization of the overall training process. It does, however, constitute an extra step compared to methods such as Federated Averaging [67]. In terms of time complexity, the proposed method involves reporting the local statistics, computing the global statistics and subsequently computing the local representativeness, all of which is done in $O(n)$ operations where n corresponds to the number of clients. Sorting the resulting values to select the top k clients requires $O(n \log n)$. Therefore, our proposed method adds $O(n \log n)$ complexity to the end-to-end process, under the assumption that this additional step is offset by the benefits of reduced communication and computational overhead during training.

3.2.2 Experimental setup

To evaluate the efficacy and utility of the client recruitment approach proposed in section 3.2.1, two vastly different experimental settings are defined. Both of which are of significant importance in the healthcare domain. In the first setting, client recruitment is evaluated on the LoS problem whereas the second setting assesses client recruitment for accurate multi-label classification of CXRs. The former calls for a regression approach and is more managerial in nature, whereas the latter constitutes a classification problem which is foremost clinical in nature. With this, a broad spectrum of problems in the healthcare domain is covered for which the client recruitment approach can be evaluated using different data modalities and models.

Figure 3.1 depicts a holistic overview of the experimental setup in which the classification and regression models along with the recruitment approach are visually represented. Figure 3.1 shows how the central server is in charge of establishing a pool of recruited clients by obtaining high level information from the potential client pool. The recruitment in itself is performed according to the method described in section 3.2.1. Subsequently, the established pool of recruited clients is used for the remainder of the standard federated learning procedure in accordance with the depicted models, for which the data, architectures, and experimental settings are presented in the subsequent sections, 3.2.3 and 3.2.4.

3.2.3 Regression

The performance and utility of the client recruitment approach for federated models, as described in 3.2.1, in the regression setting, is evaluated on the LoS problem, where LoS is defined as the remaining time in ICU for a given patient. Here, LoS is calculated as the difference between time at discharge or death T_D

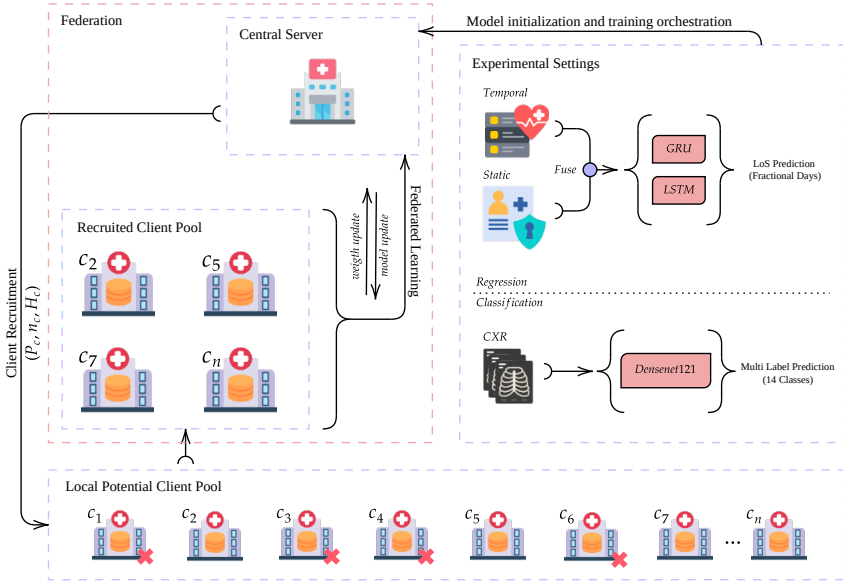


Figure 3.1: Overview of the experimental setup of the client recruitment approach for federated learning in the regression and classification settings.

and time at admission into ICU T_A, and is expressed in terms of fractional days. Specifically, the task is to accurately predict patient LoS in ICU, similar to the work in [1, 90, 91, 107]. Formally, LoS is defined as follows, given a dataset D of patient records, each consisting of features x_i and corresponding LoS y_i for each individual patient, where $y_i = T_{Di} - T_{Ai}$, the goal is to learn a function $f(\mathbf{x})$ that yields $\hat{y}$ as a prediction for y with minimal error.

Data

To evaluate the proposed approach in the regression setting on real-world, multi-center data, the eICU data set is used [34, 78, 79]. The eICU data set contains data pertaining to 208 US hospitals which jointly cover 139.000 uniquely admitted patients with over 200.000 registered patient stays between 2014 and 2015.

The benefit of using the eICU data set with respect to FL is two-fold in that, (i) it is comprised of real world data and (ii) all of the instances and observations contained within the data set can be mapped to the originating institution.

The latter is of particular interest as this results in true non-IID data splits across the various institutions, which is the core envisioned application domain of the client recruitment approach. To obtain a final, workable cohort of data from the raw set, the data is preprocessed in accordance with the preprocessing pipeline proposed in [91]. The cohort is restricted to adult patients only, for which the first 24 hours of data into ICU are extracted and used to predict LoS. The extracted data contains a mix of temporal and static features that are fused into a single set of inputs. The set of constructed inputs is cleaned, re-sampled, one-hot encoded and imputed as needed. To track the time since last true observation for the imputed values, a decay mask is added for the temporal features. For an extensive discussion of the preprocessing pipeline we refer to the work by Rocheteau et al. [91].

The obtained data cohort, shown in Table 3.2, contains 89,127 unique stays which stem from a total of 189 hospitals. A detailed overview of the static and temporal features in the data cohort is provided in Table A.1.

Table 3.2: Overview of the extracted and preprocessed data cohort for the eICU dataset.

Number of patient stays	89,127
Train	62,375
Validation	13,376
Test	13,376
Mean LoS	3.69
Median LoS	2.27
Number of features	35
Temporal	17
Demographic	18
Number of hospitals (clients)	189

As indicated, LoS is expressed in terms of fractional days, which is continuous in nature. However, to obtain client level representativeness as shown in (3.2), the divergence of the local target distribution compared to the global target distribution (β) needs to be calculated. Therefore, we discretize the continuous target by constructing ten bins, such that the bins represent the frequency of target values that fall within the given bounds. The bins are defined as: $[(0,1),[1,2),[2,3),...,[7,8),[8,14),[14,+\infty)]$ and are used to perform class counting to obtain local target distributions.

Figure 3.2 displays the global target distribution, corresponding to P_g as the aggregation of all local target distributions, compared to a subset of local

Table 3.3: Extracted temporal and static features from eICU dataset

Type	Feature	Description
Temporal	FiO2	Patient's FiO2 value
	Bedside glucose	Patient's glucose level
	Cvp	Patient's cvp value
	Heartrate	Patient's heart rate value
	Noninvasivediastolic	Patient's non invasive diastolic value
	Noninvasivemean	Patient's non invasive mean value
	Noninvasivesystolic	Patient's non invasive systolic value
	Respiration	Patient's respiration value
	Sao2	Patient's spO2 value
	St1	Patient's st1 value
	St2	Patient's st2 value
	St3	Patient's st3 value
	Systemicdiastolic	Patient's diastolic value
	Systemicmean	Patient's mean pressure
	Systemicsystolic	Patient's systolic value
	Temperature	Patient's temperature value in celsius
	Hour	Time since admission
Static	Hospitalid	Surrogate key for the hospital
	Gender	Gender of the patient
	Age	Patient's age in full years
	Admissionheight	Admission height of the patient in cm
	Admissionweight	Admission weight of the patient in kg
	Intubated	Whether patient is intubated at the time of the worst ABG result
	Vent	Whether patient is ventilated at the worst respiratory rate
	Dialysis	Whether patient is on dialysis
	Eyes	GCS score (1 to 4)
	Motor	GCS score (1 to 5)
	Verbal	GCS score (1 to 6)
	Meds	Whether GCS score could not be obtained due to meds
	Ethnicity	Patient's ethnicity
	Unittype	The picklist unit type of the unit
	Unitadmitsource	Picklist location from where the patient was admitted
	Unitstaytype	Patient's unit stay type
	Physicianspeciality	Picklist specialty of the care provider
	> 89	Whether patient is over 89 years old

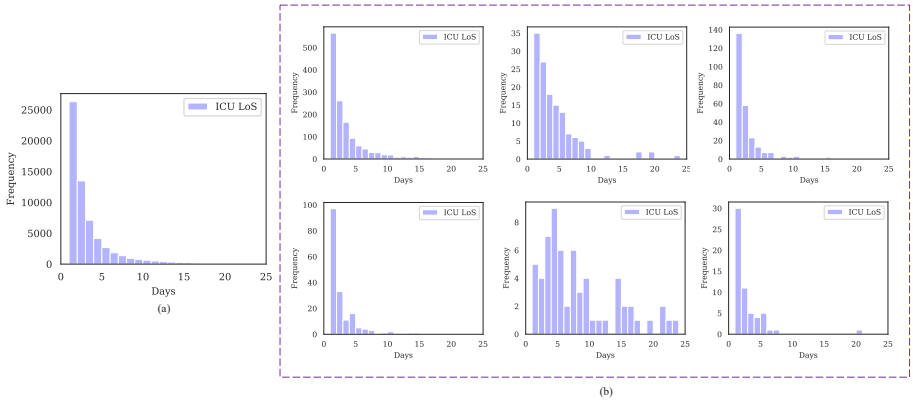


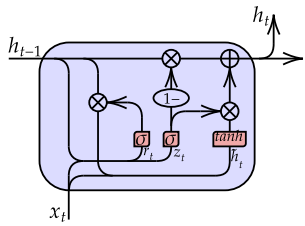
Figure 3.2: The global target distribution calculated as P_g, (a), and a subset of six local target distributions of the, in total, 189 target distributions, (b). Each of the distributions shows the frequency in terms of patients with an observed LoS, restricted until 25 days, locally and globally.

target distributions. Intuitively, the client recruitment approach aims to pre-exclude those clients for which the local target distribution vastly diverges from the global target distribution, the sample size is not sufficiently large or the underlying hardware results in inefficient training. Figure 3.2 illustrates the divergence in some of the local target distributions with respect to the global target distribution. Clients with the most divergent target distributions are pre-excluded from the federation as they can potentially contribute non-representative updates. The underlying assumption is that, in the federated setting, when local model updates are driven by non-representative data, the contribution of the local model update to the federation is of lesser value. When the local model update is considered at face value for aggregation into the global update, the effect of a non-representative update is not accounted for in the standard FedAvg setting. To account for the negative effects of such an update, weighting schemes for the aggregation procedure should be introduced. This, however, does not further enhance the privacy preserving aspect, nor does it reduce the pool of clients in the federation, resulting in reduced training time. Nevertheless, each excluded client will still obtain a trained model for local use at inference time, as will be discussed in Section 4.5.

Model architecture

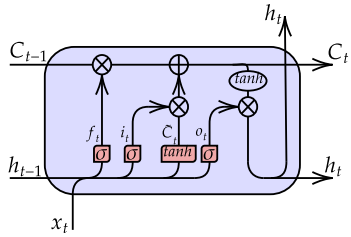
For both the central and federated training procedures, two models are employed, the Gated Recurrent Unit (GRU) [17], and the Long Short-Term Memory (LSTM) [41] networks. Both models belong to the class of deep learning models coined as the Recurrent Neural Networks (RNN) and have shown good performance when dealing with sequential data containing temporal relations. Both the GRU and LSTM cells are comprised of a set of gates that define the information flow and retention. As such, the GRU cell is made up of two sole gates. The reset and update gates, respectively denoted as r_t and z_t in (3.5). Here, z_t controls the information that is to be retained from the previous state whereas r_t controls the amount of information from the previous state that is to be forgotten. The reduced computational complexity stemming from the two gates is a desirable characteristic for FL where communication overhead and local computational efficiency pose major challenges.

The governing equations for the GRU cell are shown in (3.5), along with the visual representation. For each discrete time step t in the input sequence, the computations outlined in the governing equations occur.



$$
\begin{aligned}
r_t &= \sigma(W_r \cdot [h_{t-1}, x_t] + b_r) \\
z_t &= \sigma(W_z \cdot [h_{t-1}, x_t] + b_z) \\
\tilde{h}_t &= \tanh(W \cdot [r_t \odot h_{t-1}, x_t] + b) \\
h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t
\end{aligned}
\tag{3.5}
$$

Similarly, the governing equations for the LSTM cell are outlined in (3.6), along with the visual representation. Compared to GRU, LSTM counts one additional gate for a total of three gates, denoted as f_t, i_t and o_t, which correspond to the forget gate, input gate and output gate.



$$
\begin{aligned}
f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\
i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\
\tilde{C}_t &= \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \\
o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\
h_t &= o_t \odot \tanh(c_t)
\end{aligned}
\tag{3.6}
$$

The input gate, i_t, defines how much new information is transferred to the current state at time t, whereas the forget gate, f_t, controls the amount of information from the previous state that is to be disregarded. The output gate, o_t, controls how much information of the new cell state can be contained in the new hidden state.

In both the GRU and LSTM architectures as outlined in (3.5) and (3.6), x_t denotes the input at time step t, h_{t-1} is the hidden state at the previous time step $t-1$ and h_t denotes the hidden state at the current time step t. Furthermore, σ is the sigmoid activation function, tanh corresponds to the hyperbolic tangent activation function and $\odot$ represents element-wise multiplication.

The hidden state h_t, i.e., the output of the cell, is provided as the input of a nonlinear Fully Connected Network (FCN), which yields a single output value representing the predicted LoS. The nonlinearity stems from the ReLU activation function leveraged in the FCN as shown in (3.7),

$$
\hat{y}_t = ReLU\left(W_{y_t} h_t + b_{y_t}\right), \tag{3.7}
$$

with $\hat{y}_t$ the predicted value for LoS at time t. Employing $ReLU(x) = max(0, x)$ forces the outcome to be strictly positive. It is impossible for a patient to have a negative LoS, therefore we restrict the model to only yield predictions in the positive domain.

Experimental settings

To allow for comparison between the traditional training procedure in which data is centrally aggregated, the federated setting with standard federated averaging and the federated setting with the proposed client recruitment approach, we

Table 3.4: Hyperparameter settings used for both central and federated training.

Model	L	N	η	m	wd	r
GRU	2	32	0.005	128	0.005	0.05
LSTM	2	32	0.005	128	0.005	0.05

define experimental settings for the central and federated approaches in this section. The defined experimental setup serves the main purpose to answer three central questions related to the proposed approach; (i) Can client recruitment improve efficiency without sacrificing performance compared to alternative approaches? (ii) Are findings consistent across data modalities and learning tasks? and (iii) Can non-contributing clients still benefit from a resulting global model? To facilitate comparison, several of the settings are fixed over the different settings such that, at training time, all training procedures use *AdamW* [64] for optimization and the Mean Squared Logarithmic Error (MSLE) as the loss function, which is calculated as:

$$MSLE = \frac{1}{n}\sum_{i=1}^{n}(\log(y_i + 1) - \log(\hat{y}_i + 1))^2, \tag{3.8}$$

with y_i the true target value and $\hat{y}_i$ the predicted value. Furthermore, the model hyperparameters are fixed over all training iterations, both central and federated. The exact settings are shown in Table 3.4, with L the number of layers, N the hidden dimension for each of the layers, η the learning rate, m the batch size, wd the weight decay for the *AdamW* optimizer and r the dropout rate. To obtain the hyperparameters reported in Table 3.4, grid search was performed for GRU to determine the optimal number of hidden layers L and the number of hidden units N per layer. Subsequently, batch size (m), learning rate (η), weight decay (wd), and dropout rate (r) were obtained using population-based training [47]. The hyperparameters optimized for GRU were then kept constant for LSTM models to maintain consistency and comparability across experiments.

For evaluation of the performance, all resulting models are evaluated against the hold-out test set containing data from all 189 hospitals. In addition to the MSLE, models are evaluated using the Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and the Mean Squared Error (MSE), shown in (3.9). As an indication of the time complexity, the training time, denoted as τ, is reported in seconds.

$$
\begin{aligned}
MAE &= \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \\
MAPE &= \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \\
MSE &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2
\end{aligned}
\tag{3.9}
$$

Central Training is performed consistent with the traditional DL procedure in which all data is assumed to be centrally available. The architectures presented in (3.5) and (3.6) are trained for a predetermined amount of 15 epochs using the global train and validation sets, which consist of the accumulated data over all potential clients, i.e., the 189 originating hospitals. The resulting central model is subsequently evaluated against the hold-out test set.

Federated Training with and without client recruitment is simulated as a single process using the *FedML* framework as proposed in [38]. In this work, for FL without client recruitment, clients are either all considered in each round of training or a subset is randomly sampled consistent with the standard client selection implementation described in *FedAvg* [67]. For FL with client recruitment, the client recruitment process described in 4.3 is invoked prior to initiating the federation. Following the described implementation, the recruitment of clients is influenced by four user-defined hyperparameters $\gamma_{dv}, \gamma_{sa}, \gamma_{sa}$ and γ_{th} which respectively define the importance of the divergence in the target distribution, the local sample size, the importance of the efficiency of the underlying hardware, and the fraction of the global representativeness to be covered by the recruited clients.

For LoS prediction, four different strategies are implemented. The four approaches differ in terms of the number of clients that partake in the federation, denoted as ϵ, the percentage of clients in the federation that contribute to each training round, denoted as δ, and whether the federation is comprised of recruited clients, or all clients. For each of the federated models, each client trains for four epochs per round of server-client communication for a total of 15 rounds. The resulting model is subsequently evaluated against the hold-out test set.

The specifications for the four different FL strategies are; (i) **Fed-AC** (Federated All Clients): all clients make up the federation and partake in each training

round, (ii) **Fed-SC** (Federated Selected Clients): all clients make up the federation, 10% of which are randomly sampled to partake in each training round, (iii) **Fed-ARC** (Federated All Recruited Clients): recruited clients make up the federation and partake in each training round and (iv) **Fed-SRC** (Federated Selected Recruited Clients): recruited clients make up the federation, 10% of which are randomly sampled to partake in each training round.

Following the introduction of the strategies, we revisit and extend Algorithm 1 to create a formal representation of the Fed-SRC strategy which reflects the client recruitment procedure, nested in the entire federated training approach. This formal representation can be found in Algorithm 2. Similarly, the formal, algorithmic representations for Fed-AC and Fed-ARC can be found in Appendix A.2. We note that Algorithm 1 constitutes a formal representation of the Fed-SC strategy.

Time complexity: As indicated in the description for each of the individual strategies, each of the included strategies considers a different number of clients at each round of training depending on the clients in the federation and sampling strategy. Therefore, the overall time complexity for each of the strategies is different. To that extent, Fed-SC and Fed-SRC have the added advantage that for each round of training, only a subset of the clients is selected for training, which lowers the required communication and computational overhead. For example, the Fed-AC strategy considers all clients in every round, resulting in the highest overhead with a time complexity of $O(T \cdot n \cdot E \cdot f)$, where T is the number of global iterations, n is the number of clients, E is the number of local epochs, and f represents the complexity of local training. The Fed-SC strategy randomly selects a subset of clients each round, reducing the overhead to $O(T \cdot k \cdot E \cdot f)$, where k is the number of sampled clients, making it more efficient. The Fed-ARC strategy recruits a subset of representative clients, maintaining a similar complexity to Fed-AC but with a smaller n, as the total pool of clients in the federation has already been reduced by the recruitment procedure. Finally, the Fed-SRC strategy optimizes further by sampling from the recruited subset, significantly reducing training time and complexity to $O(T \cdot m \cdot E \cdot f)$, where m is the number of sampled recruited clients, thus achieving the lowest overhead among the proposed methods.

Additional analysis: To provide more extensive insight with respect to the proposed client recruitment approach, the approach is evaluated in two additional settings, using the GRU based Fed-SRC strategy. Without interfering in the parameter settings for Fed-SRC, the client recruitment approach is tweaked such that in the first additional setting, a subset, equally large as in the normal Fed-SRC approach, of random clients is recruited. In the second

Algorithm 2 Fed-SRC: Federated Learning with Selected Recruited Clients

Server-side execution:
Initialize global model parameters, w
for each client $c \in C$ **do**
  $P_c, n_c, H_c \leftarrow \text{ClientReport}(c)$
end for
Calculate global statistics: $P_g = \sum_{c \in C} P_c$, $n_g = \sum_{c \in C} n_c$
for each client $c \in C$ **do**
  Calculate local representativeness:

$$\nu_c = \gamma_{dv} \left| \frac{P_g}{n_g} - \frac{P_c}{n_c} \right| + \gamma_{sa} n_c^{-0.5} + \gamma_{tr} \theta_c,$$

end for
Select top k clients to yield C_k

for each training round $t = 1, 2, \ldots, T$ **do**
  Sample a fraction $C_{k,t}$ of C_k clients
  for each client $c \in C_{k,t}$ **do**
    $g_c \leftarrow \text{ClientUpdate}(c, w_t)$
  end for
  $\bar{g} \leftarrow \frac{1}{|C_{k,t}|} \sum_{c \in C_{k,t}} g_c$ ▷ Aggregate local updates
  $w_{t+1} \leftarrow$ Update global model using $\bar{g}$
end for

procedure ClientReport(c)
  $P_c \leftarrow$ Target distribution at client c
  $n_c \leftarrow$ Number of data points at client c
  $H_c \leftarrow$ Computational capacity of client c
  return (P_c, n_c, H_c)
end procedure

procedure ClientUpdate(c, w)
  $w_c \leftarrow$ Update local parameters with w
  for local epoch e from 1 to E **do**
    $g_c \leftarrow \nabla f_c(w_c)$ ▷ Local parameter update
  end for
  Send g_c back to the server
end procedure

setting, the subset, again equally large compared to the subset in the normal Fed-SRC approach, of least representative clients is recruited. Intuitively, if the standard Fed-SRC approach does in fact yield a more representative subset of clients, proposed approach should outperform the two additional experimental strategies.

Given that the client recruitment mechanism pre-excludes a set of clients to take part in the federation for training on the premise that, the locally hosted data is not sufficiently large, diverges significantly in terms of the target distribution or has inefficient hardware resources for training, we, in addition, evaluate the performance of Fed-SRC on the subset of non-recruited clients with and without a single round of domain adaptation. More specifically, the globally obtained federated model is tested against the locally hosted test sets at the non-recruited clients, only. Subsequently, each of the non-recruited clients is allowed to run a single round of fine-tuning on the local data after which the fine-tuned model is again tested against local test data only.

3.2.4 Classification

In addition to the regression setting, the client recruitment approach is evaluated on the multi-label CXR classification problem. To this extent, the MIMIC CXR data set with structured labels [49] is used. Each of the CXR images in the data set is labeled with a multi-label, containing information on the presence or absence of the 14 possible pathologies. The task is to accurately classify any given CXR into any of the 14 classes. The task is multi-label in nature with 14 dimensions in the outcome, pertaining to the 14 classes. Similar to the work in [45], we restrict the performance evaluation to five classes of interest, namely, Atelectasis, Cardiomegaly, Consolidation, Edema and Pleural Effusion.

Data

The MIMIC CXR data set [49] with structured labels contains 227,827 observations, corresponding to 377,095 multi-view chest X-Rays. In Table 3.5, a descriptive summary is provided for the data set and the corresponding train, validation and test split used throughout this work. The splits are performed such that multi-view CXRs corresponding to a single study belong to the same split to avoid information leakage.

The data is labeled such that for each observation, a corresponding vector of length 14 exists for which each entry in the vector constitutes the indication of whether the corresponding class is observed (1), not observed (0), or it is

Table 3.5: CXR data set observation count and corresponding splits

CXR images	377,095
Observations	227,827
Train	265,019
Validation	55,807
Test	56,269
Number of clients	100

uncertain (-1) as to whether the pathology is present. Following the work presented in [45], the uncertainty labels are dealt with by replacing (-1) with (1), which has been shown to yield better predictive performance. In Table 3.6, an overview is provided of the corresponding label counts on each of the 14 classes.

Table 3.6: Overview of the positive, negative and uncertain label observations for each of the 14 classes in the MIMIC CXR data set.

Observation	Positive	Negative	Uncertain
Atelectasis	45,808	1,531	10,327
Cardiomegaly	44,845	15,911	6,043
Consolidation	10,778	7,967	4,331
Edema	27,018	25,641	13,174
Enlarged Cardiomediastinum	7,179	5,283	9,375
Fracture	4,390	886	555
Lung Lesion	6,284	862	1,141
Lung Opacity	51,525	3,069	3,831
No Finding	75,455	152,373	0
Pleural Effusion	54,300	27,158	5,814
Pleural Other	2,011	126	765
Pneumonia	16,556	24,338	18,291
Pneumothorax	10,358	42,356	1,134
Support Devices	66,558	3,486	237

The CXR images contained in the data set are derived from the raw DICOM files. To that extent the pixel values were normalized to the range $[0, 255]$, and subsequently histogram equalized to enhance contrast. To do so, pixels are forced towards 0 or 255 such that all pixels in $[0, 255]$ appear equally frequent in a given CXR. Subsequently, the images are converted to JPG. In addition to dealing with the uncertainty labels, the raw JPG files are preprocessed by center

cropping each CXR to the dimension 224×224. This is done to fit the input dimensions expected by the deep neural network that is employed, without losing valuable information in the image. In addition, a random horizontal flip is performed on the input in the train set to allow for more robust training.

To facilitate federated learning on the CXR data, we construct 100 silos, each of which is assigned with a sample rate. Subsequently, the data is split into 1000 data shards from which each of the silos samples shards in accordance to the previously assigned sample rate. More specifically, each silo samples between 1 and 22 shards of data. This yields a data set that is imbalanced in terms of local sample sizes. Nevertheless, slight label imbalance as a result of the structure of the data is also present. To facilitate client recruitment, and more specifically, the calculation of β in (3.2), the local class counts for each of the 14 classes in the target are considered as the local target distribution.

Model architecture

For the central and federated training approaches in the classification setting, a single deep convolutional neural network, DenseNet-121 [44], is used. The DenseNet-121 architecture is based on the Densely Connected Convolutional Networks design principles and consists of several core components. The major building blocks the architecture is comprised of, are a combination of convolutional layers, pooling layers, transition layers, dense blocks and a final classification layer. The dense blocks constitute a sequence of convolutional layers with batch normalization and activations to densely connect the feature maps. The dense connections signify shorter connections between layers close to the input and output by connecting each layer to every other subsequent layer. By doing so, for any layer, all feature maps from previous layers are used as input to that layer [44]. The task of the transition layers is to downsample the feature maps to fit to the input of the next dense block through convolutions, batch normalization and pooling. The consecutive combination of each of these components for the DenseNet-121 architecture used in this work is summarized in Table 3.7, for which the last layer corresponds to the classification layer that yields the predictions for each of the 14 class labels.

Experimental settings

For classification, we again maintain the same model and hyperparameter settings across the different learning approaches. As such, we utilize the Binary Cross Entropy (BCE) loss with logits, which is equivalent to wrapping regular BCE in a sigmoid activation function. In addition, *AdamW* [64] is used for

Table 3.7: DenseNet-121 model architecture.

Layer Type	In/Out Size	Architecture
Convolution	112×112	7×7 conv, stride 2
Pooling	56×56	3×3 max pool, stride 2
Dense Block 1	56×56	$\begin{bmatrix} 1 \times 1 & conv \\ 3 \times 3 & conv \end{bmatrix} \times 6$
Transition Layer 1	56×56 28×28	
Dense Block 2	28×28	$\begin{bmatrix} 1 \times 1 & conv \\ 3 \times 3 & conv \end{bmatrix} \times 12$
Transition Layer 2	28×28 14×14	
Dense Block 3	14×14	$\begin{bmatrix} 1 \times 1 & conv \\ 3 \times 3 & conv \end{bmatrix} \times 24$
Transition Layer 3	14×14 7×7	
Dense Block 4	7×7	$\begin{bmatrix} 1 \times 1 & conv \\ 3 \times 3 & conv \end{bmatrix} \times 16$
Classification Layer	1×1	7×7 global average pool 14 dimensional fully-connected

optimization throughout the training iterations with the learning rate set to $1e-3$ and the weight decay set to $1e-5$. Both the learning rate and decay, were selected based on population-based training [47] for hyperparameter selection in the central setting. The achieved values were found to provide a good balance between convergence speed and model accuracy, centrally. Upon centrally establishing the parameters, they were kept constant throughout all federated experiments. Furthermore, all training procedures are performed with a batch size of 128. For evaluation of the model performance, the area under the receiver operator curve (AUC) is calculated for each of the 5 classes of interest and the training time, denoted as τ, is reported to assess time complexity of the different training procedures.

Central Training: Similar to the regression setting, all CXR data is assumed to be centrally available. The DenseNet-121 architecture is fully fine-tuned on the preprocessed CXR data for 3 consecutive epochs and evaluated against the

hold out test set, shown in Table 3.5.

Federated Training: Here, two different federated strategies are considered. Fed-AC and Fed-ARC are omitted, the reason for which is two-fold; (i) these strategies are less likely to occur in a real world setting with many clients as they rapidly become infeasible to compute within reasonable and actionable time, and (ii), due to computational limitations it is not feasible to compute these approaches in the experimental setup. As such, the two remaining strategies entail, (i) **Fed-SC**, for which all of the clients make up the federation and a subset is randomly selected to contribute to each round of training and (ii) **Fed-SRC**, for which the recruited clients make up the federation and a subset is randomly selected to contribute to each round of training. The global parameters for training are fixed across both approaches to ensure fair comparison. To that extent, during each round of training, each of the contributing clients to that given round locally trains the model for two subsequent epochs before providing the central server with a parameter update for aggregation. The total number of training rounds is set to 20. Upon completion, the models are evaluated against the hold-out test set for which the per-class AUC values are reported.

Additional analysis: As an additional step in the analysis of the resulting classification models, a multi-view CXR labeled with Cardiomegaly, Edema and Pleural Effusion is selected from the test set for which the Gradient-weighted Class Activation Maps (Grad-CAMs) [96] are computed. Grad-CAM yields visual explanations with respect to the important regions in an image for a corresponding prediction. These visualizations, or activation maps, portray which areas of an image the model attributes importance to with respect to a given outcome. To this extent, the Grad-CAMs for a multi-view CXR with postero anterior and corresponding lateral views are computed, and shown for each of the approaches in the experiment. This allows for visual evaluation of the behaviour of each of the methods. More specifically, this allows to determine whether the Fed-SRC approach learns to attribute importance similar to the central and Fed-SC approaches, whilst relying on significantly less training data.

To verify whether the observations based on the visual analysis hold for all data, we quantify the similarity of the activation maps for a larger subset of images in the test set. To this extent, we compute the Structural Similarity Index (SSIM) [114] and Frechet Inception Distance (FID) [39] for the activation maps with respect to each of the classes of interest. Both of these metrics measure similarity between two sets of images. SSIM and FID are both often used in the evaluation of image synthesis methods, such as Generative Adverserial Networks. They can, however, also be used to assess similarity and quality between two sets of images, of which one, should constitute the ground truth. SSIM evaluates

the structural similarity between two specific images. By averaging the obtained similarity values over a set of images, the average similarity between two sets of images can be obtained. FID looks at the similarity, or dissimilarity between two sets of images in high dimensional space. That high dimensional space is obtained by extracting feature embeddings from both sets of images using the Inception v3 network. The actual distance is computed as:

$$FID = \|\mu - \mu_w\|^2 + tr(\Sigma + \Sigma_w - 2(\Sigma\Sigma_w)^{\frac{1}{2}}), \tag{3.10}$$

with $\mathcal{N}(\mu, \Sigma)$ the multivariate normal distribution estimated from the obtained feature embeddings on the Grad-CAMs for the central model and $\mathcal{N}(\mu_w, \Sigma_w)$, the multivariate normal distribution estimated from the obtained feature embeddings on the Grad-CAMs for either Fed-SC or Fed-SRC.

To analyze the similarity in the activation maps, a random subset of 2000 images from the test set is selected. For each of the CXRs in the subset, the activation maps using Grad-CAM are computed for the respective models and corresponding classes of interest. The resulting analysis is such that the obtained scores correspond to the similarity between (i) Central and Fed-SC, and (ii) Central and Fed-SRC, allowing for comparison between Fed-SC and Fed-SRC.

3.3 Results

In Table 3.8, an overview is provided of the parameter settings for each of the different federated approaches, with or without client recruitment, in the regression and classification settings as described in 3.2.3 and 3.2.4. This section outlines the obtained results in terms of predictive performance and time complexity and, in addition, provides the obtained results for the respective additional analysis, further validating the proposed client recruitment approach.

3.3.1 Regression

Table 3.9 outlines the results for the training procedures in the regression setting for both GRU, LSTM and the corresponding central and federated approaches. The best performing approaches per metric have been highlighted in bold and the statistical significance, with respect to the Fed-SC approach, is indicated at the 1% and 5% confidence level.

The results in Table 3.9 reflect the performances of the federated approaches as well as the central approach for both GRU and LSTM. For GRU, our proposed

Table 3.8: Paramater settings for the client recruitment approach in both the regression and classification settings, with ϵ the total number of clients in the federation, δ the number of clients randomly sampled from ϵ in each round of training and (γ_{dv}, γ_{sa}, γ_{tr}, γ_{th}) the hyperparameters for client recruitment.

Setting	Strategy	ϵ	δ	γ_{dv}	γ_{sa}	γ_{tr}	γ_{th}
Regression	Fed-AC	189	189	-	-	-	-
	Fed-SC	189	19	-	-	-	-
	Fed-ARC	47	47	0.4	0.2	0.1	0.1
	Fed-SRC	47	5	0.4	0.2	0.1	0.1
Classification	Fed-SC	100	10	-	-	-	-
	Fed-SRC	29	6	0.5	0.5	0.4	0.2

approach, Fed-SRC, yields the best performance amongst the federated models for MAE, MAPE and τ. Whereas Fed-AC returns the best performance on MSE and MSLE. For LSTM, Fed-AC yields the best MAE, MAPE and MSLE. Fed-SRC obtains similar predictive performance with the best time to convergence, τ. We note that for both models, Fed-ARC has a higher time to convergence compared to Fed-SC. Even though the Fed-ARC strategy establishes a federation containing recruited clients, the strategy mandates that all recruited clients partake in every round of training. In Fed-SC only 10% partakes in training each round. In Table 3.8, δ denotes the number of clients in each round. For Fed-SC and Fed-ARC this amounts to 19 and 47 respectively, which is in turn why Fed-SC converges faster.

For both GRU and LSTM, the proposed Fed-SRC strategy significantly outperforms the traditional FL approach, Fed-SC, in terms of MAE. For both models, the observed performance for MAE is similar to, or on par with the central approaches at approximately a fourth of the training time.

Additional insights

The subsequent section outlines the results pertaining to the additional experiment in which the client recruitment strategy is informed to (i) randomly recruit a subset of clients, or (ii) recruit the subset of least representative clients. For both approaches, the parameter settings reported in Table 3.8 remain unchanged, yielding a subset of 47 random clients, and a subset of the 47 least representative clients respectively. The results in Table 3.10 show that when GRU is trained using the Fed-SRC strategy with either random or the worst clients in the recruited subset, the models yield a predictive performance

Table 3.9: Model performance for central and federated models with and without client recruitment. Statistical significance among the federated models in comparison to Fed-SC is indicated as * at the 5% significance level and ** at the 1% significance level.

Model	Strategy	MAE	MAPE	MSE	MSLE	τ(s)
GRU	Central	2.21 ± 0.02	0.58 ± 0.06	21.94 ± 0.63	0.33 ± 0.01	2129 ± 18
	Fed-AC	2.26 ± 0.05	0.64 ± 0.07**	**21.58 ± 0.69****	**0.33 ± 0.02****	5232 ± 27**
	Fed-SC	2.26 ± 0.06	0.46 ± 0.06	23.98 ± 1.27	0.41 ± 0.06	1470 ± 35
	Fed-ARC	2.27 ± 0.12	0.57 ± 0.17	22.67 ± 1.83	0.37 ± 0.05	3359 ± 25**
	Fed-SRC	**2.21 ± 0.03****	**0.46 ± 0.05**	23.44 ± 0.85	0.38 ± 0.04*	**546 ± 26****
LSTM	Central	2.19 ± 0.02	0.53 ± 0.05	22.39 ± 0.47	0.34 ± 0.01	1892 ± 15
	Fed-AC	**2.20 ± 0.03****	0.47 ± 0.05*	**23.21 ± 0.49****	**0.37 ± 0.02****	4668 ± 50**
	Fed-SC	2.27 ± 0.08	0.45 ± 0.03	24.22 ± 1.16	0.43 ± 0.06	1313 ± 20
	Fed-ARC	2.26 ± 0.04	**0.43 ± 0.01**	24.42 ± 0.46	0.43 ± 0.03	1493 ± 34**
	Fed-SRC	2.22 ± 0.03*	0.45 ± 0.02	23.81 ± 0.47	0.40 ± 0.02	**616 ± 45****

with a MAE of 2.31 days and 2.33 days respectively. Overall, the performance stemming from the procedure with randomly selected clients is observably better compared to the worst clients strategy.

Table 3.11 presents the results with respect to the second additional experiment in which the obtained federated model is tested against data from non-contributing clients only, yielding a MAE of 2.30 days. Notably, after a single round of fine-tuning on the local data, the obtained average performance on the local test sets, for the non-contributing clients, amounts to a MAE of 2.05 days. This performance increase, as a result of the local fine-tuning, is also reflected in the improved MSE and MSLE.

3.3.2 Classification

The results for classification are reported in Tables 3.12 and 3.13. Table 3.12 outlines the AUC scores for each of the classes of interest, whereas Table 3.13 shows the training time for each of the corresponding approaches along with the average AUC over the classes of interest.

The results in Table 3.12, represent the achieved per-class AUC for the central, Fed-SC and Fed-SRC approaches. The central approach and FED-SC approaches perform on nearly on par across the board. Amongst the federated approaches, the proposed Fed-SRC approach performs slightly worse compared to the traditional Fed-SC approach. In this instance, the observed performance difference between the federated approaches, across all the classes of interest, is limited to 0.02 in terms of the AUC.

In Table 3.13, the average AUC over the classes of interest is presented along with the training time required for each of approaches. The proposed Fed-SRC approach outperforms both the central and Fed-SC approaches in terms of the required training time to achieve the reported predictive performance, by a factor of 4. The traditional Fed-SC approach, however, reaches convergence at a slower rate than the central approach.

Additional insights

For classification, Figure 3.3 and Table 3.14 jointly represent the findings for the additional experiment, in which we visually and numerically analyze how the different approaches attribute importance to the input with respect to a certain outcome.

Table 3.10: Model performance for the Fed-SRC approach retrained with a subset of recruited clients that either constitute the least representative clients according the the recruitment approach or a set of random clients.

Model	Strategy	MAE	MAPE	MSE	MSLE
GRU (Fed-SRC)	Random	2.31 ± 0.17	0.46 ± 0.04	24.48 ± 1.63	0.45 ± 0.12
	Least representative	2.33 ± 0.12	0.44 ± 0.02	24.84 ± 1.22	0.47 ± 0.08

Table 3.11: Model performance for Fed-SRC tested against data from non-recruited clients only with and without a single round of domain adaptation (DA).

Model	Strategy	MAE	MAPE	MSE	MSLE
GRU (Fed-SRC)	Non-recruited	2.30 ± 0.03	0.47 ± 0.07	26.25 ± 1.24	0.40 ± 0.05
	Non-recruited + DA	2.05 ± 0.06	0.47 ± 0.05	23.41 ± 0.85	0.38 ± 0.05

Table 3.12: Model performance expressed in terms of AUC scores for each of the classes of interest for central and federated models with and without client recruitment.

Model	Strategy	Atelectasis	Cardiomegaly	Consolidation	Edema	Pleural Effusion
DenseNet-121	Central	0.77 ± 0.01	0.77 ± 0.01	0.74 ± 0.00	0.84 ± 0.00	0.88 ± 0.01
	Fed-SC	0.78 ± 0.00	0.77 ± 0.01	0.74 ± 0.00	0.85 ± 0.01	0.89 ± 0.00
	Fed-SRC	0.76 ± 0.06	0.75 ± 0.05	0.72 ± 0.08	0.83 ± 0.04	0.87 ± 0.03

Table 3.13: Classification training time and AUC over the classes of interest

Model	Strategy	Average AUC	τ(s)
DenseNet-121	Central	0.80 ± 0.06	12607 ± 1137
	Fed-SC	**0.80 ± 0.05**	13330 ± 1221
	Fed-SRC	0.79 ± 0.06	**3141 ± 496**

Figure 3.3 shows for a single multi-view CXR observation in the test set that each of the models consistently attribute importance to the same sub regions. Regardless of the minor variations, the general importance attribution across models and across pathologies is consistent with respect to the outcome.

As a means to corroborate the visual results shown in Figure 3.3 Table 3.14 presents the quantified results with respect to similarity in the importance maps for the different methods over a subset of 2000 images in the test set. As such, for both the Fed-SC and Fed-SRC, the obtained results for SSIM and FID represent the similarity in the activation maps from the respective model and the central model. The results show that, depending on the class of interest, irrespective of the similarity metric, Fed-SRC performs slightly better, or worse compared to Fed-SC. In general, no significant discrepancies, that would signify divergent behaviour in how one of the federated models attributes importance, are observed.

3.4 Discussion

3.4.1 Regression

Based on the results shown in Table 3.9, we note that when considering MAE, Fed-SRC for both GRU and LSTM achieves comparable performance compared to the central approaches for the respective models. With respect to the federated approaches, a discrepancy is observed for Fed-AC, in that for GRU it performs comparable to Fed-SC and Fed-SRC, whereas for LSTM, the Fed-AC approach yields the best performance. As LSTMs are prone to overfitting, a partial explanation can be found in the quantity of data used for Fed-AC. In Fed-AC, all data is used, in contrast to the remaining approaches that leverage a mere fraction of all available data, which could potentially lead to overfitting. In addition, for both models, Fed-AC obtains the best MSLE score, which can again be attributed to a better overall fit to the data, as in Fed-AC, the models have been presented with data from all clients, including those with diverging

Table 3.14: Similarity in terms of FID (0 would signify perfect similarity) and SSIM (1 would signify perfect similarity) for the Grad-CAM activation maps corresponding to 2000 images in the test set and the Fed-SC and Fed-SRC models respectively with activation maps from the central model serving as the ground truth.

Metric	Strategy	Atelectasis	Cardiomegaly	Consolidation	Edema	Pleural Effusion
↑ SSIM	Fed-SC	0.864	0.887	0.991	0.884	0.881
	Fed-SRC	0.831	0.897	0.959	0.886	0.870
↓ FID	Fed-SC	67.88	46.20	6.21	45.11	46.82
	Fed-SRC	70.31	36.92	15.10	39.51	43.85

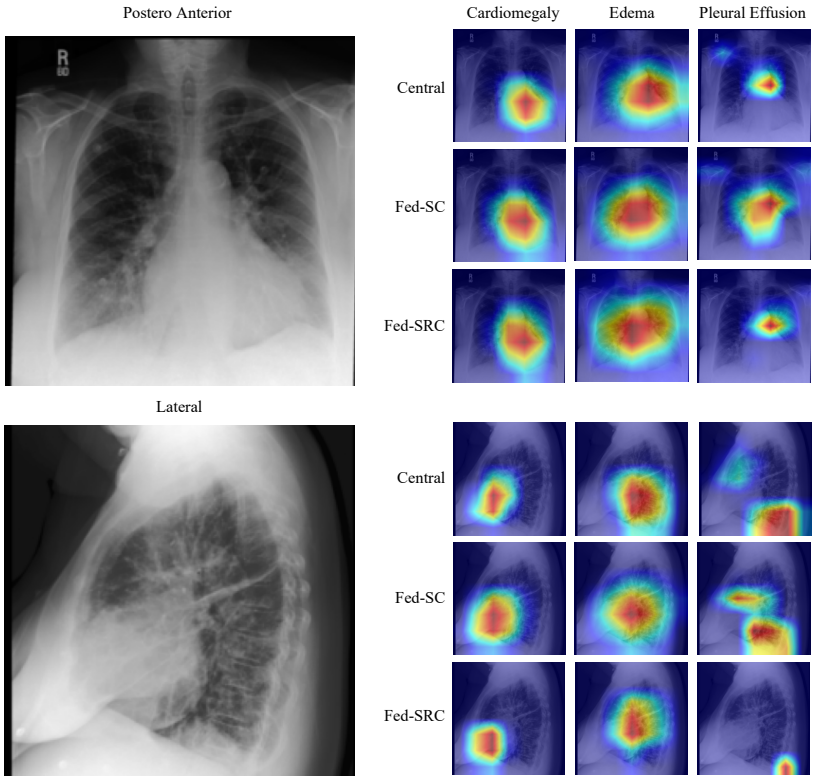


Figure 3.3: Grad-CAM visualizations of the pixel importance for each of the 3 labels corresponding to the multi-view CXR selected from the test set. Grad-CAM visualizations are shown on the right and, the raw postero anterior and lateral views are shown on the left.

target distributions. When considering Fed-SC and Fed-ARC, consistency is observed across all metrics when compared to Fed-SRC. Both Fed-SC and Fed-AC underperform compared to our proposed approach.

In this setting, the main comparative discussion revolves around the central approach, Fed-SC and Fed-SRC. From the results in Table 3.9, we note that for both GRU and LSTM, the client recruitment approach, Fed-SRC, performs significantly better than the standard approach, Fed-SC, with respect to all metrics. Most importantly, the training time reduces significantly whilst the predictive performance improves or remains on par compared to Fed-SC. We further note that, for GRU, Fed-SRC outperforms the central approach, potentially as a result of reduced noise in the training set. As such, the client

recruitment approach in the regression setting, for both GRU and LSTM, outperforms the standard FL approach in a fraction of the required training time. In addition, this means that the individuals behind the data hosted at the non-recruited clients, do not run any privacy risk whatsoever, which constitutes a significant improvement over Fed-SC where all data is subject to some form of privacy risk.

The results in Table 3.10 corroborate that the client recruitment approach does in fact recruit a subset of clients that results in better predictive performance. Both of the additional approaches, in which the least representative clients or a set of random clients make up the federation, yield significantly worse predictive performance compared to the proposed Fed-SRC approach, shown in Table 3.9.

In practice, non-recruited clients would still be provided with a copy of the resulting federated model, which is primarily optimized in function of the global target distribution. This can yield a model that does not generalize well to the non-recruited clients, given that they were excluded from partaking in the federation. As expected, the results in Table 3.11, show that when evaluating the resulting federated model, as-is, against non-recruited clients only, performance is subpar compared to the results in Table 3.9. However, when allowing for a single round of domain adaptation on the local data and subsequent evaluation on the corresponding test set for the local client, performance improves to an extent that, locally, the fine-tuned model yields significantly better predictive performance with an average MAE of 2.05 days across all non-recruited clients. This shows how, when obtaining a global federated model, running one additional round of fine-tuning can, locally, yield performances that significantly outperform the performance of the global federated model. Thus, even for excluded clients, the federated model to which they did not contribute, can still be of significant value. Another benefit relates to the privacy enhancing aspect. The proposed methodology ensures that non-recruited clients do not participate in any of the training rounds and only share a tuple of non-sensitive information when the process is initiated. During recruitment, each client reports the target distribution (P_c), the number of data points (n_c), and the computational capacity (H_c), which are used to calculate local representativeness (ν_c). This shared information does not include any privacy-sensitive details, maintaining the confidentiality of local datasets. Formally, let C be the set of all clients, C_k the set of recruited clients, and $C_{nr} = C \setminus C_k$ the set of non-recruited clients. The non-recruited clients C_{nr} are excluded from training, ensuring that the recruited clients C_k, and the overall training procedure do not access or depend on any information stored at the clients in C_{nr}. This approach ensures that non-recruited clients do not transmit any individual records. Although they do not contribute data during training, they receive the global model for local fine-tuning. This merely requires one-way communication from server to client,

where the non-recruited client does not reveal any information with respect to its local data. Finetuning can then be performed locally, with no further data exchange or requirement to be connected to the federation, preserving the privacy of non-recruited client data.

3.4.2 Classification

The main observation based on Table 3.12, is that each of the different approaches, central, Fed-SC and Fed-SRC, yield similar predictive performance with Fed-SRC performing slightly worse or on par depending on the class of interest. Noteworthy, however, is that Fed-SRC obtains similar performance in a fraction of the time required for the Fed-SC or central approach. More precisely, Fed-SRC reduces the training time by a factor of approximately 4 compared to the other approaches, as shown in Table 3.13. A major drawback from the Fed-SC approach, resulting in high training time, is that the validation during the training procedure runs against all clients in the federation. This imposes significant communication and computational overhead. This setting is identical for Fed-SRC with the benefit that for Fed-SRC the recruited client pool only contains 29 clients, resulting in faster validation times, and therefore, reduced overall training time.

The resulting visualizations in Figure 3.3 show that there is consistency across the models in terms of activated regions in each of the images for the respective pathologies. This representation shows that the approaches identify the pathologies consistently independent of the orientation. We note that the activation maps on the remaining 11 classes are mainly empty. Naturally, the models are not perfect, and thus, misclassifications occur, for which the activation maps will incorrectly show importance. From this single representation and exploratory analysis, we note that the models consistently attribute importance to the same sub-regions to make the prediction. It should be noted here, that the model resulting from the Fed-SRC approach does so whilst only having seen a fraction of the data available at training time compared to the other methods, and additionally, converges in less than a fourth of the required training time.

The results presented in Table 3.14 show that, depending on the class of interest, either Fed-SC or Fed-SRC performs slightly better. It should be noted that a SSIM of 1 represents perfect similarity whereas a lower FID represents higher similarity between the sets of images. Nevertheless, performance across the board is noted to be similar. When comparing with the activation maps obtained from the central model, both Fed-SC and Fed-SRC, attribute importance to similar regions of interest in the images with respect to a certain prediction.

This quantitative approach confirms that the visual observations from Figure 3.3 hold for a larger subset of data. This shows that Fed-SRC, whilst having seen significantly less data, still learnt to correctly attribute importance to the same sub-regions compared to the central and Fed-SC approach, in a fraction of the training time. For some of the classes, Fed-SRC shows higher similarity compared to the ground truth, especially when looking at FID. Similar to the Fed-SRC model in regression outperforming the central approach, this can also be partially attributed to the reduction in noisy data as a consequence of the client recruitment approach.

3.5 Conclusions

In this work, we present the extended client recruitment approach such that clients are recruited in function of the local sample size, the local target divergence and a proxy of the local training time calculated based on the hardware information. Furthermore, we evaluate the proposed approach in the regression and classification settings, encompassing both managerial and clinical problem sets in the healthcare domain. In both settings, we show how models trained with federations made up of recruited clients, outperform, or at least perform on par, in terms of predictive power, compared to the standard federated or central approaches at a fraction of the training time. In addition, we provide evidence with respect to the validity of the client recruitment approaches by retraining the Fed-SRC model with a subset of randomly recruited clients, as well as the subset of least representative clients. In both scenarios, Fed-SRC, with the normal set of recruited clients, outperforms the two alternative approaches. Furthermore, in the classification setting, we visually show how, for a single image in the test set, each of the models attribute importance to similar regions of the image with respect to a given prediction using Grad-CAM importance maps. We further quantify the similarity in the activation maps for each of the models on a subset of 2000 test images and show, how Fed-SRC performs similar to the central model and Fed-SC at a fraction of the training time whilst only having seen a limited subset of the most representative data. These experiments and corresponding additional insights, show that target informed client recruitment yields models that perform better or on par compared to alternative approaches in terms of predictive power in different, relevant healthcare settings whilst significantly reducing the required training time, and providing improved privacy enhancing characteristics.

By introducing the hardware architecture component in the client recruitment approach, we open up potential future research avenues that allow for more detailed assessment of how local training efforts can be expressed in terms of

more complex cost functions. These cost functions can encompass hardware efficiency, related CO_2 emissions, physical distance, demographic factors, etc., among others. This approach could eventually yield a framework that allows to outline the cost structure of intra- and inter-institutional FL endeavours. Future work will assess the performance of the client recruitment approach in a non-simulated setting, which is the current main drawback of this work. The findings presented herein, will still hold in real world implementations as the core concept of client recruitment does not change. However, network latency and communication overhead become an important factor, which can be additional components to be considered in the client recruitment approach.

3.6 Supplementary context

This section provides additional context for the extended client recruitment approach with hardware-aware efficiency and excluded-client analysis. It clarifies the relationship to Chapter 2, addresses key methodological questions, and discusses equity implications.

3.6.1 Hardware-aware recruitment

Incorporating hardware capability (H_c) into the representativeness score raises a legitimate concern: hardware configurations change more rapidly than data characteristics, and a recruitment decision based on current hardware may become outdated within months. Moreover, expressing computational capability as a single scalar (a proxy for expected local training time) is a simplification that obscures important distinctions. More floating-point operations per second do not always translate to proportionally faster training due to memory bandwidth constraints, I/O bottlenecks, or software optimization differences.

In practice, H_c is best understood as a proxy for the expected local training cost rather than a precise hardware specification. The representativeness score uses H_c as one of three terms, with the data-related terms (divergence and sample size) typically dominating the ranking. Hardware information therefore acts as a tiebreaker among data-comparable clients rather than a primary selection criterion. For long-lived federations, H_c should be periodically re-assessed, ideally at the start of each retraining cycle.

3.6.2 Batch size and convergence

In the federated training setup, batch size is a global hyperparameter applied uniformly across all clients. Clients with smaller local datasets therefore complete local epochs faster (fewer gradient steps per epoch) but also see less diverse data per training round. This raises the question of whether smaller clients converge at a different rate and whether they contribute under-trained updates to the global model.

In the current implementation, the number of local epochs (E) is uniform across clients, which means that clients with fewer samples perform fewer gradient updates per round. This can result in smaller clients being systematically under-converged relative to larger ones. Normalization strategies such as FedNova [112] address this by re-weighting updates according to the number of local steps, but were not employed here. The recruitment approach partially mitigates this issue: by excluding clients with very small sample sizes (through the $n_c^{-0.5}$ term), the federation avoids the most extreme cases of under-convergence. A more principled solution would allow heterogeneous local training configurations, where clients with less data train for more epochs or use smaller batch sizes to achieve comparable convergence. We identify this as a direction for future refinement.

3.6.3 Hardware-aware selection

Hardware capability correlates with institutional resources. Compute-aware recruitment may therefore systematically exclude smaller or under-resourced hospitals, which are often precisely the institutions that would benefit most from participating in a collaborative learning effort. This creates a tension between optimization efficiency (selecting the fastest clients) and equity (ensuring broad institutional representation).

We do not resolve this tension fully but offer two mitigating observations. First, excluded clients still receive the global model and can fine-tune locally, as demonstrated in the excluded-client experiments. They are not left without a solution; rather, they receive a solution with a lower degree of collaborative input. Second, the hardware term in the representativeness score is weighted by a user-configurable parameter (γ_{tr}), which can be set to zero or reduced if the deploying consortium prioritizes equity over speed. In practice, the deploying consortium should report both aggregate and per-site performance metrics, as well as the fraction of the total data covered by the recruited clients, to enable transparent assessment of the trade-offs. Future work should develop explicit fairness constraints that can be incorporated directly into the recruitment

optimization, for example by requiring that the recruited set covers a minimum number of institutions or a minimum geographic diversity.

Chapter 4

RAFA: Robust Adaptive Federated Aggregation for 3D MRI Segmentation

This chapter is submitted for publication as:

Scheltjens, V., Verbeke, W., De Moor, B. (2025). **RAFA: Robust Adaptive Federated Aggregation for 3D MRI Segmentation**. *Submitted at BMC Medical Informatics and Decision Making*

Paper abstract:

Background: In recent years, Federated Learning (FL) has enabled collaborative learning across decentralized data sources with improved privacy characteristics. However, real-world data is more often than not characterized as non-Independent and Identically Distributed (non-IID) across clients. Particularly in medical healthcare settings, where larger hospitals or research centers often host the majority of the available data, this poses challenges for training deep learning models in the federated setting. The skew in the data distribution across clients often results in sub-par predictive power and poor generalization of the resulting federated model. Although several aggregation and client selection strategies can cope with non-IID data to some extent, dealing with extreme cases of non-IID data distributions remains an open problem.

Methods: In this work, we introduce a Robust, Adaptive, Federated

Aggregation (RAFA) scheme that copes well with the most extreme cases of non-IID data and show how it outperforms widely used aggregation and client selection strategies on the CIFAR-10 and MNIST datasets.

Results: In addition, we show that for the real-world FeTS MRI segmentation dataset, RAFA yields better results compared to prominent aggregation and selection strategies among which FedProx and FedAvgM.

Conclusions: We emphasize the importance of such strategies in healthcare, where data heterogeneity is pronounced, and the robustness and generalization of models are of significant importance.

4.1 Introduction

Federated Learning (FL) has been widely studied as a suitable framework for dealing with challenges that arise from privacy sensitive data in the healthcare domain [89]. FL allows multiple clients, such as mobile devices or institutions, to collaboratively train a global model without directly sharing their local data. This decentralized approach to model training is particularly interesting in domains such as healthcare, where patient data is privacy sensitive and central aggregation for the sole purpose of model training results difficult due to ethical and legal constraints [2]. FL, thus, provides practitioners and researchers with a suitable tool that alleviates these challenges as data centralization is no longer needed. However, real-world, medical data is often characterized as non-Independent and Identically Distributed (non-IID) across different institutions [65, 120]. The non-IID nature of medical data poses a challenge for FL methods [65]. More specifically, the variations in decentralized client data distributions can lead to divergent model updates with respect to the global model, resulting in sub-par predictive performance and convergence [130]. In the medical domain, and for the purpose of this work, non-IID data refers to the variations in data distribution across different, decentralized clients. In real-world applications, especially in healthcare, the data collected from different sources (e.g., hospitals, research centers) can vary due to demographic differences, local practices, and equipment used. Without proper aggregation and client selection strategies, this can heavily affect resulting model performance [130].

The introduction of FL resulted in the creation of a new research domain, focused on efficient aggregation and client selection strategies. The key research areas in this domain concern the reduction of communication overhead, more efficient and informative parameter aggregations as well as intelligent client selection with the aim to reach model convergence faster. Proposed methods

in this domain often tackle one or a mixture of these challenges but still exhibit sub-par predictive performance, weak generalization capabilities and poor robustness for extreme cases of non-IID data. As indicated, the data distributions across medical institutions are often heavily skewed and directly proportionate to the size and expertise of the institutions. Standard methods such as FedAvg [67], FedProx [93] or FedAvgM [43] perform well with moderate cases of non-IID data but suffer in extreme cases, often resulting in unstable training, i.e., high variance on the error over training iterations, and potentially sub-par performance in terms of predictive power and generalization of the model.

In this work, we introduce RAFA (Robust Adaptive Federated Averaging), specifically applied to the data distributions in the healthcare domain, for which, as indicated, the distributions across decentralized clients are often heavily skewed and aggregation is not always feasible due to stringent regulations related to the privacy-sensitive nature of the data. RAFA constitutes a dynamic client selection and aggregation strategy aimed at more robust, stable learning and better predictive performance in extreme cases of non-IID data.

We show how RAFA outperforms methods such as FedProx and FedAvgM on the benchmark MNIST [53] and CIFAR-10 [51] datasets for which we simulate extreme cases of non-IID data. In addition, we train a federated 3D segmentation model using the RAFA strategy using the real-world FeTS [4, 77, 87] data for which we adhere to the real-world non-IID distributions of the data.

The remainder of this work is structured as follows: in Section 5.2 an overview of the related work is provided followed by the proposed methodology in Section 4.3 and the experimental setup in Section 4.4. In Section 4.5 the main findings are presented and subsequently discussed in Section 4.6. Finally, concluding remarks are given in Section 4.7. The implementation of RAFA can be found in the following repository: github.com/vscheltjens/rafa.

4.2 Related work

Federated Learning (FL) was first introduced by McMahan et al. [67] along with the development of the Federated Averaging (FedAvg) algorithm. FedAvg aggregates locally computed updates from multiple clients into a global update. FedAvg has been foundational in FL research, and has been shown to be robust against some level of non-IID data [67]. However, FedAvg is not sufficiently capable for dealing with most real-world non-IID data distributions across clients. The non-IID nature of most real-world datasets introduces significant challenges, it can lead to biased updates and slow convergence [130], ultimately

reducing model performance. The foundational work by McMahan et al. [67], therefore, also is a driver for subsequent research aimed at addressing these challenges related to the non-IID nature of real-world data.

To address the challenges related to non-IID data in FL, FedProx was introduced by Li *et al.* [93] as an extension to FedAvg. FedProx modifies the local objective function by adding a proximal term, which penalizes large deviations from the global model. This adjustment mitigates the impact of heterogeneous data by ensuring that local updates do not diverge significantly from the global model, thus improving stability and convergence.

Another improvement over FedAvg is Federated Averaging with Momentum (FedAvgM), which incorporates momentum into the federated averaging process [43]. Momentum in optimization algorithms is typically used to accelerate convergence and smooth out updates by considering past gradients. FedAvgM leverages this concept to enhance the performance of FL in non-IID settings. By maintaining a momentum term that accumulates previous updates, FedAvgM can provide a more consistent direction for the global model updates.

Current research efforts continue to build on these foundations. New methods are continuously developed by integrating advanced machine learning concepts such as adaptive learning rates, personalized models, and meta-learning into the FL framework [29, 48].

One of the main challenges for successfully implementing FL in medical settings is the inherent non-IID nature of medical data [55]. Non-IID data arises from variations in data distributions across clients, which, in the medical context, can be attributed to differences in demographics, equipment, imaging protocols, and disease prevalence [2, 89]. For instance, hospitals in different regions may treat patients with varying demographic profiles, leading to differences in the types of medical conditions encountered and the distribution of data. Additionally, imaging devices, such as MRI or CT scanners, may produce data with varying resolutions or noise levels across institutions, further enhancing data heterogeneity. In addition to the non-IID nature of the data, medical data is also privacy-sensitive, which poses challenges (both regulatory and ethical) when it comes to data aggregation for model training. Medical data is often spread across multiple hospitals, research institutions, and diagnostic centers, and are usually subject to stringent privacy regulations, such as HIPAA [14] in the U.S. and GDPR [28] in Europe. These regulations make centralized data storage and processing difficult.

Fl provides a solution by enabling collaborative model training without the need to share patient data across institutions. Therefore, FL is being increasingly studied as a paradigm to mitigate the issues that arise from the non-IID nature

of decentralized data and regulatory constraints for data sharing [89].

In conclusion, the application of FL in the medical field provides a promising avenue for collaborative healthcare solutions whilst improving patient privacy and alleviating issues related to data sharing and aggregation. However, the non-IID data, typically found in this domain poses a challenge for efficiently implementing FL. These challenges motivate the development of advanced methods, such as RAFA as proposed in this work, to robustly aggregate updates from diverse data sources while maintaining model stability and predictive performance.

4.3 Methods

In this section, we present a Robust, Adaptive Federated Averaging strategy (RAFA). With RAFA, we aim to achieve stable training over iterations, so as to improve predictive performance which is validated in the benchmark settings as well as the real-world segmentation setting. In what follows, a detailed description is provided of (i) the client selection strategy for RAFA, (ii) the aggregation approach and (iii) the end-to-end RAFA pipeline.

4.3.1 RAFA: Robust adaptive federated averaging

Formally, the goal of RAFA is to iteratively select a subset of clients from a larger pool to perform federated training and construct an aggregated update for the global model. Let $\mathcal{C} = \{C_1, C_2, \ldots, C_N\}$ represent the set of N clients. At each communication round t, a subset of clients $\mathcal{S}_t \subseteq \mathcal{C}$ is selected based on a joint contribution score, which is a weighted sum of historical contributions, significance of updates, and informativeness of updates. Each selected client $C_i \in \mathcal{S}_t$ trains a local model using its dataset and sends its local update Δw_i^t to the server. The global model update Δw_g^t is then computed as a weighted aggregation of the local updates:

$$\Delta w_g^t = \sum_{C_i \in \mathcal{S}_t} \frac{J_i^t}{\sum_{C_j \in \mathcal{S}_t} J_j^t} \Delta w_i^t \tag{4.1}$$

where J_i^t is the joint contribution score of client C_i at round t. The global model parameters w_g^t are updated as $w_g^{t+1} = w_g^t + \Delta w_g^t$. This constitutes a single round of training which is iterated until some objective is met or training is completed.

4.3.2 Client selection

As indicated, RAFA at its core relies on the joint contribution score, which is denoted here as (J_i), and is computed for each client i as a weighted sum of three core components: historical contributions (H_i), the significance of updates (S_i), and the informativeness of updates (I_i). Thus, the joint contribution score is given as:

$$J_i = \alpha H_i + \beta S_i + \gamma I_i \tag{4.2}$$

where α, β, and γ are adaptive weights which automatically update in favor of the component that contributed most to the update in the previous round of training. By defining the joint contributions score as such, a subset of selected clients is obtained containing clients whose updates are significant and informative while maintaining a balanced representation of clients based on past contributions. In the next paragraphs, each of the core components that make up the joint contribution score are discussed in detail.

The historical contribution component, H_i, constitutes the cumulative impact of each client's updates on the global model across previous training rounds. More specifically, H_i is calculated as the average contribution over all rounds in which the client has participated. For a client i that has participated in T_i rounds, the historical contribution is expressed as:

$$H_i = \frac{1}{T_i} \sum_{t=1}^{T_i} C_i^t \tag{4.3}$$

where C_i^t represents the contribution of client i during round t.

The second component denoted as S_i, or more precisely, the significance of an update, measures the magnitude of each client's parameter updates, indicating how significant their updates are for the global model. The significance is determined by averaging the norm of the updates for each layer in the model. If L represents the total number of layers in the model and $\Delta w_{i,l}$ is the update to layer l from client i, the significance of client i's updates is calculated as:

$$S_i = \frac{1}{L} \sum_{l=1}^{L} \|\Delta w_{i,l}\|. \tag{4.4}$$

As a third and last component, the contribution score considers the informativeness of given update with respect to a client, denoted as I_i. Different

to the significance of the update, which measures how much an update contributes, the informativeness quantifies how well an update aligns with the current direction of the global model. I_i is calculated as the cosine similarity between a client's parameter updates and the global model parameters. For each layer l, the informativeness is given by:

$$I_i = \frac{1}{L} \sum_{l=1}^{L} \frac{\Delta w_{i,l} \cdot w_{g,l}}{\|\Delta w_{i,l}\| \|w_{g,l}\|} \tag{4.5}$$

where $w_{g,l}$ represents the parameters of the global model at layer l. Similar to what momentum is typically used for, this third component allows for the strategy to prefer clients for which the updates are better aligned with the global model.

In this approach, historical contributions are of importance under the assumption that clients that initially yielded significant updates still contain valuable information. Even when seemingly exhausted, we do not fully want to neglect this information. Nevertheless, we consider it desirable to prioritize the more recent contributions, wich we denote as the recency of an update, such that they have a greater impact. As such, the joint contribution score is weighted by recency. The recency-weighted average (RW_i) for client i is computed as:

$$RW_i = \frac{\sum_{k=0}^{T_i-1} \lambda^k J_i^{(T_i-k)}}{\sum_{k=0}^{T_i-1} \lambda^k} \tag{4.6}$$

where $\lambda \in (0, 1]$ is the recency weighting factor, and $J_i^{(T_i-k)}$ is the contribution score of client i at round $(T_i - k)$. This ensures that more recent updates have a higher influence on client selection.

Based on the recency-weighted joint contribution scores, RAFA incorporates a metric to evaluate client-level contributions and select the top-performing clients, while a small subset of clients is chosen randomly. More specifically, at the operational level, RAFA selects a fraction of clients from a client pool. Initially, this selection is random, as from the second round, RAFA (i) evaluates the contribution for each client in the selection, (ii) ranks and carries over the top performers to the next round and tracks historical contributions and, (iii) randomly selects a fraction of clients from the client pool to substitute the dropped clients.

In this work, we found that a fraction of .5 of the selection to be carried over consistently works well. Nevertheless, these fractions can be completely determined by the user. For example, in a federation with 100 clients for which

it is determined that for each round 10% of clients contribute to training, 10 random clients will be initially randomly selected for which, after round 1, the 5 top-performers are carried over and 5 new clients are randomly selected for round 2, and so on. This combination of prioritizing high-contribution clients and injecting randomness ensures that the model remains robust and generalizes well to unseen data.

Bringing this all together yields the client selection strategy that governs RAFA. To that end, Algorithm 3, provides an overview of the end-to-end client selection procedure herein defined.

Algorithm 3 RAFA: Client Selection

Require: $\mathcal{C}$: Set of all clients
Require: f: Fraction of clients to be selected each round
Require: f_c: Fraction of clients to be carried over from the previous round
Require: λ: Recency weighting factor
Require: $\mathcal{S}$: Set of selected clients
Ensure: $\mathcal{S}_{\text{new}}$: Set of clients selected for current round

if $\mathcal{S} = \emptyset$ **then**
    $\mathcal{S}_{\text{new}} \leftarrow$ random sample of size $f \cdot |\mathcal{C}|$ from $\mathcal{C}$
    return $\mathcal{S}_{\text{new}}$
end if
for each client $i \in \mathcal{S}$ **do**
    $R_i \leftarrow \frac{\sum_{t=1}^{T_i} \lambda^{T_i - t} C_i^t}{\sum_{t=1}^{T_i} \lambda^{T_i - t}}$ ▷ Recency-weighted contribution
end for
$\mathcal{S}_{\text{sorted}} \leftarrow$ clients in $\mathcal{S}$ sorted by R_i in descending order
$\mathcal{S}_{\text{top}} \leftarrow$ top $f_c \cdot |\mathcal{S}|$ clients from $\mathcal{S}_{\text{sorted}}$
$\mathcal{C}_{\text{remaining}} \leftarrow \mathcal{C} \setminus \mathcal{S}$
$\mathcal{C}_{\text{random}} \leftarrow$ random sample of size $(f \cdot |\mathcal{C}| - |\mathcal{S}_{\text{top}}|)$ from $\mathcal{C}_{\text{remaining}}$
$\mathcal{S}_{\text{new}} \leftarrow \mathcal{S}_{\text{top}} \cup \mathcal{C}_{\text{random}}$
return $\mathcal{S}_{\text{new}}$

4.3.3 Weight aggregation

Similar to the standard Federated Averaging (FedAvg) [67] approach, the aggregation process in RAFA combines the updates from selected clients using a weighted average. Given that the selection strategy provides us with a client-level score for each client in the selection, we can now leverage this metric as the weighting factor for the weighted averaging in which the global update is constructed based on the local updates from the clients in the selection at a

given round. Equation (4.1), outlines the construction of the global weight update, using the contribution score, J_i as the weighting factor. In Algorithm 4, a condensed overview is provided of the construction of the global update in the context of the FL procedure.

4.3.4 Adaptive weighting

In Section 4.3.2, we indicated that the weights α, β, and γ in J_i are of adaptive nature, which up until this point we did not yet address. In standard FL, relying on client selection strategies that require parameter input or hyperparameter tuning, is often not feasible. In real-world environments, institutions do often not have the time nor resources to run multiple iterations of a single FL strategy for the sole purpose of tuning the parameters that are inputs to the FL procedure itself. Therefore, in this work, we propose an adaptive weighting scheme that neatly aligns with the conceptual logic behind the contribution score and overall client selection procedure. More specifically, in each round we observe which of the three components contributed most to the global update and allow the strategy to shift the weights in favor of that component for the next round. Given that the top-performing clients are carried over to the next round, the logic is that for the next round we allow for a greater update in the area where those clients bring the most value. If in the subsequent round, a different component-client combination becomes informative the weights will be updated in favor of that observed shift for the next round.

Formally, we define this mechanism as follows; for each round, RAFA updates the weights α, β, and γ based on the relative performance improvement of the global model. Here, this performance improvement is measured using the respective model loss and is denoted as:

$$\Delta P = P_t - P_{t-1}, \tag{4.7}$$

where P_t represents the performance, in terms of the Loss of the global model, at round t, and similarly, P_{t-1} is the performance at the previous round.

To update the weights, we employ a gradient-based approach. Let L denote the performance loss, defined as the negative performance improvement. The gradients of the loss with respect to each of the weights, α, β, and γ, are:

$$\nabla_\alpha L = \frac{\partial L}{\partial \alpha}, \quad \nabla_\beta L = \frac{\partial L}{\partial \beta}, \quad \nabla_\gamma L = \frac{\partial L}{\partial \gamma} \tag{4.8}$$

Using a gradient descent-like approach, the weights are updated as follows:

Algorithm 4 Contribution-Based Weighted Aggregation

Require: w_g: Current global model parameters
Require: $\{\Delta w_i\}$: Local updates from selected clients
Require: $\{J_i\}$: Contribution scores of selected clients
1: **Server executes:**
2: Initialize aggregated updates $\Delta w \leftarrow \{\mathbf{0}\}_{l=1}^{L}$
3: Compute total contribution score:

$$J_{\text{total}} \leftarrow \sum_{i \in \text{selected clients}} J_i$$

4: **for** each layer $l = 1$ to L **do**
5: Compute the aggregated update for layer l:

$$\Delta w_l \leftarrow \sum_{i \in \text{selected clients}} \left(\frac{J_i}{J_{\text{total}}} \cdot \Delta w_{i,l} \right)$$

6: Update the global model parameters for layer l:

$$w_{g,l} \leftarrow w_{g,l} + \Delta w_l$$

7: **end for**
8: **return** Updated global model parameters w_g

$$\alpha \leftarrow \alpha - \eta \nabla_\alpha L, \quad \beta \leftarrow \beta - \eta \nabla_\beta L, \quad \gamma \leftarrow \gamma - \eta \nabla_\gamma L \tag{4.9}$$

where η is the learning rate that controls the magnitude of the weight updates. We approximate the gradients using the relative performance improvement of each component in the joint contribution scores. Therefore, we define three separate performance metrics for each component; P_H^t, P_S^t and P_I^t, where each of these metrics reflects the part of the overall performance improvement that can be attributed to each of the individual components. As such, the gradients are approximated as:

$$\begin{aligned} \nabla_\alpha L &\approx P_H^t - P_H^{t-1}, \\ \nabla_\beta L &\approx P_S^t - P_S^{t-1}, \\ \nabla_\gamma L &\approx P_I^t - P_I^{t-1} \end{aligned} \tag{4.10}$$

Intuitively, this allows RAFA to update α, β, and γ based on the relative effectiveness of each component in improving the global model according to the information retrieved from the latest round. In addition, to ensure that no component is neglected, a minimum contribution threshold is enforced for each weight. After updating the weights, we ensure that:

$$\alpha, \beta, \gamma \geq min_c \tag{4.11}$$

where min_c is a predefined lower bound. After enforcing the minimum contribution, the weights are normalized to ensure they sum to 1:

$$\alpha + \beta + \gamma = 1 \tag{4.12}$$

In doing so, we force the joint contribution score remains balanced across all components, with each component maintaining a minimum level of influence.

4.3.5 The RAFA pipeline

At this point, we discussed all elements of the RAFA strategy, being, client selection 4.3.2, weight aggregation 4.3.3 as well as adaptive weighting 4.3.4. In summary, the end-to-end strategy is comprised of (i) initialization of the global model as well as the weights α, β, γ, (ii) select clients based on the recency-weighted contribution score, (iii) client run a local round of training and share the update with the orchestrating server, (iv) the central server aggregates the local updates into the global model update in accordance with the contribution-based weighted aggregation, and (v) α, β, γ are dynamically updated based on the observed performance improvement at the model level. The full RAFA strategy is holistically represented in Algorithm 5.

4.4 Experimental setup

To assess the performance of RAFA as a Federated Learning strategy, we conduct experiments on the benchmark datasets MNIST [53] and CIFAR-10 [51], as well as on the real world FeTS MRI segmentation dataset [4, 77, 87]. For the former two, we compare performance of RAFA against the widely used FL stratetegies FedProx [93] and FedAvgM [43] in combination with different types of non-IID data. For the experiments on the real-world FeTS dataset, we adhere to the real distributions and compare performance with FedAvgM, which, given

Algorithm 5 RAFA Pipeline

Require: T: Number of communication rounds
Require: N: Total number of clients
Require: $\mathcal{F}$: Fraction of clients selected per round
Require: $\mathbf{w}_0$: Initial global model parameters
Require: α, β, γ: Initial adaptive weights
Ensure: Global model parameters $\mathbf{w}_T$

for $t = 1$ to T **do**
  if $t = 1$ **then**
    Randomly select $\mathcal{C}_{\text{selected}} \subseteq \{1, \ldots, N\}$ such that $|\mathcal{C}_{\text{selected}}| = \mathcal{F} \cdot N$
  else
    for each client $i \in \{1, \ldots, N\}$ **do**
      Compute score $J_i \leftarrow \alpha H_i + \beta S_i + \gamma I_i$
    end for
    $\mathcal{C}_{\text{top}} \leftarrow$ top fraction of clients based on J_i
    $\mathcal{C}_{\text{random}} \leftarrow$ remaining fraction of clients randomly
    $\mathcal{C}_{\text{selected}} \leftarrow \mathcal{C}_{\text{top}} \cup \mathcal{C}_{\text{random}}$
  end if
  for each client $i \in \mathcal{C}_{\text{selected}}$ **do**
    Client i performs local training and sends update $\Delta\mathbf{w}_i$ to server
  end for
  $\Delta\mathbf{w} \leftarrow \sum_{i \in \mathcal{C}_{\text{selected}}} \left(\frac{J_i}{\sum_{j \in \mathcal{C}_{\text{selected}}} J_j} \cdot \Delta\mathbf{w}_i \right)$
  Update global model: $\mathbf{w}_t \leftarrow \mathbf{w}_{t-1} + \Delta\mathbf{w}$
  Update H_i, S_i, and I_i for all clients
  Adjust adaptive weights α, β, γ
end for
  return $\mathbf{w}_T$

the momentum term has been shown to be an effective strategy for FeTS [43]. The remainder of this section describes the data, models, metrics and different settings used to evaluate RAFA.

4.4.1 Data

MNIST

The MNIST dataset consists of grayscale images of handwritten digits between 0 and 9, with a resolution of 28x28 pixels. It contains 60,000 training samples

and 10,000 test samples, each belonging to one of the 10 digit classes. MNIST often serves as a benchmark dataset in computer vision tasks, more specifically for classification. In Figure 4.1, the top row represents a set of randomly drawn entries from the MNIST dataset [53].

CIFAR-10

CIFAR-10 is a dataset containing 60,000 color images in 10 different classes, such as airplanes, cars, birds, cats, dogs, etc. Each image is of size 32x32 pixels and contains RGB color channels. The dataset is divided into 50,000 training images and 10,000 test images. Compared to MNIST, CIFAR-10 is a more complex benchmark dataset given the complexity of the images. The bottom row in Figure 4.1 represents a subset of 5 randomly drawn images from the CIFAR-10 dataset [51]. This depiction also shows the contrast in inherent complexity between the MNIST and CIFAR-10 dataset.

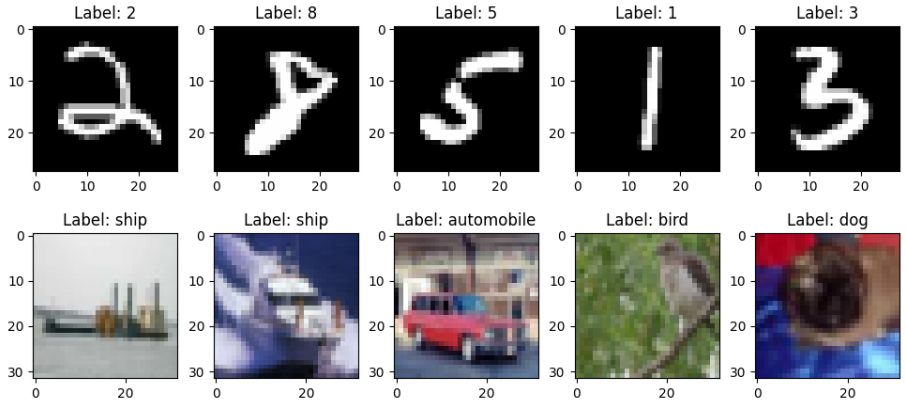


Figure 4.1: Five randomly selected images from MNIST (top row) as well as five randomly drawn images from CIFAR-10 (bottom row)

FeTS

The FeTS (Federated Tumor Segmentation) dataset comprises multi-institutional pre-operative multi-parametric MRI scans, focusing on glioblastoma (GBM) cases. The dataset includes four MRI modalities: native (T1), post-contrast T1-weighted (T1Gd), T2-weighted, and T2 Fluid-Attenuated Inversion Recovery (T2-FLAIR) [4, 77, 87]. These scans are available in NIfTI format and were collected from various institutions, which we will refer to as the originating

institutions in the remainder of this work. Apart from the raw image, the dataset contains ground truth annotations for each of the images in the set. The ground truth annotation, or segmentation mask, contains information on the enhancing tumor (ET), peritumoral edema (ED), and necrotic tumor core (NCR) [4, 77, 87]. These masks are the result of manual segmentation and validation by expert radiologists. An example for a single slice pertaining to a random MRI image in the dataset is shown in Figure 4.2.

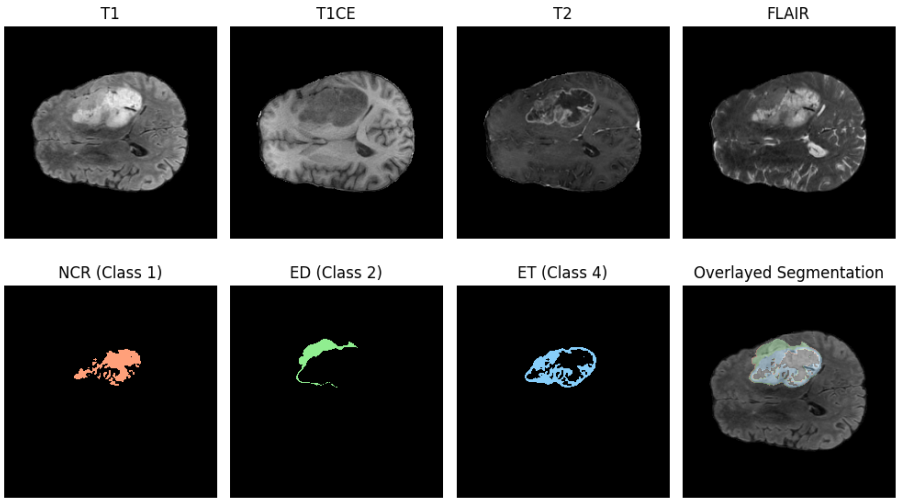


Figure 4.2: A single slice from the FeTS dataset with the 4 main modalities, the corresponding multi-label maks and the representative overlay of the mask on top of the ground truth T1 image.

The dataset is partitioned into train, validation, and test sets. As the dataset is part of a public challenge, the hold-out test set is not made publicly available. Therefore, we set aside the public validation set for testing and split the train set into a train and validation set. The data has undergone preprocessing, including co-registration to a common anatomical template, skull stripping, and interpolation to a resolution of 1 mm^3 prior to release in the public domain, therefore we do not define an additional preprocessing pipeline other than normalization for crafting the input to our models. For FeTS, we define the 3D segmentation task, meaning that we preserve the 3D nature of an MRI image throughout the network architecture. Each image is of shape 4x224x224x155, where the dimensions subsequently correspond to the 4 MRI modalities, the height, width and number of slices or depth. All 150 slices stacked on top of each other form the 3D representation of the MRI image. A 2D slice-by-slice

sequential representation with the ground truth annotations as overlay is shown in Figure 4.3.

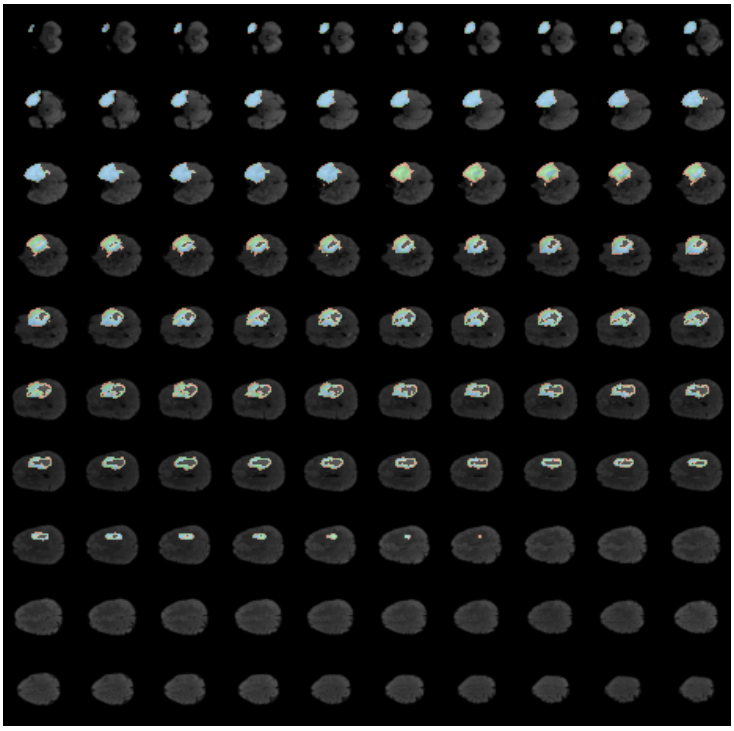


Figure 4.3: Raw T1 MRI scan with the segmentation mask overlay for 100 subsequent slices (slice 25 - 125).

4.4.2 Data distributions

To evaluate and compare the efficacy of RAFA against other FL strategies, we define different data distributions that artificially mimic different forms of IID and non-IID data. More specifically, we define one IID distribution, a linear partitioning, an exponential distribution as well as a Dirichlet based data distribution for the MNIST and CIFAR datasets. To do so, we rely on the partitioners available in the Flower framework [7], with exception of the exponential partitioner, for which we define a custom implementation. In the following paragraphs, we provide a high level description of the partitioners and the resulting data distribution.

IID Partitioner: The IID partitioner splits the dataset in a uniform manner, meaning that each client receives a random sample of the dataset, preserving the overall data distribution across clients. Intuitively, each client receives a well-balanced set of samples from all classes, such that the dataset characteristics are equally represented across the clients. This constitutes a baseline for this work in that it is not non-IID but fully IID, which is not representative of the real-world setting. Nevertheless, including the IID setting is relevant as a baseline.

Linear Partitioner: The linear partitioner generates a non-IID data distribution by sequentially partitioning the dataset in such a way that each client receives a continuous range of data points. Simply put, the number of samples assigned to a client is linear, and proportional to the extent of the full dataset, correlated with the client ID. The resulting data distribution over 100 clients for MNIST is shown in Figure 4.4.

Exponential Partitioner: The exponential partitioner aims to simulate the real-world non-IID distribution that is often seen in the healthcare domain where there is homogeneity in the feature space across clients but a small subset of the clients in the federation host most of the data. To that extent we defined a distribution that mimics exponential growth in the number of instances for the top 20 clients in the federation. A lower bound of 200 instances per client was enforced to ensure that each client has a minimal set of data available for training. Figure 4.4 depicts this tpye of distribution for the MNIST dataset with 100 partitions, corresponding to the 100 simulated clients.

Dirichlet Partitioner: The dirichlet partitioner [124] yields the most complex distribution in this work. By sampling from a Dirichlet distribution based on the label such that each client only hosts data that pertains to a subset of the classes. More specifically, this approach partitions the data such that there is heterogeneity in the label space across clients and a small portion of the clients host the majority of the the data. For MNIST and CIFAR. one client might predominantly receive samples from two or three classes, while others receive different subsets of classes. Here again, a visual representation of the resulting distribution is shown in Figure 4.4, for reference.

Contrary to the defined data partitions for MNIST and CIFAR-10, we do not define additional distributions for FeTS. In this work and the enclosed experiments, we adhere to the distributions as share by the data provider. To that extent, 2 main ways of partitioning the data are available. In the first, data is split into 23 partitions which correspond to the originating institution. In the second, additional partitions are artificially created, for a total of 33 partitions, and data is partitioned based on some, unknown, underlying characteristic of the data different from the originating institution. Both resulting distributions

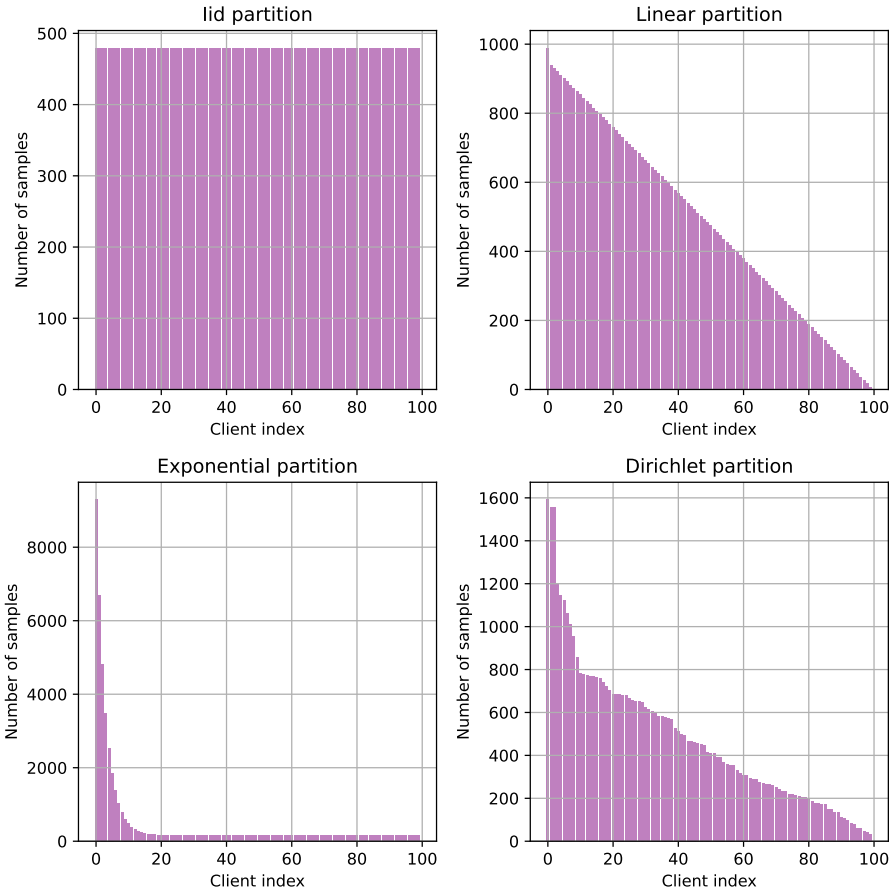


Figure 4.4: Overview of the data distributions used in this work for the MNIST and CIFAR-10 datasets. The plots show here reflect the MNIST distributions and absolute values. The same characteristics, however, hold for the CIFAR-10 distributions.

are shown in Figure 4.5, for reference.

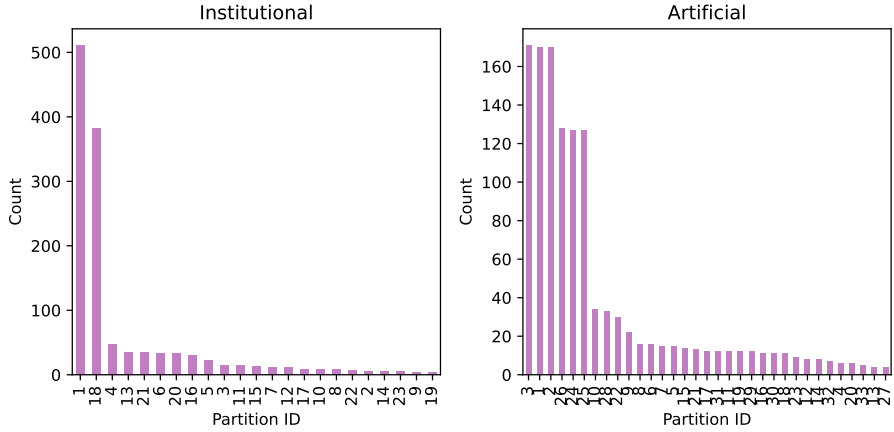


Figure 4.5: Overview of the institutional and artificial data partitioning as per the FeTS meta data.

4.4.3 Models

In this work, we define two main experimental settings to evaluate the performance of RAFA. The first of which relates to classification for the MNIST and CIFAR-10 data, whereas the second setting relates to the 3D image segmentation problem on the FeTS data. Both require vastly different modelling approaches, for the former we define a simple Convolutional Neural Network (CNN) [52], whereas for the latter we leverage the 3D attention U-Net architecture [46]. Both models are described in more depth in the subsequent paragraphs.

CNN

For MNIST and CIFAR-10, the objective is to classify input images into predefined categories based. For the MNIST dataset, the task involves classifying grayscale images of handwritten digits into one of 10 classes (0-9). Formally, given an input image $\mathbf{x} \in \mathbb{R}^{1\times28\times28}$, the task is to assign it a label $y \in \{0, 1, ..., 9\}$. For CIFAR-10, the classification task remains the same, however, the input consists of color images $\mathbf{x} \in \mathbb{R}^{3\times32\times32}$ from 10 different categories, and the goal is to map each image to one of the 10 class labels $y \in \{0, 1, ..., 9\}$.

For this specific classification task, a simple but performant convolutional neural network (CNN) is defined, consisting of two convolutional layers followed by three fully connected layers. The model is defined such that an image $\mathbf{x}$ first passes through a convolutional layer, after which a ReLU activation function is applied, followed by max-pooling which results in the first feature map representation. The second convolutional layer applies a similar structure, for which the output is subsequently flattened into a vector, $\mathbf{z}_2''$, which serves as the input to the fully connected layers. These layers perform further transformations through linear mappings and ReLU activations. The eventual output is then passed through a softmax function to predict the class probabilities. A detailed overview of the CNN architecture adopted in this work can be found in Table 4.1.

Conceptually, the CNN architecture allows for hierarchical feature learning, where lower layers capture local patterns like edges, and deeper layers learn more abstract representations. The fully connected layers then allow the model to map the learned features to the final dimension which corresponds to the classification task.

Table 4.1: CNN Architecture for MNIST and CIFAR-10 Classification

Layer	**Type**	**Kernel/Units**	**Output Size**
Input	-	$1 \times 28 \times 28$	$1 \times 28 \times 28$
Conv1	Conv2D	6 filters, 5×5	$6 \times 24 \times 24$
ReLU	Activation	-	$6 \times 24 \times 24$
Pool1	MaxPool2D	2×2	$6 \times 12 \times 12$
Conv2	Conv2D	16 filters, 5×5	$16 \times 8 \times 8$
ReLU	Activation	-	$16 \times 8 \times 8$
Pool2	MaxPool2D	2×2	$16 \times 4 \times 4$
Flatten	Reshape	-	256
FC1	Linear	120 units	120
ReLU	Activation	-	120
FC2	Linear	84 units	84
ReLU	Activation	-	84
FC3	Linear	10 classes	10

3D Attention UNET

The second model in this work is fully tailored to the segmentation task for the FeTS dataset. More specifically, the main objective is to perform 3D image segmentation, where we aim to classify each voxel (3D pixel) of the input MRI

scans into different tissue types or tumor regions as annotated in the ground truth masks. To that end, we define a model architecture that is aimed at predicting the segmentation mask from volumetric input data. To do so, we rely on the U-Net architecture which is further extended to allow for volumetric input and attention mechanisms. The 3D Attention U-Net in this work is based on the model proposed in [46], which in itself is based on the standard 3D U-Net [115]. The key innovation comes from the integration of attention mechanisms that allow the network to focus on the most informative regions of the feature maps.

The architecture follows a typical encoder-decoder structure, where the encoder reduces the spatial resolution through 3D convolutional layers and pooling operations, progressively capturing more abstract features. The bottleneck layer represents the most condensed form of the data, processing the abstract and global features. The decoder restores the image resolution through transposed convolutions while integrating attention blocks. These blocks allow the model to focus on the most informative regions of the feature maps. This is done through computation of the gating signal, which enhances the flow of relevant information through the model. Skip connections ensure that details from the encoder are passed directly to the decoder, which adds to the models capability to precisely delineate boundaries. The final output layer produces the segmentation maps, with a 3D convolution reducing the number of channels to match the number of segmentation classes. Table 4.2, provides a detailed overview of the different building blocks of the architecture adopted for the 3D segmentation task in this work.

4.4.4 Metrics

To evaluate model performance we use a combination of standard classification and segmentation metrics which provide a insights with respect to the predictive accuracy, segmentation accuracy, and computational efficiency.

Given that this work evaluates RAFA in the benchmark classification setting as well as the real-world 3D segmentation task, the metrics are different for each of these settings. For the classification tasks, the main metric is accuracy, which measures the proportion of correctly classified instances. More specfically, given a set of true labels y and corresponding predictions $\hat{y}$, the accuracy is formally defined as:

$$\text{Accuracy} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(\hat{y}_i = y_i)$$

where n is the total number of samples and $\mathbb{1}$ is the indicator function, which returns 1 if the prediction is correct, and 0 otherwise, this notation is equivalent

Table 4.2: Attention U-Net 3D Architecture Overview

Layer	**Input Size**	**Output Size**	**Operation**
Conv3D + ReLU	$64 \times 240 \times 240 \times 155$	$64 \times 240 \times 240 \times 155$	$3 \times 3 \times 3$ convolution
MaxPool3D	$64 \times 240 \times 240 \times 155$	$64 \times 120 \times 120 \times 77$	Max pooling $2 \times 2 \times 2$
Conv3D + ReLU	$128 \times 120 \times 120 \times 77$	$128 \times 120 \times 120 \times 77$	$3 \times 3 \times 3$ convolution
MaxPool3D	$128 \times 120 \times 120 \times 77$	$128 \times 60 \times 60 \times 38$	Max pooling $2 \times 2 \times 2$
Conv3D + ReLU	$256 \times 60 \times 60 \times 38$	$256 \times 60 \times 60 \times 38$	$3 \times 3 \times 3$ convolution
MaxPool3D	$256 \times 60 \times 60 \times 38$	$256 \times 30 \times 30 \times 19$	Max pooling $2 \times 2 \times 2$
Conv3D + ReLU	$512 \times 30 \times 30 \times 19$	$512 \times 30 \times 30 \times 19$	$3 \times 3 \times 3$ convolution
MaxPool3D	$512 \times 30 \times 30 \times 19$	$512 \times 15 \times 15 \times 9$	Max pooling $2 \times 2 \times 2$
Bottleneck	$1024 \times 15 \times 15 \times 9$	$1024 \times 15 \times 15 \times 9$	$3 \times 3 \times 3$ convolution
Attention Block	$512 \times 30 \times 30 \times 19$	$512 \times 30 \times 30 \times 19$	Gated Attention
ConvTranspose3D + Concat	$512 \times 30 \times 30 \times 19$	$512 \times 60 \times 60 \times 38$	Upsampling + Skip Connections
Output	$3 \times 240 \times 240 \times 155$	$3 \times 240 \times 240 \times 155$	$1 \times 1 \times 1$ convolution

to expressing accuracy as the fraction between correct predictions ($TP + TN$), and total number of predictions ($TP + TN + FP + FN$).

For the 3D medical image segmentation task, accuracy is no longer a suitable metric or proxy for the overall model performance, therefore, we use the Dice coefficient [131], which measures the overlap between the predicted segmentation mask P and the ground truth G. Formally, the Dice coefficient is defined as:

$$\text{Dice}(P, G) = \frac{2|P \cap G|}{|P| + |G|}$$

where $|P|$ and $|G|$ represent the size of the predicted and ground-truth sets, respectively.

In addition to the Dice coefficient as a general metric for evaluation, we use Dice loss as the loss function to optimize the model during training which is formally defined as:

$$\text{Dice Loss}(P, G) = 1 - \text{Dice}(P, G)$$

Intuitively, using Dice loss penalizes the dissimilarity between the predicted and ground-truth masks, driving the model to improve segmentation accuracy.

As an additional ad-hoc metric for performance evaluation, we employ the Hausdorff distance [9], which calculates the maximum discrepancy between the boundaries of the predicted and ground-truth segmentations, and is therefore a proxy to the spatial accuracy of the segmentation model. Formally, the Hausdorff distance between two sets of points P and G is defined as:

$$H(P, G) = \max \left\{ \sup_{p \in P} \inf_{g \in G} \|p - g\|, \sup_{g \in G} \inf_{p \in P} \|p - g\| \right\}$$

where $\| \cdot \|$ denotes the Euclidean distance between points. Intuitively, the Hausdorff distance captures the largest boundary mismatch between the prediction and the ground truth. The Hausdorff distance is, however, known to be sensitive to outliers, therefore, we rely on the more robust version, the 95th percentile Hausdorff distance (H_{95}), which focuses on reducing the influence of outliers by computing the 95th percentile of the boundary distances as follows:

$$H_{95}(P, G) = \text{percentile}_{95} \left\{ \inf_{g \in G} \|p - g\| : p \in P \right\}$$

This in turn yields a more stable proxy for the spatial boundary precision of the evaluated segmentation model.

4.4.5 Federated learning

In our experiments, we simulate a federated learning environment to assess the performance of different FL strategies for the MNIST, CIFAR-10 and FeTS datasets using the Flower framework. In what follows, we outline the experimental FL setup for each dataset and the federated learning strategies employed.

MNIST and CIFAR-10 Setup

For both MNIST and CIFAR-10, we simulate a federated learning scenario with 100 clients, each receiving a partition of the dataset in line with the predefined partitions as shown in Figure 4.4. We fix the following environment variables over all experiments: (i) for each round, a subset of 10% of clients participates, the constitution of this subset depends fully on the employed FL strategy. (ii) The models are trained for 30 communication rounds, (iii) with each client performing one local epoch per round. At last, (iv) this process is repeated 10 times to compute the standard deviation of the results.

Furthermore, we do we do not perform hyperparamter optimization but rely on a set of predetermined parameter settings given that the goal here is not to perform better than SOTA, but rather to show and compare performance of RAFA against other FL strategies. To that extent, we use the Adam optimizer with a learning rate of 0.001 and the binary cross-entropy loss function to train the models. The models are evaluated using accuracy as the primary performance metric, with the standard deviation being reported over the 10 iterations.

We compare the following federated learning strategies:

- **FedProx:** An extension of FedAvg, which introduces a proximal term in the local objective to address heterogeneity in client data distributions [93].
- **FedAvgM:** Another variant of FedAvg that incorporates momentum into the update process, accelerating convergence and improving model generalization in non-IID settings [43].
- **RAFA:** Our proposed strategy that dynamically adjusts client selection and utilizes adaptive weighting based on a performance indicator to enhance model performance and robustness.

FeTS setup

Contrary to the FL settings for MNIST and CIFAR-10, we do not longer simulate an arbitrary number of 100 clients, for FeTS, however, we adhere to the number of clients in the real-world and artificial partitions as defined by the data providers. Here, training is terminated after 20 rounds of communication for which each client locally trains the model for 2 subsequent epochs. For each round, 10% of the clients participate to training, the constitution of that subset of clients is again fully dependent on the FL strategy. Given the computational complexity, only 5 training iterations are performed to obtain average performance and the resulting standard deviations.

For FETS, we compare performance of RAFA against the FedAvgM strategy, which has been shown to be effective in the FL MRI segmentation setting. Table 4.3, provides a condensed overview of the settings with repsect to each of the different datasets in this work.

Table 4.3: Experimental Setup Overview

Dataset	Clients	Rounds	Local Epochs	Iterations
MNIST	100	30	1	10
CIFAR-10	100	30	1	10
FeTS	23/33	20	2	5

This setup ensures consistency in training across datasets, while allowing us to explore the differences in federated learning strategies under varying data complexities and model architectures.

4.5 Results

This section presents the performance of the federated learning strategies FedAvgM, FedProx, and RAFA on the MNIST and CIFAR-10 datasets across different data distribution scenarios (IID, Linear, Exponential, and Dirichlet). Subsequently, we also report the performance on the 3D segmentation task using the FeTS dataset under institutional and artificial data partitions. As indicated in 4.4.4, MNIST and CIFAR-10 are evaluated mainly based on accuracy, however, we also report loss as well as the time complexity (τ) in seconds. For FeTS, the performance is evaluated by means of the Dice score, Dice loss, and Hausdorff 95.

From the results for MNIST and CIFAR-10, as shown in Table 4.4, we observe that across all datasets and scenarios, RAFA consistently demonstrates

Table 4.4: Performance Metrics for Federated Learning Strategies on MNIST and CIFAR-10 Datasets

Data Distribution	Dataset	Strategy	Accuracy (%)	Loss	τ (s)
IID	MNIST	FedAvgM	96.31 ± 0.58	0.66 ± 0.09	245.83 ± 1.16
		FedProx	95.90 ± 0.38	0.57 ± 0.10	242.63 ± 1.83
		RAFA	95.35 ± 0.35	0.59 ± 0.01	356.73 ± 133.89
	CIFAR-10	FedAvgM	35.56 ± 5.31	6.83 ± 0.35	281.73 ± 4.23
		FedProx	33.21 ± 6.42	7.21 ± 0.67	279.62 ± 0.90
		RAFA	33.46 ± 4.61	8.04 ± 1.67	270.82 ± 15.54
Linear	MNIST	FedAvgM	95.04 ± 0.13	0.45 ± 0.09	238.77 ± 0.76
		FedProx	96.39 ± 1.22	0.69 ± 0.18	241.90 ± 0.87
		RAFA	95.90 ± 0.71	0.79 ± 0.10	233.51 ± 15.79
	CIFAR-10	FedAvgM	37.35 ± 0.35	8.33 ± 0.71	364.41 ± 12.13
		FedProx	10.00 ± 0.00	11.35 ± 0.81	328.90 ± 33.36
		RAFA	37.18 ± 0.21	7.50 ± 0.13	264.75 ± 17.79
Exponential	MNIST	FedAvgM	98.39 ± 0.58	0.79 ± 0.50	262.68 ± 2.40
		FedProx	98.38 ± 0.00	0.85 ± 0.25	254.92 ± 0.28
		RAFA	97.82 ± 0.51	0.24 ± 0.01	275.48 ± 40.30
	CIFAR-10	FedAvgM	39.09 ± 1.50	30.12 ± 0.70	297.85 ± 5.81
		FedProx	36.98 ± 8.84	24.51 ± 1.39	287.30 ± 0.04
		RAFA	41.96 ± 1.22	24.46 ± 0.39	327.16 ± 17.50
Dirichlet	MNIST	FedAvgM	95.76 ± 0.06	1.39 ± 0.26	245.79 ± 0.74
		FedProx	93.54 ± 0.69	1.33 ± 0.76	240.83 ± 1.10
		RAFA	95.20 ± 0.07	1.66 ± 0.21	236.33 ± 11.46
	CIFAR-10	FedAvgM	20.99 ± 5.33	12.69 ± 4.90	275.73 ± 2.75
		FedProx	19.16 ± 0.38	10.86 ± 5.85	277.39 ± 1.07
		RAFA	21.02 ± 1.76	12.71 ± 0.86	276.65 ± 9.67

competitive performance, on several occasions RAFA outperforms the other methods or at least performs on par in terms of each of the reported metrics. Given that the exponential setting provides the closest proxy to the real-world medical data partitions, it is particularly interesting to observe that under the Exponential distribution on MNIST, RAFA achieved a high accuracy of 97.82% with minimal distributed loss (0.24), surpassing both FedAvgM and FedProx in terms of stability whilst yielding similar performance in terms of accuracy.

For the more complex CIFAR-10 dataset, the performance for RAFA is mostly on par with FedAvgM and FedProx, showing similar accuracy in the Linear and Exponential settings. However, RAFA provides more stable results across iteration under the Exponential distribution, where it yields higher accuracy (41.96%) and shows competitive performance with FedAvgM. These findings are visually represented in Figure 4.6.

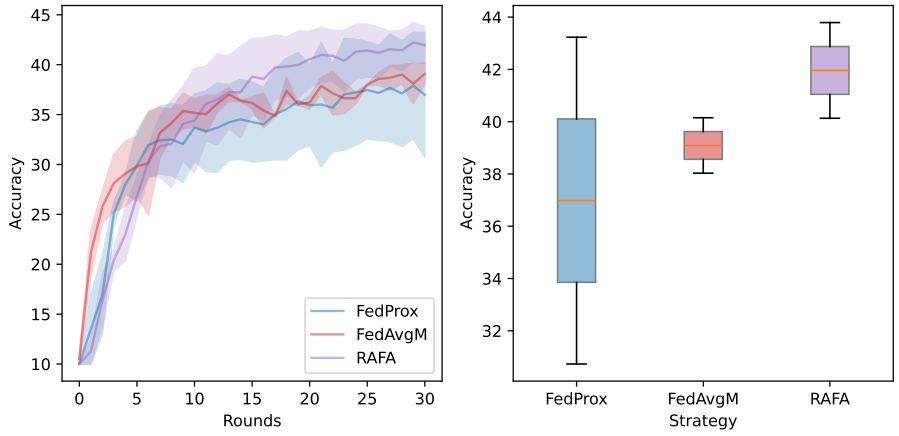


Figure 4.6: Results for the three FL strategies after 30 rounds of training on the exponential datasplit over 100 clients for the CIFAR-10 dataset. The image shows the training curves in terms of accuracies on the left whereas the resulting performance after 30 rounds is shown on the right along with the achieved standard deviation.

In Table 4.5, the results for the 3D segmentation task on the FeTS dataset are provided, RAFA consistently outperforms FedAvgM for both institutional and artificial data distributions. Under the institutional distribution, RAFA achieves a higher Dice score (79.41%) and lower Dice loss (12.91) compared to FedAvgM (78.21% Dice score and 13.73 Dice loss), along with a lower Hausdorff 95 value (17.83 vs. 20.58). A similar trend can be observed for the artificial

Table 4.5: Performance Metrics for 3D Segmentation on Test Set with Institutional and Artificial Data Distributions

Split	Strategy	Dice Score	Dice Loss	Hausdorff 95
Inst.	FedAvgM	78.21 ± 0.76	13.73 ± 1.31	20.58 ± 0.52
	RAFA	79.41 ± 0.43	12.91 ± 1.22	17.83 ± 0.23
Art.	FedAvgM	73.85 ± 0.92	16.83 ± 1.92	23.66 ± 0.56
	RAFA	75.01 ± 0.73	14.94 ± 1.68	20.38 ± 0.44

data distribution, where RAFA outperforms FedAvgM with a Dice score of 75.01% and a lower Dice loss of 14.94.

4.6 Discussion

Based on the results provided in Section 4.5, we observe that RAFA is a robust and stable strategy across different datasets and distribution settings, providing competitive performance in federated learning tasks when compared to other, widely used FL strategies. Under robust and stable we specifically understand that RAFA yields comparable predictive performance with less variance in the standard deviation on the error, which in turn showcases more stable training over iterations. On simpler datasets like MNIST, RAFA consistently yields high accuracy and low distributed loss. More important, is how under challenging distributions such as the Exponential and Dirichlet cases, RAFA is able to yield good predictive performance with less variance in terms of standard deviation compared to FedAvgM and FedProx. Especially in the non-IID setting, this is a desirable characteristic. Intuiatively, this means that RAFA, as an FL strategy, provides more stable and reliable results over iterations.

With respect to the 3D segmentation task on FeTS, RAFA consistently outperformed FedAvgM in both the institutional and artificial data distributions, achieving higher Dice scores and lower Dice losses. Additionally, the lower Hausdorff 95 value achieved by RAFA demonstrates its ability to provide more precise segmentation results. In addition, it does so with lower standard deviations on the results overall, meaning that, when selecting a strategy in the real-world, RAFA would be a more suitable option compared to FedAvgM and FedProx, at least for this specific use case. This in turn, highlights the strength of RAFA in complex, real-world applications.

Overall, RAFA proves to be a reliable and competitive federated learning strategy, offering comparable, or better, performance in highly distributed environments characterized by non-IID data.

4.7 Conclusion

Federated learning has become increasingly important for training models in distributed environments where data sensitivity poses a risk. However, the challenge of non-IID data remains an open problem in terms of obtaining optimal performance and stable convergence for the resulting model. Existing approaches, such as FedAvg and its variants, including FedAvgM and FedProx, have been proposed to mitigate the effects of data heterogeneity but still face challenges in scenarios with extreme cases of non-IID data.

In this work, we introduce RAFA (Robust Adaptive Federated Aggregation), a novel method that aims to address these gaps in federated learning with non-IID data. RAFA extends existing methodologies by employing an adaptive weighting mechanism that dynamically adjusts the influence of client updates based on the relevance of their local model to the global objective.

The proposed method is evaluated on several datasets, including standard benchmarks such as MNIST and CIFAR-10, as well as the real-world FeTS medical segmentation dataset. In all settings, RAFA shows competitive or better performance compared to FedAvgM and FedProx. We show how RAFA provides more consistent results over iterations as observed by the standard deviations for accuracy and loss across rounds.

Even though RAFA shows promising results, the improvements are only marginal, to that extent future work will focus on extending the RAFA strategy with client clustering approaches to further enhance performance and robustness. We note that RAFA performs well in the non-IID settings defined in this work, however, further research is required to evaluate and finetune the performance of RAFA in settings with extreme label imbalance.

In conclusion, RAFA addresses the key challenges of non-IID data distributions, as typically observed in the medical setting, with respect to federated learning by incorporating an adaptive and robust aggregation mechanism and dynamic client selection strategy. We show how this approach consistently yields, robust results in terms of predictive performance under different settings in terms of data partitions and deep learning tasks, across different datasets.

4.8 Supplementary context

This section provides additional context for the RAFA framework, addressing the theoretical justification for key design choices, the stability of the adaptive weighting mechanism, and the relationship to alternative approaches.

4.8.1 Design choices

Several design choices in the RAFA framework, including the use of cosine similarity for the informativeness score, weight decay in the adaptive mechanism, and the specific formulation of the contribution score, were guided by empirical experimentation rather than derived from first principles. We provide explicit motivation for these choices here and discuss their limitations.

We chose cosine similarity for the informativeness component because it measures directional alignment between the updates of a given client and the current global update direction, independent of magnitude. In the federated setting, update magnitudes vary considerably across clients due to differences in local dataset size and number of local gradient steps. Cosine similarity normalizes for this variation, focusing on whether a client is pushing in the right direction. However, this choice has known limitations. First, it is biased towards small updates from clients that are already well-aligned with the global model, potentially undervaluing large, novel updates from clients with genuinely different data. Second, cosine similarity is numerically unstable for near-zero weight vectors, which can occur early in training or for parameters that have not yet been meaningfully updated. The initialization strategy and the minimum-magnitude clamping in the implementation mitigate this risk in practice. Third, by rewarding alignment with the current global direction, cosine similarity may suppress useful disagreement: a client whose data reveals a genuine limitation of the current model would produce updates that are directionally misaligned, and would therefore receive a low informativeness score even though its contribution could improve the global model. This is a fundamental limitation of alignment-based metrics and motivates future exploration of disagreement-aware scoring.

We apply weight decay to the adaptive parameters (α, β, γ) to prevent them from growing without bound across rounds. Without weight decay, a parameter that is beneficial in early rounds could accumulate a disproportionately large value, making the system slow to adapt when the training dynamics change. The decay rate was set empirically. Alternatives such as projection onto a simplex (constraining $\alpha + \beta + \gamma = 1$) or periodic resetting were considered but not adopted due to the added complexity.

4.8.2 Interaction between α, β, and γ

The three adaptive weight parameters α (significance/magnitude), β (informativeness/alignment), and γ (historical contribution) are updated independently using a gradient-like procedure based on the performance change attributable to each component. A valid concern is that these parameters are highly interactive:

the gradient of the loss with respect to α is not independent of the current values of β and γ, because the contribution score is a multiplicative function of all three components. Updating each parameter as if it were independent can lead to oscillatory behavior, where increasing α reduces the effective importance of β, which then triggers a compensating increase in β in the next round.

The gradient approximation used equates the finite difference of a performance metric over training rounds with the gradient of the loss. This approximation is valid only if the performance metric is smoothly related to the loss and the time steps are sufficiently small. In practice, neither condition is guaranteed, particularly under non-stationary client participation. The empirical results suggest that the adaptive mechanism is sufficiently stable for the datasets studied, but the lack of formal convergence guarantees is a limitation that should be addressed in future theoretical work. A more principled approach could treat the weight selection as a bandit problem or employ Bayesian optimization with appropriate priors, which would naturally handle the interaction between parameters.

4.8.3 Stability versus reactivity

A key concern for any adaptive system operating on healthcare data is whether it is stable (responding to genuine shifts in the data landscape) or merely reactive (amplifying noise). Healthcare data can exhibit short-term fluctuations due to seasonal effects, staffing changes, or data quality issues that do not represent meaningful distributional shifts. An overly reactive adaptive mechanism could chase these fluctuations, leading to volatile client weighting and potentially degrading model stability.

RAFA's recency-weighted historical contribution provides a degree of inertia that guards against pure reactivity: the exponential decay factor λ means that recent contributions are weighted more heavily but historical performance still exerts influence. The carry-over mechanism (retaining the top 50% of clients between rounds while injecting 50% new random clients) further stabilizes the process by ensuring continuity. However, the current implementation does not include explicit stability diagnostics. For deployment, we recommend logging the following signals at each training round: (i) the α, β, γ values over time to detect oscillatory behavior; (ii) the client rankings to detect pathological patterns such as a fixed set of clients dominating indefinitely; (iii) the variance of the global model's validation performance over a rolling window to detect instability. If oscillation is detected, the learning rate η for the adaptive weight updates should be reduced, or the weight decay increased.

4.8.4 Significance versus informativeness

The contribution score in RAFA consists of three components: significance (magnitude of the update), informativeness (directional alignment with the global update), and historical contribution. These components may overlap conceptually: a client that produces large, well-aligned updates will score highly on both significance and informativeness, potentially double-counting its contribution. Conversely, a client with a large but misaligned update will score high on significance but low on informativeness, which is the intended behavior for identifying clients that are different but not necessarily harmful.

The partial overlap between significance and informativeness is a design feature: significance captures how much a client has to contribute (magnitude), while informativeness captures whether that contribution is useful in the current training context (direction). A client with a large magnitude in an unhelpful direction should be downweighted, which the multiplicative combination achieves. That said, the three-component formulation introduces more flexibility than may be strictly necessary, and a simpler two-component formulation (e.g., combining magnitude and alignment into a single projected-gradient score) could achieve comparable results with fewer adaptive parameters. We consider this simplification a promising direction for future work.

4.8.5 Relationship to reinforcement learning

The dynamic client selection problem in RAFA bears resemblance to the multi-armed bandit problem in reinforcement learning: the server must choose which clients to pull (select for participation) based on uncertain estimates of their value, balancing exploitation (selecting known good clients) with exploration (trying new clients to discover potentially better contributors). The carry-over mechanism combined with random injection implements a simple ε-greedy-like strategy. More sophisticated bandit algorithms such as UCB (Upper Confidence Bound) or Thompson Sampling could improve the exploration–exploitation trade-off by maintaining uncertainty estimates for each client's contribution value.

4.8.6 Handling new clients

In a real-world deployment, new institutions may join the federation after the initial training has commenced. RAFA accommodates this naturally: a new client enters the random selection pool and will be evaluated alongside existing clients. However, because the new client has no historical contribution, its

initial γ component is zero, which may disadvantage it relative to established clients with accumulated history. To mitigate this cold-start problem, new clients could be given a grace period during which they are always included in the selection, or their historical score could be initialized to the median of existing clients. The current implementation does not address this explicitly, and we note it as a practical consideration for deployment.

4.8.7 Exponential smoothing

The recency-weighted historical contribution in RAFA uses a fixed exponential decay factor λ applied to all past rounds. This is conceptually similar to exponential smoothing in time-series analysis, where recent observations receive higher weight. An alternative to the fixed decay would be to let the lookback window adapt to the client's participation pattern: clients that participate frequently could have a shorter effective memory (since their recent behavior is well-characterized), while clients that participate sporadically could retain longer memory to compensate for the sparse signal. This adaptive-window approach would be analogous to event-driven models, where the excitation kernel determines how long past events influence the present. Implementing such a mechanism would require tracking per-client participation frequencies and adjusting λ accordingly, which adds complexity but could improve robustness in heterogeneous participation settings.

Chapter 5

From Pixels to Insights: Leveraging Class-Dependent Activation Maps for Augmented Medical Image Captioning

This chapter is submitted for publication as:

Scheltjens, V., Verbeke, W., De Moor, B. (2025). **From Pixels to Insights: Leveraging Class-Dependent Activation Maps for Augmented Medical Image Captioning**. *Submitted at BMC Medical Informatics and Decision Making*

Paper abstract:

Background: In medical imaging, radiologists face an increasing workload as a direct result of the growing number of medical images that require interpretation and reporting, particularly for Chest X-Rays. Consequently, automated medical image captioning has gained importance in recent years, driven by advances in vision-language architectures and underlying large language models. Existing approaches typically rely on raw image inputs combined with feature extraction pipelines to encode the relevant information needed for report generation.

However, these pipelines often fail to offer transparency into the internal mechanisms of the model, thereby limiting clinical adoption.

Methods: This work proposes an end-to-end, two-stage, learning approach in which during the first stage, a multi-label classifier is trained to predict pathological observations. The predictions from this classifier are then used to extract Gradient-weighted Class Activation Maps (Grad-CAM) corresponding to the top-k most likely pathologies for each input image. These saliency maps are subsequently fused with the original 3-channel CXR to construct a $(3 + k)$-channel augmented input representation that encodes both raw pixel-level information and class-specific activation features. This augmented image is then passed to the second stage of the pipeline: a vision encoder-decoder captioning model, to generate radiology reports. By doing so, we do not rely solely on learned feature representations within the encoder, but instead explicitly provide targeted semantic information regarding disease presence into the model input, offering a form of guided supervision. Additionally, we define the end-to-end approach in the Federated Learning setting to (i) showcase the feasibility complex pipelines in the federated, medical setting and (ii) mitigate limitations that would otherwise result from data sharing and aggregation limitations in practice.

Results: The full pipeline, implemented and evaluated on the MIMIC-CXR dataset, demonstrates that this augmentation strategy leads to captioning performance on-par with the SOTA, with the added benefit that the feature space is tractable and interpretable end-to-end, whilst mitigating challenges related to data-privacy restrictions that would otherwise occur. More specifically, the results show consistent and on-par performance in caption quality as measured by CIDEr, BLEU, and ROUGE metrics.

Conclusion: This work proposes and validates an interpretable, two-stage federated learning framework for medical image captioning, showing that enriching visual input with class-dependent information results in on-par predictive performance and enhanced model transparency.

5.1 Introduction

Recent advances in Artificial Intelligence (AI) in healthcare have shown significant promise, especially in the field of radiology. Radiologists are faced with an increasing amount of medical images that need interpretation and reporting, a task that is time-consuming. Chest X-rays (CXR) are the most commonly used imaging techniques for diagnosis of lung diseases. Interpretation

of CXR images requires significant expertise and, as stated, is time-consuming for the radiologists. Therefore, radiologists often represent a bottleneck in a patients care trajectory as they are faced with vast amounts of images that need interpretation, and annotation. To this extent, AI-driven clinical decision support systems (CDSS) aim to address these challenges by automating the analysis and report generation, thereby reducing the workload faced by radiologists which in turn can improve efficiency.

Automated medical image captioning has emerged as a tool for the reporting of radiological imaging. As the volume and complexity of medical imaging data continue to grow, the workload radiologists face increases as well. Therefore, research and industry are shifting to the development of automated systems that can provide descriptive insights from these images. Radiology reports, traditionally relying on textual descriptions provided by radiologists, often serve as the foundation for the clinical decision-making processes. However, manual generation of these reports is time-consuming, labor-intensive, and prone to variability [32, 88, 103].

Significant advances in the image captioning domain have been achieved [105, 126–128]. Driven by this work, the medical field has started to devote more research efforts to medical image captioning. By learning to generate natural language descriptions directly from visual inputs, these models and techniques are suited to, at least partially, automate the radiology reporting process. However, the adoption of image captioning models in the domain of medical imaging presents several challenges. Medical images are complex in nature, show different anatomical structures and pathological observations. Therefore, a deeper understanding of the underlying visual features to generate accurate and clinically relevant captions is needed [99]. Transformers, initially introduced for natural language processing (NLP) tasks, have shown promising results in text generation applications [108]. Specifically, models such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) have set new benchmarks in various NLP tasks, including text generation, summarization, and translation [22, 82]. These advances have contributed to the development of the Show, Attend, Tell approaches where the models provide a textual, descriptive output corresponding to the provided visual cues as input. In the medical domain, the application of transformer-based models to medical image analysis and report generation has shown promising results when it comes to in AI-driven CDSS.

Despite these advancements, a critical aspect of deploying AI in clinical settings is ensuring the interpretability and transparency of the models. Current results are promising, but to further pave the road to widespread adoption of AI-driven CDSS all end-to-end pipelines should incorporate explainability methods, such as Gradient-weighted Class Activation Mapping (Grad-CAM), which provides

insights into which regions of the image contribute to the model's predictions [96]. By incorporating these techniques across the board, we can bridge the gap between AI-driven predictions and clinical practice, making the models more trustworthy and reliable for healthcare practitioners.

Another challenge relates to data availability. Medical institutions remain reluctant to data sharing initiatives in light of patient privacy. The AI-driven CDSS we discuss here, benefit from a larger pool of available data at training time, hence, the challenge. Medical institutions would benefit from data sharing initiatives as it would aid predictive power of the resulting models, however, these same institutions are reluctant to share data to protect patient privacy. As in this work we aim to propose a valid end-to-end framework, we tackle this challenge by introducing the Federated Learning (FL) paradigm [67]. By doing so, we alleviate concerns and issues related to data privacy, and show the feasibility of adopting the FL framework in the medical setting for complex learning pipelines.

In this work, we propose an end-to-end, two-stage federated decision support system that leverages class-dependent activation maps (Grad-CAM) to enhance the interpretability of CXR report generation pipeline at all stages. The proposed approach involves augmenting the input image with Grad-CAM activation maps as additional channels, serving as the input to the image captioning model.

We show that, by explicitly encoding class-dependent information into the input space for the text generation methods, our approach can yields captions on-par with the SOTA with the added benefit that the proposed pipeline becomes more interpretable over all stages and mitigates data-privacy concerns through the federated setup. The aim of this work is to work towards models and more complex approaches that result in better, easier adoption in practice, bridging the gap between the data science and clinical practice.

5.2 Related work

The latest advances in generative modeling have sparked research efforts and a larger adoption rate of AI-driven applications in healthcare. Especially in medical imaging, the advantages are clear, nevertheless stringent regulations that aim to regulate usage of AI in privacy sensitive environments increasingly require for models to be interpretable across the board. In this section an overview of the literature with respect to the different components of the end-to-end, AI-driven CDSS pipeline is provided within the scope of CXR classification and report generation. Specifically, we review the latest work related to CXR

classification, vision encoder-decoder architectures, text generation models, and explainability in AI.

5.2.1 CXR classification

CXR classification involves using deep learning models to identify and classify abnormalities in chest X-ray images. One of the seminal works related to CXR classification was presented by Rajpurkar et al., in their work they proposed the CheXNet model [86], which utilized a 121-layer DenseNet architecture to achieve radiologist-level performance in detecting pneumonia. The model was trained on the NIH Chest X-ray dataset, which contains over 100,000 CXR images labeled with 14 different pathologies [113]. DenseNet-121, used in CheXNet, has become a standard architecture for medical image classification as it can capture fine-grained details through densely connected layers [44] in complex input spaces. Subsequent studies built on this work by incorporating additional datasets, improving the training process, and exploring other CNN architectures. To that extent, Irvin et al. introduced the CheXpert dataset, which includes over 220,000 CXR images and provides more robust annotations for model training and evaluation [45]. In our work, we leverage the Densenet-121 architecture and, indirectly, the CheXpert dataset by relying on models pretrained on CheXpert for both the classification and report generation tasks.

5.2.2 Vision encoder-decoder architectures

While CNNs constitute suitable architectures for image classification, generating detailed and contextually accurate reports from medical images requires a different approach. Vision encoder-decoder architectures, which combine CNNs for image feature extraction and transformers for text generation, have shown promise in addressing this challenge. These architectures leverage the strengths of CNNs in capturing visual information and transformers in handling sequential data. One notable example is the Show, Attend and Tell (SAT) model, which uses a CNN encoder to extract features from images and a recurrent neural network (RNN) decoder to generate descriptive text [110], this is one of the foundational and seminal works for image captioning in general. However, RNNs have limitations in capturing long-range dependencies and parallel processing, which transformers overcome with their self-attention mechanisms. As a result, generative models have shifted towards transformers in which the sequential aspect of natural language can be captured through the attention mechanisms.

The Convolutional vision Transformer (CvT) further advances this concept by integrating convolutional layers with transformer encoders, allowing for better

spatial understanding and feature extraction from images [116]. CvT has been successfully applied in various computer vision tasks, demonstrating improved performance over traditional CNNs and vanilla transformers.

As indicated, transformers are currently the main driving force when it comes to text generation tasks with their ability to model complex dependencies and generate corresponding text. BERT and GPT are among the most well-known and best performing transformer-based models in this domain. BERT, introduced, is a bidirectional transformer that captures context from both directions [21]. GPT, on the other hand, is a unidirectional transformer designed for generative tasks, capable of producing high-quality text, based on given prompts [81]. The combination of vision transformers such as CvT with transformer-based language models such as BERT enables the development of end-to-end systems for generating accurate radiology reports. Recent studies have explored the use of these models for medical image captioning and report generation, showing promising results in terms of both accuracy and interpretability [16].

5.2.3 Explainability in AI

As AI systems become more integrated into clinical workflows, the need for explainability grows in light of further, widespread adoption. Methods such as Grad-CAM provide visual explanations for model predictions by highlighting the regions of an image that contribute most to the output [96]. Incorporating such techniques is required for building trust in AI-driven systems among healthcare professionals. Grad-CAM has been widely used to explain CNN-based models, and recent work has extended its application to transformer architectures. By visualizing the attention weights in transformers, researchers can gain insights into how these models process and generate text based on visual inputs. In this work, we leverage Grad-CAM to enhance the interpretability of the proposed CXR report generation system. Specifically, the activation maps for the top predicted classes are used as explicit features in the input for the text generation model.

In this work we build on the advancements in CXR classification, vision encoder-decoder architectures, and transformer-based text generation models. By integrating Grad-CAM activation maps into the report generation pipeline, we show that we can generate reports on par with the current literature whilst providing an end-to-end explainable clinical decision support system.

5.2.4 Federated learning for medical image captioning

In recent years, federated learning (FL) has emerged as a paradigm for privacy-enahncing machine learning in healthcare applications, where data governance, regulatory compliance, and institutional boundaries prevent centralized data aggregation [100,122]. The core principle of FL is to collaboratively train models across multiple data silos without aggregating or sharing raw data, instead sharing only model updates during the training procedure. This is particularly relevant in radiology, where hospitals generate large volumes of imaging data which is considered privacy sensitive information, and therefore, not easily shared.

Within the domain of chest X-ray (CXR) analysis, FL has seen applications primarily in classification tasks [89], focusing on developing models that can detect pathologies like pneumonia, edema, or pleural effusion from distributed CXR datasets. However, the application of FL to medical image captioning, remains largely unexplored. Especially, encapsulating a full, end-to-end pipeline in an adoptable federated and explainable framework has not yet been widely covered for such complex learning tasks.

In this work, we propose a two-stage FL framework tailored to CXR report generation. The first stage consists of training a multi-label classifier across federated clients, each with their own local imaging datasets, to learn class-specific discriminative features. In the second stage, these features are exploited to guide the training of a federated vision-to-text model.

5.3 Methods

In this work, we define an end-to-end pipeline for the classification of CXR images and generation of a corresponding report whilst safeguarding the interpretability throughout and mitigating data-privacy restrictions and concerns. To that extent, we define two-core parts in our proposed methodology which are subsequently integrated as a single pipeline; (i) federated classification and (ii) augmented federated report generation. Each of the required components for this two-stage pipeline are discussed in depth in the subsequent sections, along with the data and specifications of the federated approach.



Findings: A hazy opacity is present in the right lung which may represent aspiration, pleural effusion or hemorrhage. Retrocardiac opacity at the left base is unchanged. Moderate cardiomegaly is stable. Slight prominence of the pulmonary vasculature with cephalization and enlarged pulmonary arteries are consistent with mild pulmonary edema. Tracheostomy tube is in place. There are no displaced rib fractures.

Impression: 1. Hazy opacity in the right lung which may represent aspiration versus pleural effusion or hemorrhage. 2. Mild pulmonary edema. 3. No displaced rib fractures.

Figure 5.1: Example of an original, raw CXR image in the dataset with the corresponding findings and impression sections from the radiology report.

5.3.1 Data

Throughout this work, the MIMIC-CXR (Medical Information Mart for Intensive Care - Chest X-ray) dataset [34, 49] is used. MIMIC-CXR constitutes a publicly available database specifically for medical imaging research. The dataset contains a comprehensive collection of chest X-ray images accompanied by associated radiological reports and clinical metadata. With 377 110 CXR images from 227 835 unique studies, corresponding to 65 379 patients, the dataset provides a large of data for training and evaluating machine learning models in various medical imaging tasks. Each image is linked to one or more radiological reports, providing detailed textual descriptions of observed findings. In addition, the dataset includes clinical metadata for each patient, enabling research into critical care medicine and healthcare informatics. The meta data for the dataset contains a predefined train, validate and test split to which we adhere in this work. The MIMIC-CXR dataset, as indicated, is multi-modal, containing both images and text. For this work, we extract up to 5 images per study and the corresponding written report. The written report is processed such that only the cleaned *findings* and *impression* sections remain, which are clinically of most relevance. Figure 5.1, depicts an example of an extracted, preprocessed, image-text pair.

5.3.2 Classification

To obtain class-dependent activation maps, a deep convolutional neural network, DenseNet-121 [44], is trained on the MIMIC-CXR data. As indicated in 5.2, we make use of the DenseNet-121 architecture as it can capture complex artifacts and patterns that occur in medical images. It does so by relying on the structure

of densely connected layers, which help in effective feature propagation and mitigating the vanishing gradient problem [44].

The architecture of DenseNet-121 consists of multiple dense blocks, each containing convolutional layers with batch normalization, ReLU activation, and dropout for regularization. The final dense block is followed by a global average pooling layer, which reduces the spatial dimensions and outputs a feature vector. The last layer of the model is a fully connected classification layer designed to yield predictions for each of the 14 possible class labels, corresponding to different pathological conditions in the CXRs. The global architecture used in this work is shown in Fig. 5.2.

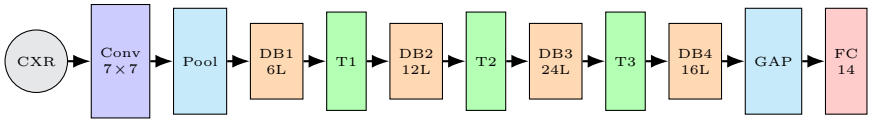


Figure 5.2: The figure represents the DenseNet-121 architecture as applied to CXR classification. The network starts with a convolution followed by a series of four Dense Blocks (DB1–DB4) containing 6, 12, 24, and 16 convolutional layers, respectively. Each Dense Block is preceded by a transition layer consisting of a 1×1 convolution followed by 2×2 average pooling. The model concludes with global average pooling and a fully connected output layer producing multi-label predictions for 14 chest related pathologies.

More formally, let $\mathbf{X} \in \mathbb{R}^{B\times 3\times H\times W}$ denote a batch of input CXR images, where B is the batch size, and H and W are the height and width of the images, respectively. The DenseNet-121 model $f_{\text{DenseNet}}(\cdot)$ processes the input images to produce a feature map $\mathbf{F} \in \mathbb{R}^{B\times D}$, where D is the dimensionality of the feature vector obtained after the global average pooling layer. The classification layer $\mathbf{W} \in \mathbb{R}^{D\times 14}$ then maps these features to the final output $\mathbf{Y} \in \mathbb{R}^{B\times 14}$, which contains the predicted probabilities for each of the 14 class labels:

$$\mathbf{Y} = \sigma(\mathbf{FW} + \mathbf{b}) \tag{5.1}$$

where $\sigma(\cdot)$ denotes the sigmoid activation function applied element-wise, and $\mathbf{b} \in \mathbb{R}^{14}$ is the bias term.

The data is labeled such that for each observation, a corresponding vector of length 14 exists. Each entry in the vector constitutes the indication of whether the corresponding class is observed (1), not observed (0), or uncertain (-1) as to whether the pathology is present. Following the methodology presented in [45], the uncertainty labels are dealt with by replacing -1 with 1, which has been shown to yield better predictive performance.

The classifier is trained on this multi-label classification using the binary cross-entropy loss, defined for each class c as:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i^c \log(\hat{y}_i^c) + (1 - y_i^c)\log(1 - \hat{y}_i^c)\right], \tag{5.2}$$

where N is the number of samples, y_i^c is the ground truth label for class c for the i-th sample, and $\hat{y}_i^c$ is the predicted probability for class c.

Once trained, this classifier serves as the pre-trained model to obtain a prediction, for which we can also obtain the confidence and subsequently the activation maps that correspond to the prediction. The activation maps, generated using Grad-CAM, highlight the regions of the image that contribute most to the classifier's decision for each class, thereby providing valuable insights into how the model attributes importance.

Grad-CAM

To obtain both insights into how the classifier attributes importance as well as explicitly crafted features that we can use to augment the input for the text generation component of the pipeline, we use Gradient-weighted Class Activation Maps (Grad-CAM). The technique is used to visualize the regions of an input image that are most important for a prediction. By providing insights into which parts of the image contribute most to the prediction of a particular class, practitioners can visually verify predictions, which in turn enhances interpretability of model decisions. Especially given the increasingly stringent regulations in healthcare related to the adoption of AI-driven applications, explainability and interpretability have become a core component of every CDSS. Grad-CAM specifically, generates a localization map that highlights important regions in an image for predicting a specific class, leveraging the gradients of the target class score with respect to the feature maps of a convolutional layer.

In the context of our proposed methodology, by augmenting the input images with these activation maps, we aim to generate accurate reports and improve the overal interpretability of an end-to-end pipeline which in turn aids in further adoption of these methods in practice.

Formally, the Grad-CAM operates in the following manner; consider a convolutional neural network processing an input image $\mathbf{I}$. Let A^k denote the k-th feature map obtained from a convolutional layer. The feature map A^k is a two-dimensional matrix representing the activations at each spatial location for the k-th channel. For a given class c, the network outputs a score

y^c, indicating the confidence that the input image belongs to class c. This score is typically obtained after applying a series of fully connected layers and a softmax function to the feature maps from the final convolutional layer.

Grad-CAM calculates the gradient of the class score y^c with respect to the feature maps A^k:

$$\frac{\partial y^c}{\partial A^k} \tag{5.3}$$

These gradients measure the sensitivity of the class score concerning the activations in the feature maps. The importance of each feature map A^k for the class c is then obtained by computing the average of these gradients across all spatial locations:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{5.4}$$

where α_k^c represents the importance weight of the k-th feature map for class c, and Z is a normalization factor equal to the number of spatial locations in the feature map.

The Grad-CAM activation map $L^c_{\text{Grad-CAM}}$ is obtained by performing a weighted combination of the feature maps A^k using the importance weights α_k^c:

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right) \tag{5.5}$$

The ReLU function $\text{ReLU}(x) = \max(0, x)$ is applied to ensure that only the features with a positive influence on the class score are included in the activation map.

In our methodology, these Grad-CAM activation maps are used as additional input channels to the vision encoder-decoder model. By incorporating these class-specific activation maps, we aim to enrich the visual representation and provide the model with more detailed and relevant information.

5.3.3 Report generation: vision encoder-decoder

In the second stage of our methodology, structured radiology reports are generated by combining visual information extracted from chest X-ray (CXR)

images with clinically informative activation maps derived from a DenseNet-121 classifier. Our approach uses a Vision Transformer (ViT) encoder [23] coupled with a GPT-2 language decoder [83]. To enable effective cross-modal learning and enhance interpretability, we augment the original visual input with Grad-CAM activation maps, thereby explicitly highlighting the regions of the image that contributed most significantly to the classifier's predictions.

A central design decision pertains to the selection of the number of Grad-CAM maps used. Empirical analysis, as visualized in Figure 5.3, indicates that both the actual pathological findings and the clinical topics discussed within radiology reports typically involve fewer than or equal to four distinct pathologies. Limiting the augmentation to the four most probable class-specific Grad-CAM maps effectively captures the most significant pathological information while ensuring computational feasibility. These activation maps offer visual guidance to the model, explicitly emphasizing areas associated with critical pathologies.

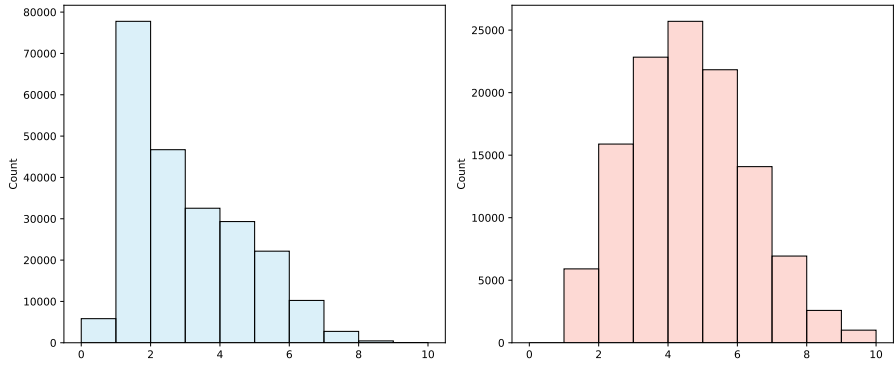


Figure 5.3: Histogram showing the distribution of the number of observed classes per instance (left) and the number of observed topics in the corresponding reports (right).

The augmented visual input for the ViT encoder is formally constructed as follows: We start by considering the original greyscale CXR image, represented by $\mathbf{x} \in \mathbb{R}^{1\times 224\times 224}$. From the DenseNet-121 classifier, we identify the four pathologies with the highest predicted probabilities. The Grad-CAM technique [96] is applied individually to these four classes, producing four corresponding activation maps, each indicating regions that significantly influenced the respective pathology prediction. These maps, $L^{c_i}_{\text{Grad-CAM}} \in \mathbb{R}^{1\times 224\times 224}$ for $i = 1, \ldots, 4$, are individually normalized using min-max scaling to ensure consistency and numerical stability. The augmented input, denoted as $\mathbf{x}^{\text{aug}}$, is

then represented as:

$$\mathbf{x}^{\text{aug}} = [\,\mathbf{x};\; L^{c_1}_{\text{Grad-CAM}};\; L^{c_2}_{\text{Grad-CAM}};\; L^{c_3}_{\text{Grad-CAM}};\; L^{c_4}_{\text{Grad-CAM}}\,] \in \mathbb{R}^{7\times 224\times 224}. \tag{5.6}$$

This explicit fusion of raw image data with the clinically informed attention maps encourages the model to prioritize clinically relevant visual features from the earliest encoding stages.

The Vision Transformer encoder used in our pipeline is ViT-Base, initially pretrained on ImageNet-21k and further fine-tuned on ImageNet-1k [24]. Given our augmented input comprises seven channels instead of the original three (RGB), the initial embedding convolutional layer is modified accordingly. To leverage pretrained weights effectively, the original three-channel weights of the stem layer remain intact, while the additional 4 channels' weights are initialized to zero. This initialization preserves the pretrained model's general image feature extraction capabilities, while allowing the model to gradually learn the importance of the newly introduced Grad-CAM channels during fine-tuning.

The textual component of the pipeline is managed by a GPT-2 decoder, pretrained on general-domain textual corpora. To ensure the generated radiology reports adhere to the standard clinical reporting structure, we incorporate specialized tokens explicitly delineating the sections within each report. Specifically, each training example comprises a concatenation of two primary sections, explicitly marked as follows:

$$\mathcal{Y} = \texttt{<FINDINGS>}\ \mathcal{F}\ \texttt{</FINDINGS>}\ \texttt{<IMPRESSION>}\ \mathcal{I}\ \texttt{</IMPRESSION>}, \tag{5.7}$$

where $\mathcal{F}$ and $\mathcal{I}$ denote the findings and impression sections, respectively.

Prior to tokenization, all reports undergo preprocessing to standardize formatting. This includes the removal of whitespace and newlines, ensuring that the model does not unnecessarily focus on formatting artifacts. After standardization, a byte-level Byte Pair Encoding (BPE) tokenizer with a vocabulary size of 30,000 tokens tokenizes the processed textual input.

During model training, we optimize the cross-entropy loss computed between generated and ground-truth token sequences, explicitly masking padding tokens to prevent their influence on training dynamics. The loss is formally expressed as:

$$\mathcal{L}_{\text{CE}} = -\sum_{t=1}^{T} \log p_\phi(y_t \mid y_{<t}, \mathbf{x}^{\text{aug}}), \tag{5.8}$$

where y_t denotes the ground truth token at position t, p_ϕ represents the conditional probability modeled by the GPT-2 decoder, and T indicates the total token length of the training instance.

To enforce clinically coherent and structured report generation during inference, we constrain the decoder to initiate report generation with the `<FINDINGS>` token and terminate the sequence upon generating the `</IMPRESSION>` token. To enhance generation quality, we employ beam search decoding with beam width $b = 4$, length penalty $\alpha = 0.8$, and repetition penalty $\gamma = 2.0$. These parameters were selected empirically to produce reports that balance brevity, clinical coherence, and readability.

A comprehensive summary of the training hyperparameters employed for the ViT–GPT2 fine-tuning process is provided in Table 5.1.

Table 5.1: ViT–GPT2 Training Hyperparameters

Hyperparameter	Notation	Setting
Initial learning rate	η_0	5×10^{-5}
Optimizer	–	AdamW ($\beta_1 = 0.9, \beta_2 = 0.95$)
Epochs	E	8 (early stopping ROUGE-L)
Effective batch size	B_{eff}	8
Beam width	b	4
Length penalty	α	0.8
Repetition penalty	γ	2.0

By combining visually interpretable Grad-CAM information with pretrained encoder-decoder models, our methodology, which we name Augmented Medical Image Captioning with Class-dependent Activation Maps (AMICCAM), aims to enhance interpretability and transparency but also holds potential for improved clinical relevance and accuracy in automated radiology reporting systems.

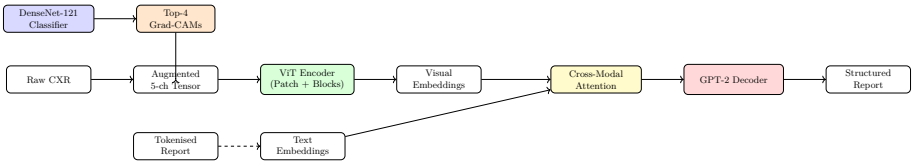


Figure 5.4: The proposed two-stage pipeline. A DenseNet-121 classifier produces class-specific Grad-CAMs that are concatenated with the raw image to form a 7-channel tensor. A ViT encoder converts this tensor into visual embeddings, which interact with text embeddings through a cross-modal attention block and a GPT-2 decoder to generate the structured radiology report.

5.4 Two-stage federated learning

A central aspect of the AMICCAM pipeline is its reliance on high-quality, pathology-specific activation maps produced by a well-trained classifier. To ensure consistent, robust activation maps across distributed data sources, we adopt a two-stage federated learning strategy implemented using the Flower federated learning framework [6]. This approach inherently supports partial privacy preservation, as data remains decentralized, with only model weights communicated between local sites and a central aggregation server. The secondary aim is to show the feasibility of implmenting federated learning approaches for complex, privacy-sensitive deep learning pipelines.

In the first stage, we train a DenseNet-121 classifier collaboratively across multiple federated clients. Each of the K participating institutions maintains a private subset of the MIMIC-CXR dataset, containing local training and validation sets. The global test set remains isolated and unseen by all institutions, reserved for final evaluation. At each federated round, local models undergo one or multiple epochs of training on their respective subsets. After local training, model weights are communicated to the central server, where they are aggregated into a global model via the FedAvg algorithm [67]. This iterative process repeats for a predefined number of rounds, resulting in a global classifier. We refer to Algorithm 6 for a structured overview of this first stage.

Once the global classifier is obtained, we initiate the second stage of training. In this stage, the global classifier is distributed to all sites to produce Grad-CAM maps from local images. Each site then locally constructs augmented image inputs by concatenating the top $k = 4$ Grad-CAM saliency maps with the original grayscale CXR image, formally represented as:

$$\mathbf{x}_{\text{aug}}^{(i)} = \left[\mathbf{x}_{\text{gray}}^{(i)}, \text{CAM}_i^{c_1}, \text{CAM}_i^{c_2}, \text{CAM}_i^{c_3}, \text{CAM}_i^{c_4} \right] \in \mathbb{R}^{7\times 224\times 224} \tag{5.9}$$

where $\mathbf{x}_{\text{gray}}^{(i)}$ denotes the normalized grayscale image at client i, and $\text{CAM}_i^{c_j}$ represents the normalized Grad-CAM map corresponding to the j^{th} most confidently predicted pathology class. The resulting seven-channel input tensor encodes both spatial image features and explicit pathology information, enhancing interpretability of the input for report generation.

In parallel, each site prepares textual radiology reports by tokenizing and structuring them using special tokens to seperate the FINDINGS and IMPRESSION sections, ensuring consistency during cross-modal decoding. Subsequently, the augmented images and tokenized texts serve as paired inputs for training a vision encoder-decoder (ViT-GPT2) model locally. Similar to stage one, local

Algorithm 6 Stage 1: Federated Classifier

Notation:
K: Number of sites.
$\mathcal{D}_i$: Dataset at site i, containing CXR images with multi-label pathologies.
w^C: Weights of DenseNet-121 classifier.
T_{clf}: Number of federated rounds for classifier training.
η_C: Learning rate for classifier.

procedure FEDERATEDCLASSIFIER($\{\mathcal{D}_i\}_{i=1}^{K},\ T_{\text{clf}},\ w_0^C,\ \eta_C$)
  $w^C \leftarrow w_0^C$ ▷ *Initialize global classifier*
  for $t \leftarrow 1$ to T_{clf} **do** ▷ *Federated rounds*
  (A) Broadcast:
    for $i \leftarrow 1$ to K **in parallel do**
      Send w^C to site i
    end for
  (B) Local Update:
    for $i \leftarrow 1$ to K **in parallel do**
      $w_i^C \leftarrow$ LOCALCLASSIFIERTRAINING($w^C,\ \mathcal{D}_i,\ \eta_C$) ▷ *Train DenseNet-121 on local data*
    end for
  (C) Aggregation:
    $w^C \leftarrow$ FEDAVG$\big(\{w_i^C\},\ \{|\mathcal{D}_i|\}\big)$ ▷ *Merge local classifier weights*
  end for
  return w^C ▷ *Final global classifier*
end procedure

ViT-GPT2 models undergo federated training rounds, with their weights being aggregated using FedAvg.

Table 5.2 summarizes the key parameters and settings utilized for federated training. [56, 100].

The described two-stage federated approach as depicted in Algorithm 6 and Algorithm 7 aims to facilitate training of a interpretable caption generation model without having to centralize sensitive patient data.

5.4.1 Evaluation metrics

To evaluate the quality and clinical accuracy of the proposed approach we use the following metrics: BLEU [75], ROUGE [60], CIDEr [109], as well as the classification-oriented metric Area Under the ROC Curve (AUC) [30]. Table 5.3 summarizes each of these metrics along with the corresponding notation.

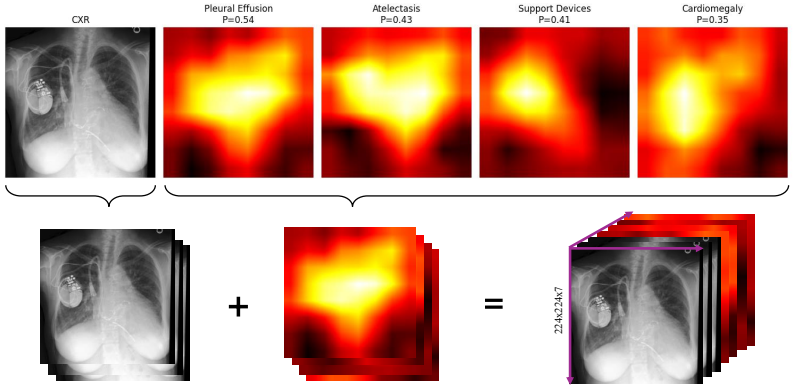


Figure 5.5: Example of the augmented input, the raw CXR is enhanced with the top 4 saliency maps serving as a 7-channel input to the Vision Encoder-Decoder model.

Table 5.2: Federated learning hyperparameter settings used in both training stages.

Parameter	**DenseNet-121**	**ViT-GPT2**
Number of federated rounds	30	20
Local epochs per round	2	1
Optimizer	AdamW	AdamW
Learning rate (η)	1×10^{-4}	5×10^{-5}
Aggregation strategy	FedAvg	FedAvg
Clients/sites (K)	10	10
Batch size	32	8

5.5 Results

In this section, we present a comprehensive evaluation of the proposed pipeline, which integrates class-specific interpretability through Grad-CAM into a vision-language model for automated radiology report generation. We assess and compare three models: a baseline ViT-GPT2 transformer model trained on raw chest X-ray images, the augmented model AMICCAM which incorporates Grad-CAM-based saliency maps, and its federated counterpart AMICCAM-FL. All

Table 5.3: Summarized representation of the different metrics used to evaluate clinical and linguistic accuracy of the two-stage federated approach.

Metric	Definition	Interpretation
BLEU [75]	$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log P_n\right)$, with $P_n = \frac{\sum \min(\text{Count}_C, \text{Count}_R)}{\sum \text{Count}_C}$ and $BP = \begin{cases} 1 & c > r \\ e^{1-\frac{r}{c}} & c \leq r \end{cases}$	Measures precision of n-grams between generated and reference text, penalizing overly brief outputs
ROUGE-L [60]	$ROUGE\text{-}L = \frac{LCS(X,Y)}{\text{length}(Y)}$	Measures recall based on longest common subsequence, indicating text fluency and coverage
CIDEr [109]	$CIDEr = \frac{1}{M} \sum_{m=1}^{M} \frac{\sum w_i g_i}{\|w\|\|g\|}$, where $w_i = tf_{i,j} \cdot \log\left(\frac{N}{df_i}\right)$	Weighs n-grams using TF-IDF to capture consensus across multiple references
BERTScore [21]	$BERTScore = \text{cosine-similarity}(\text{BERT}(x), \text{BERT}(y))$	Uses contextual embeddings from BERT to measure semantic similarity between generated and reference sentences
AUC [30]	$AUC = \int_0^1 TPR(FPR)\, d(FPR)$	Area under ROC curve, quantifying classifier's ability to discriminate across different thresholds

Algorithm 7 Stage 2: Federated AMICCAM

Notation:
Same K sites, each with $\mathcal{D}_i$ now containing *both* images $x_i^{(n)}$ and textual radiology reports $r_i^{(n)}$.
w^C: *Final* classifier weights from Stage 1.
w^R: Weights of the report-generation model (vision encoder-decoder).
T_{rep}: Number of federated rounds for report generation.
η_R: Learning rate for the report-generation model.
GradCAM$(x;\, w^C)$: Activation maps derived from the *fixed* classifier.

procedure FEDERATEDAMICCAM($\{\mathcal{D}_i\},\ T_{\text{rep}},\ w^C,\ w_0^R,\ \eta_R$)
  $w^R \leftarrow w_0^R$ ▷ *Initialize global report-generation model*
  for $t \leftarrow 1$ to T_{rep} **do** ▷ *Federated rounds for captioning*
  (A) Broadcast:
    for $i \leftarrow 1$ to K **in parallel do**
      Send w^R *(and frozen* w^C *if needed for local inference)* to site i
    end for
  (B) Local Update:
    for $i \leftarrow 1$ to K **in parallel do**
      for each (x, r) in $\mathcal{D}_i$ **do**
        $\hat{y} \leftarrow \text{DenseNet}(x; w^C)$ ▷ *Classifier is fixed/frozen from Stage 1*
        $A \leftarrow \text{GradCAM}(x;\, w^C)$ ▷ *Obtain top-k saliency maps (Augmented input:* (x, A) *for the report generator)*
      end for
      $w_i^R \leftarrow$ LOCALREPORTGENTRAINING$(w^R,\ \mathcal{D}_i,\ A,\ \eta_R)$
    end for
  (C) Aggregation:
    $w^R \leftarrow$ FEDAVG$\big(\{w_i^R\}, \{|\mathcal{D}_i|\}\big)$
  end for
  return w^R
end procedure

procedure INFERENCE(x_{new}, w^C, w^R)
  $\hat{y} \leftarrow \text{DenseNet}(x_{\text{new}};\ w^C)$
  $A_{\text{new}} \leftarrow \text{GradCAM}(x_{\text{new}};\ w^C)$
  $r_{\text{new}} \leftarrow \text{EncoderDecoder}(x_{\text{new}}, A_{\text{new}};\ w^R)$
  return r_{new}
end procedure

models are trained and evaluated on the MIMIC-CXR dataset using consistent data partitions.

5.5.1 Classification performance

The performance of the DenseNet-121 classifier used to generate Grad-CAM maps is evaluated in both centralized and federated training settings. Table 5.4 presents AUC scores for a selection of the most relevant pathology classes.

The results indicate strong and consistent classification capabilities across both training paradigms. This robustness is critical for generating reliable saliency maps that serve as inputs to the captioning pipeline. To illustrate the qualitative interpretability of these maps, Figure 5.6 shows examples of Grad-CAM overlays highlighting salient regions for three major pathology classes.

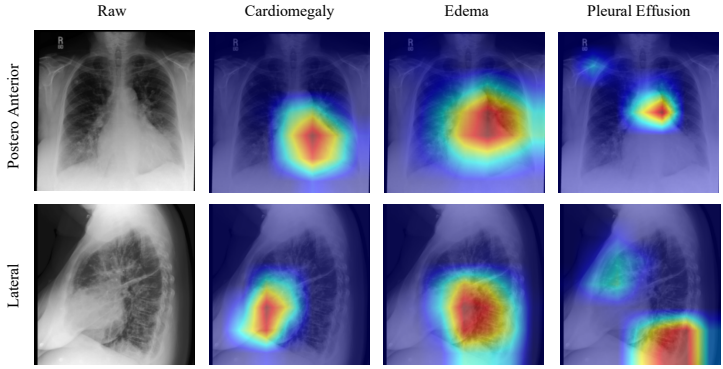


Figure 5.6: Grad-CAM activation maps for three pathology classes, overlaid on chest X-rays. These saliency maps serve as additional input channels to the captioning model, encoding pathology-specific visual semantics.

5.5.2 Radiology report generation

We evaluate the captioning performance across the three model variants using both traditional language quality metrics (BLEU-4, ROUGE-L, CIDEr) and BERTScore. The results are summarized in Table 5.5.

The AMICCAM model shows some improvements over the baseline. Most notable is that performance is very much on par with the baseline with the added benefit that the pipeline is more interpretable overall. The limited improvements indicate that, to some extent, augmenting input images with class-specific Grad-CAM maps contributes meaningful and interpretable visual signals that improve downstream report generation.

Table 5.4: Classifier performance (DenseNet-121) measured in terms of AUC scores for selected pathology classes.

Approach	**Atelectasis**	**Cardiomegaly**	**Consolidation**	**Edema**	**Pleural Effusion**
Centralized	0.82	0.87	0.80	0.89	0.92
Federated	0.80	0.85	0.78	0.87	0.90

Table 5.5: Evaluation metrics comparing the baseline ViT-GPT2 model against AMICCAM and AMICCAM-FL on the full MIMIC-CXR dataset.

Model	BERTScore	CIDEr	ROUGE-L	BLEU-4
Baseline (ViT-GPT2)	0.761	0.209	0.313	0.061
AMICCAM	0.765	0.213	0.314	0.062
AMICCAM-FL	0.752	0.205	0.301	0.057

AMICCAM-FL closely approximates the centralized performance, with small reductions across all metrics. This performance validates the feasibility of training such interpretability-aware pipelines in privacy-enhancing federated environments without sacrificing predictive performance. In this scenario, the pay-off between decentralized training and overall performance is limited, which, in itself, is a characteristic in favor of adoption in practice.

5.6 Discussion

This study explored a hybrid approach to medical image captioning by explicitly encoding class-dependent interpretability into the vision-language pipeline. Our results show that the use of Grad-CAM activation maps as additional input channels leads slightly improved performance in terms of linguistic quality and clinical relevance.

The integration of interpretability into the model input space distinguishes AMICCAM from traditional attention-based captioning frameworks, which rely solely on implicit feature extraction. By anchoring the encoder on pathology-specific image regions, our method provides an interpretable mechanism that aligns closely with clinical reasoning. Radiologists can visually verify that the model focuses on medically plausible features, addressing one of the primary barriers to trust and adoption in clinical AI-driven systems.

The AMICCAM-FL variant demonstrates that such pipelines can be deployed in distributed environments, where central data sharing is infeasible due to ethic and regoulatory constraints. This reinforces the argument that capable and interpretable AI tools can be trained across institutional boundaries while enhancing data privacy.

5.7 Conclusion and limitations

We introduced AMICCAM, a two-stage framework that incorporates pathology-specific Grad-CAM activation maps into a vision-language model for medical image captioning. The results demonstrate slight improvements in linguistic quality and clinical accuracy of generated reports over traditional models trained on raw image inputs. Through explicit visual augmentation, the model gains interpretable and semantically targeted supervision, resulting in outputs that are not only more accurate but also explainable.

We further developed and validated AMICCAM-FL, a federated variant of the pipeline that maintains performance whilst decentralized training. This contribution addresses a challenge in medical AI: enabling complex model and pipeline development under strict privacy constraints without centralizing sensitive data.

Taken together, the proposed methods contribute to the field related to interpretable, federated vision-language learning for clinical applications. They lay the groundwork for future extensions into multimodal and multiview settings, while aligning with the ethical and regulatory standards demanded in real-world healthcare environments.

Nonetheless, everal aspects require further investigation. First, while Grad-CAM maps offer transparency, their dependency on intermediate feature gradients may introduce sensitivity to noise and classifier calibration. Additionally, future work could explore alternative saliency mechanisms (e.g., Score-CAM, XGrad-CAM) or ensemble techniques to enhance map quality and stability. Moreover, our study focused on single-view frontal CXRs, whereas real-world diagnostic scenarios often involve multi-view or temporal imaging data. Extending the current framework to accommodate such multimodal inputs may further improve clinical robustness and generalizability.

5.8 Supplementary context

This section provides additional context for the AMICCAM federated medical image captioning pipeline, addressing validation methodology, error propagation risks, and the role of the federated component.

5.8.1 Validation of generated captions

The evaluation of generated radiology reports in this chapter relies on standard natural language generation metrics: BLEU, ROUGE, CIDEr, and BERTScore. While these metrics are widely used and enable comparison with prior work, they have well-documented limitations in the clinical context. BLEU measures n-gram overlap and penalizes length mismatch but does not assess semantic accuracy: a report that mentions the correct pathology in different words would receive a low BLEU score, while a report that uses similar words but attributes findings to the wrong anatomical location could receive a high score. CIDEr partially addresses this through TF-IDF weighting but remains fundamentally a lexical metric.

The absence of a clinical validation protocol is a limitation of this work. In an ideal evaluation, generated reports would be assessed by radiologists along several dimensions: (i) clinical accuracy (are all findings correctly reported and no critical findings omitted?), (ii) anatomical correctness (are findings attributed to the correct locations?), (iii) severity calibration (are findings described with appropriate severity?), and (iv) actionability (does the report support the correct clinical decision?). A critical-finding error rate, measuring how often the model hallucinates a finding that would trigger unnecessary intervention or omits a finding that requires urgent action, would be the most clinically meaningful metric. Such evaluation requires expert clinical involvement and is beyond the scope of this thesis, but we strongly recommend it as a necessary step before any deployment consideration. The automated metrics presented here should be interpreted as measuring linguistic quality and superficial content overlap, not clinical safety.

5.8.2 Saliency is not causality

The Grad-CAM activation maps used to augment the image input are saliency maps that highlight regions the classifier finds most discriminative for its predictions. Saliency does not imply causality. The classifier may learn to rely on spurious correlations or shortcut features rather than genuine pathological indicators. In a multi-site federated setting, site-specific artifacts, institutional logos, image orientation markers, scanner-specific noise patterns, or positioning devices visible in the images, could serve as discriminative features that are predictive of certain labels at specific sites but do not generalize.

Several strategies could mitigate this risk. First, preprocessing to remove or mask known artifact regions (e.g., image borders, text overlays) before classification. Second, adversarial training that penalizes the classifier for using

features that are predictive of the source institution rather than the pathology. Third, systematic comparison of activation maps across institutions for the same pathology class to identify site-specific versus pathology-specific patterns. The current implementation does not include these safeguards, which is a limitation. The Grad-CAM visualizations presented in the results (and quantified via SSIM and FID across models) provide preliminary evidence of consistency, but a rigorous evaluation would require expert annotation of whether highlighted regions correspond to genuine pathological features.

5.8.3 Error propagation

The AMICCAM pipeline operates in two stages: first, a multi-label classifier identifies probable pathologies and generates class-specific Grad-CAM maps; second, these maps are fused with the original image to form the input to the caption generation model. This sequential design creates a dependency chain in which errors in the first stage propagate to the second. If the classifier misclassifies an image (producing incorrect top-k predictions), the resulting Grad-CAM maps will highlight regions relevant to the wrong pathologies, and the caption generator will condition on misleading visual information.

The severity of this error propagation depends on the classifier's error mode. False negatives (missing a pathology) mean that the corresponding Grad-CAM map is absent, depriving the caption generator of useful guidance; the generator must then rely on the raw image features alone for that finding. False positives (incorrectly predicting a pathology) inject misleading saliency information, which could cause the generator to hallucinate findings. The latter is clinically more dangerous, as it could lead to unnecessary follow-up procedures. Potential mitigations include: (i) uncertainty gating, where the Grad-CAM augmentation is only applied for predictions above a confidence threshold, falling back to the raw image for uncertain classes; (ii) ensemble-based classification, where multiple classifiers must agree before a class-specific map is included; and (iii) training the caption generator with augmented inputs that include deliberately incorrect Grad-CAM maps, teaching it to be robust to first-stage errors. We did not implement these strategies in the current work but they represent important directions for improving pipeline robustness.

5.8.4 Separating along classes

We generate separate Grad-CAM activation maps for each of the top-k predicted classes, rather than a single aggregate saliency map. This decision is motivated by the multi-label nature of the CXR classification task. A single chest X-ray

may simultaneously exhibit multiple pathologies (e.g., cardiomegaly and pleural effusion) with distinct spatial distributions. An aggregate saliency map would merge the regions relevant to different pathologies, creating a diffuse and less informative input for the caption generator. By providing separate per-class maps as additional input channels, the caption generator receives structured spatial guidance that preserves the distinction between pathologies. This design also enhances interpretability: a clinician reviewing the system's output can inspect the per-class activation maps to understand which regions drove each finding in the generated report.

5.8.5 Improvement over single-center learning

The federated component of this chapter may appear secondary to the core methodological contribution (the Grad-CAM augmented captioning pipeline). This observation is partially valid: the primary novelty lies in the augmented image representation and its integration into the captioning pipeline, and the chapter would have substantial value even without the federated experiments. However, the federated experiments serve two purposes. First, they demonstrate feasibility: applying FL to a complex, multi-stage deep learning pipeline is non-trivial, and showing that the pipeline's performance is preserved (or only minimally degraded) under federated training is a meaningful empirical contribution. Second, they connect this chapter to the overarching thesis narrative of building towards practical, privacy-enhancing clinical decision support systems.

Regarding improvement over single-center learning: the federated setup pools knowledge from multiple institutions, which should improve model robustness compared to a model trained on a single hospital's data. However, the MIMIC-CXR dataset used in this chapter originates from a single institution (Beth Israel Deaconess Medical Center), and the federated setup is simulated by partitioning this data across virtual clients. The non-IID challenges encountered are therefore less severe than in a genuinely multi-institutional setting, and the comparison between federated and centralized training is less informative about real-world deployment. Evaluation on a genuinely multi-institutional CXR dataset (such as CheXpert combined with MIMIC-CXR, or the NIH Chest X-ray dataset) would provide a more rigorous test of the federated pipeline's value proposition. We note this as an important extension for future work.

5.8.6 Non-IID bottleneck

In the two-stage federated pipeline, non-IID data can affect each stage differently. The classifier stage, which performs multi-label classification, is affected by label distribution skew: different sites may observe different prevalence of pathologies, leading to heterogeneous gradient updates during federated classification training. The caption generation stage is affected by both label skew (which pathologies appear in the reports) and linguistic variation (different institutions may use different reporting conventions, vocabulary, or level of detail). Of these, the classifier stage is likely the dominant bottleneck under heterogeneity, because classification performance directly determines the quality of the Grad-CAM maps that condition the second stage. Approaches that could improve robustness include personalized classification heads per site (sharing the feature extractor but maintaining site-specific classification layers), site-adaptive batch normalization, or applying the RAFA aggregation strategy from Chapter 4 to the classifier training stage specifically.

Chapter 6

Conclusions and Future Perspectives

Recent developments in artificial intelligence have shown potential to advance healthcare by enhancing diagnostic accuracy, augmenting clinical decision support systems, and enabling tailored, patient-centric solutions. The increasing availability of medical datasets, combined with computational resources and methodological advancements, has accelerated the integration of AI within clinical workflows, ranging from medical imaging and radiology to predictive modeling in intensive care and personalized medicine. However, the practical integration and widespread adoption of AI-driven methodologies within healthcare systems still face several challenges. The major factor among these challenges are the strict regulatory frameworks concerning data privacy and protection, such as the General Data Protection Regulation (GDPR) in the European Union and the Health Insurance Portability and Accountability Act (HIPAA) in the United States, which govern the use and sharing of sensitive patient data. Furthermore, healthcare institutions frequently maintain isolated data repositories, often referred to as institutional data silos, due to competitive, administrative, or jurisdictional reasons, thereby limiting the aggregation of comprehensive datasets needed for developing robust AI models. The inherent heterogeneity across patient populations, medical practices, and clinical protocols further increases this issue, introducing complexities that diminish model generalizability and performance across different institutional settings.

In response to these challenges, federated learning has emerged as an innovative paradigm enabling collaborative AI model training across multiple decentralized

data sources without the need to centralize or directly share sensitive patient data. In federated frameworks, individual institutions retain full control over their respective datasets, training AI models locally and subsequently communicating only aggregated model updates or parameters to a centralized orchestrator. This decentralized approach not only aligns well with regulatory and ethical standards but also creates opportunities for international, multi-institutional collaboration that would otherwise be hindered by data-sharing constraints. Nevertheless, traditional federated learning methodologies face several technical and practical obstacles, notably data heterogeneity arising from diverse local data distributions, substantial communication overhead in transmitting model updates, and persistent privacy and security vulnerabilities. Collectively, these limitations represent bottlenecks that restrict the adoption of federated learning within real-world clinical environments.

This dissertation addresses these key challenges through a series of methodological innovations, empirical evaluations, and practical demonstrations tailored to healthcare applications of federated learning. The remainder of this chapter elaborates on the main scientific and practical contributions derived from this research, discusses their broader implications for healthcare management and clinical practice, identifies current limitations of the proposed methodologies, and outlines strategic directions for future research and development.

6.1 Contributions

Target-informed client recruitment. The first major contribution of this dissertation involves the development and validation of a data-informed client recruitment strategy. As presented in Chapter 2, this method leverages metadata, specifically the local output distribution statistics and local sample sizes, to identify and recruit the subset of federated learning clients most likely to positively impact the global learning process. Specifically, the approach assesses client-level representativeness, prioritizing participants whose data distributions align closely with the collective data characteristics across the federation. Conversely, it strategically pre-excludes clients with smaller sample sizes or those whose output distributions significantly diverge from the aggregated global data. This targeted selection not only enhances model accuracy but also improves the efficiency of federated training processes. Empirical evaluation demonstrated that federated models trained under this recruitment paradigm yielded higher predictive performance relative to baseline federated learning methods without selective client participation. Additionally, the proposed recruitment strategy reduced training times, directly attributable to the decreased communication overhead stemming from fewer active client

nodes. Consequently, this methodological innovation provides a practical and computationally efficient entry point for federated learning deployments within hospital networks, where resource availability and training costs are often highly variable and constrained.

Despite its demonstrated effectiveness, the presented client recruitment strategy does have certain limitations, most notably the dependence on a user-defined recruitment parameter, γ_{th}. This parameter directly determines the selectivity of client inclusion within the federated training, thereby substantially influencing the number of actively participating clients. In real-world healthcare scenarios, precise tuning of γ_{th} may not always be feasible due to practical constraints, dynamic clinical environments, or limited resources for parameter exploration. Additionally, the experiments conducted in this dissertation utilized a simulated, single-process computational environment in which communication overhead was approximated but not fully representative of real-world networking scenarios.

Full multi-factor framework for client recruitment in FL. Expanding on the previously discussed data-informed client recruitment methodology, Chapter 3 introduced a comprehensive multi-factor client recruitment framework for federated learning, integrating multiple dimensions beyond mere data representativeness. This extended recruitment approach incorporates local computational capabilities, explicitly assessing each hardware architecture and the corresponding expected local training duration. Additionally, the method factors in statistical representativeness, combining metrics of local sample size and divergence from global target distributions to construct a unified cost–informativeness metric. The proposed framework effectively prioritizes the most suitable institutions to participate in federated training rounds, ensuring maximal efficiency and minimal computational redundancy.

Empirical evaluations presented in Chapter 3 demonstrate the methodological robustness and practical relevance of this multi-factor recruitment framework across both regression and classification tasks within realistic healthcare scenarios, encompassing managerial as well as clinical use-cases. Experimental outcomes indicate that federated models trained exclusively with strategically recruited clients consistently achieve predictive performance either superior to, or at least comparable with, conventional federated learning approaches and fully centralized training methods. Crucially, these performance gains were realized alongside significant reductions in overall federated training time, highlighting the potential for enhanced computational efficiency and operational feasibility within resource-constrained clinical environments.

Robust adaptive aggregation for non-IID data distributions. A central and persistent challenge in federated learning, particularly within healthcare applications, is managing heterogeneous data distributions across participating

institutions. Non-IID conditions often arise naturally in medical settings due to variations in patient demographics, clinical practices, diagnostic criteria, and data acquisition protocols, thereby complicating model convergence and limiting the predictive robustness and generalization of federated models. Existing aggregation strategies, including classical methods such as FedAvg and variants such as FedAvgM and FedProx, have attempted to mitigate these effects through modifications like momentum-based aggregation or proximal terms. However, such approaches remain limited in their efficacy, especially under conditions of extreme data heterogeneity, frequently encountered in clinical environments.

To address these limitations, Chapter 4 introduced Robust Adaptive Federated Aggregation (RAFA), a novel aggregation methodology explicitly designed for enhanced performance and stable convergence under severe non-IID scenarios. RAFA incorporates an adaptive weighting mechanism that dynamically adjusts client contributions based on historical client update quality and statistical relevance of their local models relative to the global optimization objective. By assigning greater aggregation weights to client updates that show higher consistency and predictive value, RAFA mitigates the adverse impacts arising from distributional skew, while reducing the effects of noisy or less informative client contributions.

The robustness and generalizability of RAFA were evaluated across a range of benchmark datasets, including commonly employed benchmarks such as MNIST and CIFAR-10, as well as a clinically relevant real-world dataset, the FeTS brain tumor segmentation dataset derived from multi-institutional MRI imaging studies. Empirical results showed that RAFA outperformed or performed on par with, standard and state-of-the-art federated aggregation techniques (e.g., FedAvgM and FedProx) across diverse tasks and datasets. Most important, RAFA shows better convergence stability, achieving lower standard deviations in both accuracy and loss across iterative training rounds.

Framework for multi-stage, interpretable FL pipelines. In Chapter 5, we introduced and evaluated a hybrid approach to medical image captioning, AMICCAM, which explicitly integrates class-dependent interpretability into the vision-language modeling pipeline. The AMICCAM framework leverages pathology-specific Grad-CAM activation maps as additional input channels for a ViT-GPT2 architecture, facilitating direct and visually interpretable guidance during caption generation. Unlike traditional attention-based captioning frameworks that rely solely on implicit feature extraction, AMICCAM explicitly anchors the vision encoder to clinically relevant and interpretable regions of medical images, closely aligning the model's decision-making process with established clinical reasoning patterns.

Empirical results demonstrated that the inclusion of Grad-CAM saliency maps

slightly but consistently improved both the linguistic quality and clinical relevance of the generated radiology reports. Beyond quantitative improvements in performance metrics, the explicit integration of interpretability mechanisms within the input space significantly enhances model transparency. Radiologists and clinical stakeholders can visually inspect and verify that the attention aligns with medically relevant pathological features, thereby addressing challenges related to trust, adoption, and practical integration of AI-driven diagnostic tools within clinical workflows.

Extending this interpretability-driven captioning pipeline to federated settings, we developed AMICCAM-FL, a variant specifically designed for decentralized training environments where central data aggregation is infeasible due to ethical, regulatory, or institutional data-sharing constraints. Our comprehensive evaluation revealed that AMICCAM-FL successfully retained the performance benefits and interpretability characteristics of its centralized counterpart, demonstrating the feasibility and robustness of federated learning approaches even in complex and multi-stage domains.

6.2 Limitations and future research

Despite the progress presented in this dissertation, several limitations and challenges remain, highlighting directions for future research and development within the domain of FL for healthcare.

Client transparency and autonomy The client recruitment and aggregation methodologies proposed in this work inherently assume some degree of transparency and centralized coordination among participating institutions. However, in real-world clinical and healthcare contexts, these assumptions may not always hold true due to privacy regulations, competitive considerations, and concerns around intellectual property. Institutions are often reluctant to disclose detailed information regarding their computational resources, internal data characteristics, even non-privacy sensitive agglomerate statistics, or specific patient demographics, complicating transparent and efficient federated collaboration. Consequently, future research efforts can focus on developing decentralized or cryptographically secured methods for client profiling, data representativeness evaluation, and computational capability assessment. Potential directions include leveraging secure multiparty computation and differential privacy techniques. These innovations could facilitate federated collaborations without requiring explicit data disclosures, further enhancing autonomy and trust among participants. Specifically, without the target distribution P_c, the recruitment score proposed in Chapters 2 and 3 cannot

be computed, and client selection falls back to random sampling or sample-size-only weighting. Without sample size n_c, the convergence-rate term is unavailable. Without hardware information H_c, the framework degrades to the Chapter 2 formulation. For RAFA, the core mechanism requires access to model updates, the minimal information exchanged in any FL protocol, but the contribution scoring relies on the server inspecting these updates, which is, for now, incompatible with encrypted aggregation protocols.

To preserve method usefulness under limited transparency, several privacy-preserving alternatives exist. Local differential privacy could be applied to histograms P_c before transmission, with calibrated noise that preserves the ranking of clients while obscuring exact values. Secure aggregation could compute the global distribution P_g without the server learning individual contributions. For RAFA, secure contribution scoring via additive secret sharing is feasible but computationally expensive. The choice among these alternatives depends on the specific trust model and regulatory requirements of the deploying consortium.

Fairness and representation Although the Robust Adaptive Federated Aggregation method presented in this thesis mitigates performance degradation due to data heterogeneity, achieving good performance across participating institutions remains challenging. Differences in patient populations, sample sizes, and data quality across institutions pose significant risks of systemic biases, potentially disadvantaging smaller, resource-limited institutions or minority patient groups. An interesting research avenue would be to explicitly incorporate fairness-aware objectives into aggregation frameworks, ensuring that federated models do not reproduce or increase existing health inequities. Methodological advancements could include fairness-constrained aggregation algorithms, bias detection and mitigation strategies, and adaptive weighting schemes explicitly targeting fair representation. Empirical evaluation of these methods across diverse demographic and institutional contexts would be essential, ensuring federated learning solutions are free of any bias.

Complex multimodal and temporal data integration The AMICCAM-FL pipeline introduced herein effectively demonstrates federated interpretability within medical imaging tasks, however complexities arise when extending federated learning to multimodal and temporal clinical data. Healthcare involves integrating heterogeneous data sources such as electronic health records, genomic profiles, laboratory tests, and longitudinal patient records, presenting significant modeling and computational challenges. Future work should explore federated learning architectures explicitly designed for complex multimodal integration, employing advanced neural attention mechanisms, intermodal knowledge distillation, and hierarchical modeling frameworks. Additionally, incorporating longitudinal and temporal modeling techniques, such as temporal

attention transformers, and survival analysis methods, may enhance federated model ability to manage temporally structured data.

Real-world deployment, evaluation, and governance Moving from theoretical development to practical deployment of federated learning solutions in healthcare requires overcoming infrastructural, regulatory, and operational challenges. Real-world deployment scenarios require robust federated orchestration platforms that facilitate seamless, secure, and scalable federated training and model updates across institutions. Furthermore, continuous model monitoring and auditing mechanisms are essential to detect and mitigate performance drift, data distribution changes, and potential adversarial manipulations. Future research must focus on establishing standardized evaluation metrics, comprehensive model documentation frameworks, and compliance monitoring tools aligned with existing clinical governance and regulatory standards. Investigations into regulatory compliance, ethical considerations surrounding federated model deployment, and public transparency mechanisms will further improve trustworthiness, accountability, and clinical acceptance of federated solutions.

In conclusion, addressing these limitations and pursuing these research directions will further strengthen federated learning methodological robustness, clinical applicability, and real-world impact.

6.3 Future perspective: towards integrated federated decision support systems

Looking forward, federated learning holds promise for integration into clinical decision support systems, enhancing their capacity to deliver personalized and accurate medical recommendations while maintaining patient privacy standards. Future federated learning-based CDSS could actively manage uncertainty by identifying cases where model predictions exhibit lower confidence or higher variability, thereby deferring these complex scenarios to clinical experts for further review. This approach not only leverages human clinical judgment effectively but also maintains a high standard of care by reducing potential risks associated with incorrect or uncertain AI-driven decisions.

Moreover, federated CDSS frameworks have the potential to interactively solicit targeted additional data, dynamically enriching their predictive power and interpretability. For instance, federated models could systematically request specific clinical parameters, additional diagnostic imaging, or laboratory tests based on their expected informational value in refining model predictions. Such

adaptive data solicitation can enhance both clinical accuracy and operational efficiency by focusing clinical resources precisely where additional information is most impactful, thus reducing unnecessary diagnostics and healthcare expenditures.

Integrating federated learning with active experimentation methodologies, such as adaptive clinical trials or randomized controlled interventions, represents another significant area for future exploration. Federated models could proactively initiate targeted experiments within participating institutions to efficiently resolve uncertainties about treatment effectiveness or diagnostic procedures. Through this mechanism, federated CDSS can continuously refine their predictive performance and expand their knowledge base while adhering to rigorous ethical and regulatory standards. Such integration can result in greater data availability and higher quality evidence, facilitating evidence-based medical practices and improving overall patient outcomes.

The embedding of clinical domain expertise within federated workflows also constitutes a promising direction to improve interpretability and clinical relevance. By designing federated models that explicitly incorporate domain-specific constraints, guidelines, and evidence-based clinical pathways, these systems could provide clinicians with decision recommendations that are not only statistically robust but also transparent and aligned closely with established medical practice. Enhanced interpretability, particularly through explainable AI techniques such as integrated saliency mapping, attention mechanisms, and causal inference frameworks, can foster clinician trust, enhance user acceptance, and facilitate informed clinical judgment.

Several use-case scenarios show how federated CDSS integration can offer benefits. In medical imaging, federated models could enable multi-institutional collaboration on disease classification and segmentation tasks, improving diagnostic accuracy through diverse data sources while preserving patient privacy. In intensive care settings, FL-CDSS could predict adverse events, such as sepsis onset or clinical deterioration, leveraging data from heterogeneous patient populations to provide highly reliable risk stratifications and early warning indicators. Additionally, in the realm of personalized medicine, federated approaches can incorporate genomic and phenotypic data from multiple institutions, enhancing predictive accuracy for individualized strategies and reducing clinical risks associated with treatment decisions.

Furthermore, federated CDSS solutions inherently mitigate data privacy risks associated with conventional centralized systems. By design, federated frameworks minimize data exposure and reduce vulnerabilities to data breaches or unauthorized access, thereby aligning with increasingly stringent privacy regulations and ethical considerations.

The methodological innovations and empirical validations presented in this thesis provide foundational advances in the readiness and practicality of federated learning within healthcare. By addressing challenges related to client recruitment, aggregation robustness, interpretability, and data privacy, this work lays essential groundwork for the broader integration of federated AI solutions into advanced clinical decision support systems. Ultimately, such integrative FL-CDSS platforms hold potential to enhance clinical accuracy, operational efficiency, interpretability, and patient safety. A recurring theme across all contributions is the view that a well-trained global model serves as a strong initialization point from which local fine-tuning can yield high-performing, site-specific solutions, rather than as a one-size-fits-all approach. Complementing this, the thesis has argued that FL provides not only privacy in the form of protection against inference from model updates, but also control. Both properties are essential for building the trust required for multi-institutional collaboration in healthcare. By addressing challenges related to client recruitment, aggregation robustness, interpretability, and data privacy, the methods presented here contribute to making federated clinical decision support systems not only technically feasible but operationally practical, ethically sound, and strategically attractive to the institutions that must adopt them.

Appendix A

Appendix

A.1 Features

Table A.1 outlines and describes the temporal and static features that constitute the data cohort used for all training procedures described within this work.

A.2 Federated Algorithms

Algorithm 8 Fed-AC: Federated Learning with All Clients

Server-side execution:
Initialize global model parameters, w
for each training round $t = 1, 2, \ldots, T$ **do**
  for each client $c \in C$ **do**
    $g_c \leftarrow \text{ClientUpdate}(c, w_t)$
  end for
  $\bar{g} \leftarrow \frac{1}{|C|} \sum_{c \in C} g_c$ ▷ Aggregate local updates
  $w_{t+1} \leftarrow$ Update global model using $\bar{g}$
end for

procedure ClientUpdate(c, w)
  $w_c \leftarrow$ Update local parameters with w
  for local epoch e from 1 to E **do**
    $g_c \leftarrow \nabla f_c(w_c)$ ▷ Local parameter update
  end for
  Send g_c back to the server
end procedure

Table A.1: Extracted temporal and static features

Type	Feature	Description
Temporal	FiO2	Patient's FiO2 value
	Bedside glucose	Patient's glucose level
	Cvp	Patient's cvp value
	Heartrate	Patient's heart rate value
	Noninvasivediastolic	Patient's non invasive diastolic value
	Noninvasivemean	Patient's non invasive mean value
	Noninvasivesystolic	Patient's non invasive systolic value
	Respiration	Patient's respiration value
	Sao2	Patient's spO2 value
	St1	Patient's st1 value
	St2	Patient's st2 value
	St3	Patient's st3 value
	Systemicdiastolic	Patient's diastolic value
	Systemicmean	Patient's mean pressure
	Systemicsystolic	Patient's systolic value
	Temperature	Patient's temperature value in celsius
	Hour	Time since admission
Static	Hospitalid	Surrogate key for the hospital
	Gender	Gender of the patient
	Age	Patient's age in full years
	Admissionheight	Admission height of the patient in cm
	Admissionweight	Admission weight of the patient in kg
	Intubated	Whether patient is intubated at the time of the worst ABG result
	Vent	Whether patient is ventilated at the worst respiratory rate
	Dialysis	Whether patient is on dialysis
	Eyes	GCS score (1 to 4)
	Motor	GCS score (1 to 5)
	Verbal	GCS score (1 to 6)
	Meds	Whether GCS score could not be obtained due to meds
	Ethnicity	Patient's ethnicity
	Unittype	The picklist unit type of the unit
	Unitadmitsource	Picklist location from where the patient was admitted
	Unitstaytype	Patient's unit stay type
	Physicianspeciality	Picklist specialty of the care provider
	> 89	Whether patient is over 89 years old

Algorithm 9 Fed-ARC: Federated Learning with All Recruited Clients

Server-side execution:
Initialize global model parameters, w
for each client $c \in C$ **do**
  $P_c, n_c, H_c \leftarrow \text{ClientReport}(c)$
end for
Calculate global statistics: $P_g = \sum_{c \in C} P_c$, $n_g = \sum_{c \in C} n_c$
for each client $c \in C$ **do**
  Calculate local representativeness:

$$\nu_c = \gamma_{dv} \left| \frac{P_g}{n_g} - \frac{P_c}{n_c} \right| + \gamma_{sa} n_c^{-0.5} + \gamma_{tr} \theta_c,$$

end for
Select top k clients to yield C_k

for each training round $t = 1, 2, \ldots, T$ **do**
  for each client $c \in C_k$ **do**
    $g_c \leftarrow \text{ClientUpdate}(c, w_t)$
  end for
  $\bar{g} \leftarrow \frac{1}{|C_k|} \sum_{c \in C_k} g_c$ ▷ Aggregate local updates
  $w_{t+1} \leftarrow$ Update global model using $\bar{g}$
end for

procedure ClientReport(c)
  $P_c \leftarrow$ Target distribution at client c
  $n_c \leftarrow$ Number of data points at client c
  $H_c \leftarrow$ Computational capacity of client c
  return (P_c, n_c, H_c)
end procedure

procedure ClientUpdate(c, w)
  $w_c \leftarrow$ Update local parameters with w
  for local epoch e from 1 to E **do**
    $g_c \leftarrow \nabla f_c(w_c)$ ▷ Local parameter update
  end for
  Send g_c back to the server
end procedure

Bibliography

[1] Al-Dailami, A., Kuang, H., and Wang, J. Predicting length of stay in ICU and mortality with temporal dilated separable convolution and context-aware feature fusion. *Comput. Biol. Med. 151*, Pt A (Dec. 2022), 106278.

[2] Antunes, R. S., André da Costa, C., Küderle, A., Yari, I. A., and Eskofier, B. Federated learning for healthcare: Systematic review and architecture proposal. *ACM Trans. Intell. Syst. Technol. 13*, 4 (May 2022).

[3] Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., and Shmatikov, V. How to backdoor federated learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics* (26–28 Aug 2020), S. Chiappa and R. Calandra, Eds., vol. 108 of *Proceedings of Machine Learning Research*, PMLR, pp. 2938–2948.

[4] Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., Freymann, J. B., Farahani, K., and Davatzikos, C. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific Data 4*, 1 (2017), 170117.

[5] Beam, A. L., and Kohane, I. S. Big data and machine learning in health care. *JAMA 319*, 13 (Apr. 2018), 1317–1318.

[6] Beutel, D. J., Topal, T., Mathur, A., Qiu, X., Fernandez-Marques, J., Gao, Y., Sani, L., Kwing, H. L., Parcollet, T., Gusmão, P. P. d., and Lane, N. D. Flower: A friendly federated learning research framework. *arXiv preprint arXiv:2007.14390* (2020).

[7] Beutel, D. J., Topal, T., Mathur, A., Qiu, X., Parcollet, T., and Lane, N. D. Flower: A friendly federated learning research framework. *CoRR abs/2007.14390* (2020).

[8] Bhardwaj, R., Nambiar, A. R., and Dutta, D. A study of machine learning in healthcare. In *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)* (2017), vol. 2, pp. 236–241.

[9] Birsan, T., and Tiba, D. One hundred years since the introduction of the set distance by dimitrie pompeiu. In *System Modeling and Optimization* (Boston, MA, 2006), F. Ceragioli, A. Dontchev, H. Futura, K. Marti, and L. Pandolfi, Eds., Springer US, pp. 35–39.

[10] Blanchard, P., El Mhamdi, E. M., Guerraoui, R., and Stainer, J. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems* (2017), I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, Curran Associates, Inc.

[11] Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., and Seth, K. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (New York, NY, USA, 2017), CCS '17, Association for Computing Machinery, p. 1175–1191.

[12] Bonawitz, K. A., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konečný, J., Mazzocchi, S., McMahan, H. B., Overveldt, T. V., Petrou, D., Ramage, D., and Roselander, J. Towards federated learning at scale: System design. *CoRR abs/1902.01046* (2019).

[13] Callahan, A., and Shah, N. H. Chapter 19 - machine learning in healthcare. In *Key Advances in Clinical Informatics*, A. Sheikh, K. M. Cresswell, A. Wright, and D. W. Bates, Eds. Academic Press, 2017, pp. 279–291.

[14] Centers for Medicare & Medicaid Services. The Health Insurance Portability and Accountability Act of 1996 (HIPAA). Online at http://www.cms.hhs.gov/hipaa/, 1996.

[15] Ceulemans, H., Wuyts, R., Verachtert, W., Simm, J., Arany, A., Moreau, Y., and Herzeel, C. Secure broker-mediated data analysis and prediction. `https://patents.google.com/patent/US20190102670A1/en`, Apr. 2019. US Patent Application US20190102670A1, filed 2017-10-02, published 2019-04-04. Assigned to Katholieke Universiteit Leuven, Janssen Pharmaceutica NV, and IMEC.

[16] Chen, H., Zheng, G., and Ji, Y. Generating hierarchical explanations on text classification via feature interaction detection. In *Proceedings of*

the 58th Annual Meeting of the Association for Computational Linguistics (Online, July 2020), D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Association for Computational Linguistics, pp. 5578–5593.

[17] Cho, K., van Merrienboer, B., Gülçehre, Ç., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR abs/1406.1078* (2014).

[18] Cho, Y. J., Wang, J., and Joshi, G. Client selection in federated learning: Convergence analysis and power-of-choice selection strategies. *CoRR abs/2010.01243* (2020).

[19] Ciampi, M., Sicuranza, M., and Silvestri, S. A privacy-preserving and standard-based architecture for secondary use of clinical data. *Information 13*, 2 (2022).

[20] Deist, T. M., Dankers, F. J. W. M., Ojha, P., Scott Marshall, M., Janssen, T., Faivre-Finn, C., Masciocchi, C., Valentini, V., Wang, J., Chen, J., Zhang, Z., Spezi, E., Button, M., Jan Nuyttens, J., Vernhout, R., van Soest, J., Jochems, A., Monshouwer, R., Bussink, J., Price, G., Lambin, P., and Dekker, A. Distributed learning on 20 000+ lung cancer patients - the personal health train. *Radiother Oncol 144* (Jan. 2020), 189–200.

[21] Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR abs/1810.04805* (2018).

[22] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[23] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[24] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[25] Dwork, C., and Roth, A. 2014.

[26] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature 542*, 7639 (Feb 2017), 115–118.

[27] Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., and Dean, J. A guide to deep learning in healthcare. *Nature Medicine 25*, 1 (Jan 2019), 24–29.

[28] European Commission. 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.

[29] Fallah, A., Mokhtari, A., and Ozdaglar, A. Personalized federated learning with meta-learning. In *Proceedings of the 34th Conference on Neural Information Processing Systems* (2020), pp. 2221–2230.

[30] Fawcett, T. An introduction to roc analysis. *Pattern Recognition Letters 27*, 8 (2006), 861–874. ROC Analysis in Pattern Recognition.

[31] Fredrikson, M., Jha, S., and Ristenpart, T. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver, CO, USA, October 12-16, 2015* (2015), I. Ray, N. Li, and C. Kruegel, Eds., ACM, pp. 1322–1333.

[32] Ganeshan, D., Rosenkrantz, A. B., Bassett Jr, R. L., Williams, L., Lenchik, L., and Yang, W. Burnout in academic radiologists in the united states. *Academic radiology 27*, 9 (2020), 1274–1281.

[33] Geyer, R. C., Klein, T., and Nabi, M. Differentially private federated learning: A client level perspective, 2018.

[34] Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C. K., and Stanley, H. E. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation 101*, 23 (June 2000), E215–20.

[35] Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., and Webster, D. R. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA 316*, 22 (12 2016), 2402–2410.

[36] Gupta, K., Fournarakis, M., Reisser, M., Louizos, C., and Nagel, M. Quantization robust federated learning for efficient inference on heterogeneous devices, 2022.

[37] Hard, A., Kiddon, C. M., Ramage, D., Beaufays, F., Eichner, H., Rao, K., Mathews, R., and Augenstein, S. Federated learning for mobile keyboard prediction, 2018.

[38] He, C., Li, S., So, J., Zhang, M., Wang, H., Wang, X., Vepakomma, P., Singh, A., Qiu, H., Shen, L., Zhao, P., Kang, Y., Liu, Y., Raskar, R., Yang, Q., Annavaram, M., and Avestimehr, S. Fedml: A research library and benchmark for federated machine learning. *Advances in Neural Information Processing Systems, Best Paper Award at Federate Learning Workshop* (2020).

[39] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Klambauer, G., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a nash equilibrium. *CoRR abs/1706.08500* (2017).

[40] Heyndrickx, W., Mervin, L., Morawietz, T., Sturm, N., Friedrich, L., Zalewski, A., Pentina, A., Humbeck, L., Oldenhof, M., Niwayama, R., Schmidtke, P., Fechner, N., Simm, J., Arany, A., Drizard, N., Jabal, R., Afanasyeva, A., Loeb, R., Verma, S., Harnqvist, S., Holmes, M., Pejo, B., Telenczuk, M., Holway, N., Dieckmann, A., Rieke, N., Zumsande, F., Clevert, D.-A., Krug, M., Luscombe, C., Green, D., Ertl, P., Antal, P., Marcus, D., Do Huu, N., Fuji, H., Pickett, S., Acs, G., Boniface, E., Beck, B., Sun, Y., Gohier, A., Rippmann, F., Engkvist, O., Göller, A. H., Moreau, Y., Galtier, M. N., Schuffenhauer, A., and Ceulemans, H. Melloddy: Cross-pharma federated learning at unprecedented scale unlocks benefits in qsar without compromising proprietary information. *Journal of Chemical Information and Modeling 64*, 7 (Apr 2024), 2331–2344.

[41] Hochreiter, S., and Schmidhuber, J. Long short-term memory. *Neural Comput. 9*, 8 (nov 1997), 1735–1780.

[42] Homem-de Mello, T. On rates of convergence for stochastic optimization problems under non–independent and identically distributed sampling. *SIAM Journal on Optimization 19*, 2 (2008), 524–551.

[43] Hsu, T. H., Qi, H., and Brown, M. Measuring the effects of non-identical data distribution for federated visual classification. *CoRR abs/1909.06335* (2019).

[44] Huang, G., Liu, Z., and Weinberger, K. Q. Densely connected convolutional networks. *CoRR abs/1608.06993* (2016).

[45] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R. L., Shpanskaya, K. S., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., and Ng, A. Y. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *CoRR abs/1901.07031* (2019).

[46] Islam, M., Vibashan, V., Jose, V. J. M., Wijethilake, N., Utkarsh, U., and Ren, H. Brain tumor segmentation and survival prediction using 3d attention unet. In *International MICCAI Brainlesion Workshop* (2019), Springer, pp. 262–272.

[47] Jaderberg, M., Dalibard, V., Osindero, S., Czarnecki, W. M., Donahue, J., Razavi, A., Vinyals, O., Green, T., Dunning, I., Simonyan, K., Fernando, C., and Kavukcuoglu, K. Population based training of neural networks. *CoRR abs/1711.09846* (2017).

[48] Jiang, Y., Konecny, J., Rush, K., Konecný, J., and Rush, K. Improving federated learning personalization via model agnostic meta learning. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics* (2019), pp. 1733–1742.

[49] Johnson, A. E. W., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C., Mark, R. G., and Horng, S. MIMIC-CXR: A large publicly available database of labeled chest radiographs. *CoRR abs/1901.07042* (2019).

[50] Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., and Suresh, A. T. SCAFFOLD: stochastic controlled averaging for on-device federated learning. *CoRR abs/1910.06378* (2019).

[51] Krizhevsky, A., and Hinton, G. Learning multiple layers of features from tiny images. Tech. Rep. 0, University of Toronto, Toronto, Ontario, 2009.

[52] Lecun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems 2* (1990), D. S. Touretzky, Ed., Morgan Kaufmann, pp. 396–404.

[53] Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE 86*, 11 (1998), 2278–2324.

[54] Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. Sample complexity of asynchronous q-learning: Sharper analysis and variance reduction, 2021.

[55] Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. Federated learning: Challenges, methods, and future directions. *CoRR abs/1908.07873* (2019).

[56] Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine 37*, 3 (2020), 50–60.

[57] Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems 2020* (2020), vol. 2, pp. 429–450.

[58] Li, T., Sanjabi, M., Beirami, A., and Smith, V. Fair resource allocation in federated learning, 2020.

[59] Li, W., Milletari, F., Xu, D., Rieke, N., Hancox, J., Zhu, W., Baust, M., Cheng, Y., Ourselin, S., Cardoso, M. J., and Feng, A. Privacy-preserving federated brain tumour segmentation. *CoRR abs/1910.00962* (2019).

[60] Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out* (Barcelona, Spain, July 2004), Association for Computational Linguistics, pp. 74–81.

[61] Lin, Y., Han, S., Mao, H., Wang, Y., and Dally, W. J. Deep gradient compression: Reducing the communication bandwidth for distributed training. *CoRR abs/1712.01887* (2017).

[62] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A., van Ginneken, B., and Sánchez, C. I. A survey on deep learning in medical image analysis. *Medical Image Analysis 42* (Dec. 2017), 60–88.

[63] Loshchilov, I., and Hutter, F. Fixing weight decay regularization in adam. *CoRR abs/1711.05101* (2017).

[64] Loshchilov, I., and Hutter, F. Fixing weight decay regularization in adam. *CoRR abs/1711.05101* (2017).

[65] Lu, Z., Pan, H., Dai, Y., Si, X., and Zhang, Y. Federated learning with non-iid data: A survey. *IEEE Internet of Things Journal 11*, 11 (2024), 19188–19209.

[66] Ma, Z., Zhao, M., Cai, X., and Jia, Z. Fast-convergent federated learning with class-weighted aggregation. *Journal of Systems Architecture 117* (2021), 102125.

[67] McMahan, B., Moore, E., Ramage, D., Hampson, S., and Arcas, B. A. y. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (20–22 Apr 2017), A. Singh and J. Zhu, Eds., vol. 54 of *Proceedings of Machine Learning Research*, PMLR, pp. 1273–1282.

[68] McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (2017), PMLR, pp. 1273–1282.

[69] Mehrjou, A., Soleymani, A., Buchholz, A., Hetzel, J., Schwab, P., and Bauer, S. Federated learning in multi-center critical care research: A systematic case study using the eicu database. *CoRR abs/2204.09328* (2022).

[70] Messina, P., Pino, P., Parra, D., Soto, A., Besa, C., Uribe, S., Andía, M., Tejos, C., Prieto, C., and Capurro, D. A survey on deep learning and explainability for automatic report generation from medical images. *ACM Comput. Surv. 54*, 10s (Sept. 2022).

[71] Mhamdi, E. M. E., Guerraoui, R., and Rouault, S. The hidden vulnerability of distributed learning in byzantium, 2018.

[72] Nishio, T., and Yonetani, R. Client selection for federated learning with heterogeneous resources in mobile edge. In *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)* (May 2019), IEEE.

[73] Nishio, T., and Yonetani, R. Client selection for federated learning with heterogeneous resources in mobile edge. In *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)* (2019), pp. 1–7.

[74] Nishio, T., and Yonetani, R. Client selection for federated learning with heterogeneous resources in mobile edge. In *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)* (2019), pp. 1–7.

[75] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (Philadelphia, Pennsylvania, USA, July 2002), P. Isabelle, E. Charniak, and D. Lin, Eds., Association for Computational Linguistics, pp. 311–318.

[76] Paragliola, G. A federated learning-based approach to recognize subjects at a high risk of hypertension in a non-stationary scenario. *Information Sciences 622* (2023), 16–33.

[77] Pati, S., Baid, U., Zenk, M., Edwards, B., Sheller, M., Reina, G. A., Foley, P., Gruzdev, A., Martin, J., Albarqouni, S., Chen, Y., Shinohara, R. T., Reinke, A., Zimmerer, D., Freymann, J. B., Kirby, J. S., Davatzikos, C., Colen, R. R., Kotrotsou, A., Marcus, D., Milchenko, M., Nazeri, A., Fathallah-Shaykh, H., Wiest, R., Jakab, A., Weber, M.-A., Mahajan, A., Maier-Hein, L., Kleesiek, J., Menze, B., Maier-Hein, K., and Bakas, S. The federated tumor segmentation (fets) challenge, 2021.

[78] Pollard, T. J., Johnson, A. E. W., Raffa, J., and Badawi, O. The eICU collaborative research database, 2017.

[79] Pollard, T. J., Johnson, A. E. W., Raffa, J. D., Celi, L. A., Mark, R. G., and Badawi, O. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific Data 5*, 1 (Sep 2018), 180178.

[80] Price, 2nd, W. N., and Cohen, I. G. Privacy in the age of medical big data. *Nat Med 25*, 1 (Jan. 2019), 37–43.

[81] Radford, A., and Narasimhan, K. Improving language understanding by generative pre-training.

[82] Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving language understanding by generative pre-training.

[83] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners.

[84] Rajpurkar, P., Chen, E., Banerjee, O., and Topol, E. J. Ai in health and medicine. *Nature Medicine 28*, 1 (Jan 2022), 31–38.

[85] Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C. P., Patel, B. N., Yeom, K. W., Shpanskaya, K., Blankenberg, F. G., Seekins, J., Amrhein, T. J., Mong, D. A., Halabi, S. S., Zucker, E. J., Ng, A. Y., and Lungren, M. P. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med 15*, 11 (Nov. 2018), e1002686.

[86] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D. Y., Bagul, A., Langlotz, C. P., Shpanskaya, K. S.,

Lungren, M. P., and Ng, A. Y. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *CoRR abs/1711.05225* (2017).

[87] Reina, G. A., Gruzdev, A., Foley, P., Perepelkina, O., Sharma, M., Davidyuk, I., Trushkin, I., Radionov, M., Mokrov, A., Agapov, D., Martin, J., Edwards, B., Sheller, M. J., Pati, S., Moorthy, P. N., Wang, H. S., Shah, P., and Bakas, S. Openfl: An open-source framework for federated learning. *CoRR abs/2105.06413* (2021).

[88] Reiner, B. I., Knight, N., and Siegel, E. L. Radiology reporting, past, present, and future: the radiologist's perspective. *Journal of the American College of Radiology 4*, 5 (2007), 313–319.

[89] Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M., and Cardoso, M. J. The future of digital health with federated learning. *npj Digital Medicine 3*, 1 (Sep 2020), 119.

[90] Rocheteau, E., Liò, P., and Hyland, S. Temporal pointwise convolutional networks for length of stay prediction in the intensive care unit. In *Proceedings of the Conference on Health, Inference, and Learning* (New York, NY, USA, 2021), CHIL '21, Association for Computing Machinery, p. 58–68.

[91] Rocheteau, E., Tong, C., Velickovic, P., Lane, N. D., and Liò, P. Predicting patient outcomes with graph representation learning. *CoRR abs/2101.03940* (2021).

[92] Ruan, Y., Zhang, X., and Joe-Wong, C. How valuable is your data? optimizing client recruitment in federated learning. In *2021 19th International Symposium on Modeling and Optimization in Mobile, Ad hoc, and Wireless Networks (WiOpt)* (2021), pp. 1–8.

[93] Sahu, A. K., Li, T., Sanjabi, M., Zaheer, M., Talwalkar, A., and Smith, V. On the convergence of federated optimization in heterogeneous networks. *CoRR abs/1812.06127* (2018).

[94] Sattler, F., Müller, K.-R., and Samek, W. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE Transactions on Neural Networks and Learning Systems 32*, 8 (2021), 3710–3722.

[95] Scheltjens, V., Momo, L. N. W., Verbeke, W., and De Moor, B. Client recruitment for federated learning in icu length of stay prediction. In *2023 IEEE 19th International Conference on e-Science (e-Science)* (2023), pp. 1–9.

[96] Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., and Batra, D. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR abs/1610.02391* (2016).

[97] Shailaja, K., Seetharamulu, B., and Jabbar, M. A. Machine learning in healthcare: A review. In *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)* (2018), pp. 910–914.

[98] Shapiro, A. Monte carlo sampling methods. In *Stochastic Programming*, vol. 10 of *Handbooks in Operations Research and Management Science*. Elsevier, 2003, pp. 353–425.

[99] Sharma, H., Jain, J. S., Bansal, P., and Gupta, S. Feature extraction and classification of chest x-ray images using cnn to detect pneumonia. In *2020 10th international conference on cloud computing, data science & engineering (Confluence)* (2020), IEEE, pp. 227–231.

[100] Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D., Colen, R. R., and Bakas, S. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci. Rep. 10*, 1 (July 2020), 12598.

[101] Sheller, M. J., Reina, G. A., Edwards, B., Martin, J., and Bakas, S. Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* (Cham, 2019), A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, and T. van Walsum, Eds., Springer International Publishing, pp. 92–104.

[102] Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models, 2017.

[103] Simon, A. F., Holmes, J. H., and Schwartz, E. S. Decreasing radiologist burnout through informatics-based solutions. *Clinical Imaging 59*, 2 (2020), 167–171.

[104] Tan, A. Z., Yu, H., Cui, L., and Yang, Q. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems 34*, 12 (2023), 9587–9603.

[105] Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., and Ting, D. S. W. Large language models in medicine. *Nature medicine 29*, 8 (2023), 1930–1940.

[106] van Soest, J., Sun, C., Mussmann, O., Puts, M., van den Berg, B., Malic, A., van Oppen, C., Towend, D., Dekker, A., and Dumontier, M. Using the personal health train for automated and Privacy-Preserving analytics on vertically partitioned data. *Stud Health Technol Inform 247* (2018), 581–585.

[107] Vandenberghe, A., Wamba Momo, L. N., Scheltjens, V., and De Moor, B. Multimodal deep learning for early length of stay prediction using patient similarity embeddings. In *Proc. of BNAIC/BeNeLearn* (2022), Mechelen, Belgium.

[108] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2023.

[109] Vedantam, R., Zitnick, C. L., and Parikh, D. Cider: Consensus-based image description evaluation, 2015.

[110] Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. Show and tell: A neural image caption generator. *CoRR abs/1411.4555* (2014).

[111] Wang, H., Kaplan, Z., Niu, D., and Li, B. Optimizing federated learning on non-iid data with reinforcement learning. In *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications* (2020), pp. 1698–1707.

[112] Wang, J., Liu, Q., Liang, H., Joshi, G., and Poor, H. V. Tackling the objective inconsistency problem in heterogeneous federated optimization. *CoRR abs/2007.07481* (2020).

[113] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Los Alamitos, CA, USA, jul 2017), IEEE Computer Society, pp. 3462–3471.

[114] Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing 13*, 4 (2004), 600–612.

[115] Wolny, A., Cerrone, L., Vijayan, A., Tofanelli, R., Barro, A. V., Louveaux, M., Wenzl, C., Strauss, S., Wilson-Sánchez, D., Lymbouridou, R., Steigleder, S. S., Pape, C., Bailoni, A., Duran-Nebreda, S., Bassel, G. W., Lohmann, J. U., Tsiantis, M., Hamprecht, F. A., Schneitz, K., Maizel, A., and Kreshuk, A. Accurate and versatile 3d segmentation of plant tissues at cellular resolution. *eLife 9* (jul 2020), e57613.

[116] Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., and Zhang, L. Cvt: Introducing convolutions to vision transformers, 2021.

[117] Xia, W., Quek, T. Q. S., Guo, K., Wen, W., Yang, H. H., and Zhu, H. Multi-armed bandit-based client scheduling for federated learning. *IEEE Transactions on Wireless Communications 19*, 11 (2020), 7108–7123.

[118] Xie, C., Koyejo, S., and Gupta, I. Asynchronous federated optimization, 2020.

[119] Xu, J., Glicksberg, B. S., Su, C., Walker, P., Bian, J., and Wang, F. Federated learning for healthcare informatics. *Journal of Healthcare Informatics Research 5*, 1 (Mar 2021), 1–19.

[120] Xu, J., Glicksberg, B. S., Su, C., Walker, P., Bian, J., and Wang, F. Federated learning for healthcare informatics. *Journal of Healthcare Informatics Research 5*, 1 (2021), 1–19.

[121] Xu, J., and Wang, H. Client selection and bandwidth allocation in wireless federated learning networks: A long-term perspective. *IEEE Transactions on Wireless Communications 20*, 2 (2021), 1188–1200.

[122] Yang, Q., Liu, Y., Chen, T., and Tong, Y. Federated machine learning: Concept and applications, 2019.

[123] Yoshida, N., Nishio, T., Morikura, M., and Yamamoto, K. Mab-based client selection for federated learning with uncertain resources in mobile networks. In *2020 IEEE Globecom Workshops (GC Wkshps* (2020), pp. 1–6.

[124] Yurochkin, M., Agarwal, M., Ghosh, S., Greenewald, K., Hoang, T. N., and Khazaeni, Y. Bayesian nonparametric federated learning of neural networks, 2019.

[125] Zhang, H., Xie, Z., Zarei, R., Wu, T., and Chen, K. Adaptive client selection in resource constrained federated learning systems: A deep reinforcement learning approach. *IEEE Access 9* (2021), 98423–98432.

[126] Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., and Gao, J. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021), pp. 5579–5588.

[127] Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Learning to prompt for vision-language models. *International Journal of Computer Vision 130*, 9 (2022), 2337–2348.

[128] Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592* (2023).

[129] Zhu, H., Xu, J., Liu, S., and Jin, Y. Federated learning on non-iid data: A survey. *CoRR abs/2106.06843* (2021).

[130] Zhu, H., Xu, J., Liu, S., and Jin, Y. Federated learning on non-iid data: A survey. *Neurocomputing 465* (2021), 371–390.

[131] Zou, K. H., Warfield, S. K., Bharatha, A., Tempany, C. M. C., Kaus, M. R., Haker, S. J., Wells, 3rd, W. M., Jolesz, F. A., and Kikinis, R. Statistical validation of image segmentation quality based on a spatial overlap index. *Acad Radiol 11*, 2 (Feb. 2004), 178–189.

Statement on the use of Generative AI

Generative AI assistance tools were used for the purpose of language assistance and generating programming code. For language assistance, the OpenAI models were leveraged. For generating programming code a combination of the models made available by OpenAI and Anthropic were used. The primary use for code generation existed in writing boilerplate code, general structures for frameworks and modularizing and refactoring the code base.
The text, code, and images in this thesis are my own (unless otherwise specified). Generative AI has only been used in accordance with the KU Leuven guidelines and appropriate references have been added. I have reviewed and edited the content as needed and I take full responsibility for the content of the thesis.

FACULTY OF ECONOMICS AND BUSINESS
RESEARCH CENTRE FOR INFORMATION SYSTEMS ENGINEERING
Naamsestraat 69 bus 3500
3000 Leuven

