Federated learning of biomedical data in multicentric imaging studies

Santiago Smith Silva Rincon

▶ **To cite this version:**

Santiago Smith Silva Rincon. Federated learning of biomedical data in multicentric imaging studies. Medical Imaging. Université Côte d'Azur, 2023. English. NNT : 2023COAZ4056 . tel-04417044v1

HAL Id: tel-04417044

https://theses.hal.science/tel-04417044v1

Submitted on 19 Oct 2023 (v1), last revised 25 Jan 2024 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ CÔTE D'AZUR

ÉCOLE DOCTORALE
SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

Thèse de Doctorat

Apprentissage Fédéré de Données Biomédicales dans les Études d'Imagerie Multicentriques

Santiago Smith Silva Rincon

Inria, Équipe Epione

Thèse dirigée par Marco Lorenzi et co-dirigée par Barbara Bardoni

Soutenue le 18 juillet 2023

Présentée en vue de l'obtention du grade de Docteur en Automatique, Traitement du Signal et des Images de l'Université Côte d'Azur.

Devant le jury composé de :

James Cole	Dementia Research Centre, Institute of Neurology, UCL	Rapporteur
Jean-F. Mangin	Univ. Paris-Saclay, CEA, CNRS, NeuroSpin, BAOBAB	Rapporteur
Marco Lorenzi	Centre INRIA d'Université Côte d'Azur	Directeur de thèse
Barbara Bardoni	Centre INRIA d'Université Côte d'Azur	Co-directeur de thèse
Jonas Richiardi	Lausanne University Hospital & Univ. of Lausann	Invité
André Altmann	Centre for Medical Image Computing (CMIC), Dept. of Medical Physics and Biomedical Engineering, UCL	Invité

Inria

Federated Learning of Biomedical Data in Multicentric Imaging Studies

SANTIAGO SMITH SILVA RINCON

INRIA, EPIONE Team

Supervised by Marco LORENZI

Co-supervised by Barbara BARDONI

Defended on July 18, 2023

Presented to obtain the title of
DOCTEUR EN AUTOMATIQUE, TRAITEMENT DU SIGNAL ET DES IMAGES
of the UNIVERSITÉ CÔTE D'AZUR

Jury:

James COLE	Dementia Research Centre, Institute of Neurology, UCL	Reviewer
Jean-F. MANGIN	Univ. Paris-Saclay, CEA, CNRS, NeuroSpin, BAOBAB	Reviewer
Marco LORENZI	Centre INRIA d'Université Côte d'Azur	Supervisor
Barbara BARDONI	Centre INRIA d'Université Côte d'Azur	Co-supervisor
Jonas RICHIARDI	Lausanne University Hospital & Univ. of Lausann	Invited
André ALTMANN	Centre for Medical Image Computing (CMIC), Dept. of Medical Physics and Biomedical Engineering, UCL	Invited

Abstract

In order to gather sufficient sample size and representativity of clinical populations, the multi-centric analysis paradigm is often adopted for statistical and machine learning studies of biomedical data, particularly in the field of neuroimaging. Conventional multi-centric analysis paradigms are based on meta-analysis and mega-analysis, often in conjunction with data harmonization, to account for systematic biases and improve the combined analysis of data from multiple sources. However, while meta-analyses are mainly suited for standard statistical testing only, mega-analyses require centralizing the data, which can undermine data privacy and security. Today, data protection regulations such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) impose strict governance on sensitive patient information, significantly limiting researchers' access to such data.

Federated learning (FL) is an alternative paradigm to multi-centric studies enabling multiple parties to train a model collaboratively without sharing sensitive data. FL thus addresses data governance challenges while enhancing patients' data privacy. However, to facilitate real-life applications of FL, a series of challenges must be addressed: i) heterogeneity and generalization due to differences in data distributions and discrepancies across different institutions which can result in biased models that do not generalize properly, ii) occasional requirements of considerable amounts of computational resources that hospitals or institutions may not have, limiting its practicality and, ii) a common framework and infrastructure to put in place real-life applications while fulfilling research and governance demands. This thesis aims to contribute to the evolving landscape of neuroimaging research by investigating the potential of FL to transform the way researchers collaborate and analyze data, ultimately paving the way for more efficient and effective advancements in neuroimaging.

We start by addressing the issue of data heterogeneity in federated learning setups, by introducing two methods, namely "Fed-ComBat" and "federated mixed-effect modeling", which aim to perform data harmonization and modeling on heterogeneous data respectively.

Secondly, we introduce a black-box optimization scheme for FL aiming to improve the optimization process in federated setups. This method is based on gradient-free optimization of a global model, through the collaborative iterative refinement of the

cost function associated with the distributed optimization problem across clients. This approach aims to centralize computational costs and mitigate overfitting issues linked to gradient-descent-based approaches while enabling institutions and hospitals with limited computational resources to participate in federated learning setups while achieving accurate and generalizable models.

Finally, to enable and empower real-life federated applications, we introduce Fed-BioMed as an open-source framework for federated learning in healthcare, aiming to fulfill the need for a common collaborative framework that is also compliant with privacy and ethical standards.

Overall, this thesis comprises methodological and technical contributions that tackle the challenges of data heterogeneity, optimization, and infrastructure in federated learning setups for neuroimaging research, with the ultimate goal of facilitating more efficient and effective advancements in healthcare while preserving patient privacy and data governance.

Keywords: federated learning, healthcare, data protection, GDPR, CCPA, medical imaging, data harmonization, meta-analysis, mega-analysis, Fed-BioMed, Bayesian optimization, random effect models, FedComBat.

Résumé

Afin de rassembler une taille d'échantillon suffisante et une représentativité des populations cliniques, le paradigme de l'analyse multi-centrique est souvent adopté pour les études statistiques et d'apprentissage automatique des données biomédicales, en particulier dans le domaine de la neuroimagerie. Les paradigmes d'analyse multi-centrique conventionnels reposent sur la méta-analyse et la méganalyse, souvent conjointement avec l'harmonisation des données, pour tenir compte des biais systématiques et améliorer l'analyse combinée des données provenant de sources multiples. Cependant, alors que les méta-analyses sont principalement adaptées aux tests statistiques standard, les méganalyses nécessitent une centralisation des données, ce qui peut nuire à la confidentialité et à la sécurité des données. Aujourd'hui, les réglementations sur la protection des données telles que le Règlement Général sur la Protection des Données (RGPD) et la California Consumer Privacy Act (CCPA) imposent une gouvernance stricte sur les informations sensibles des patients, limitant considérablement l'accès des chercheurs à ces données.

L'apprentissage fédéré (FL) est un paradigme alternatif aux études multi-centriques permettant à plusieurs parties de former un modèle en collaboration sans partager de données sensibles. Le FL répond ainsi aux défis de la gouvernance des données tout en améliorant la confidentialité des données des patients. Cependant, pour faciliter les applications réelles du FL, une série de défis doit être relevée : i) l'hétérogénéité et la généralisation en raison des différences dans les distributions de données et les écarts entre les différentes institutions qui peuvent entraîner des modèles biaisés qui ne se généralisent pas correctement, ii) les besoins occasionnels en ressources de calcul considérables que les hôpitaux ou les institutions peuvent ne pas avoir, limitant ainsi sa praticité et, ii) un cadre et une infrastructure communs pour mettre en place des applications réelles tout en répondant aux exigences de recherche et de gouvernance. Cette thèse vise à contribuer au paysage évolutif de la recherche en neuroimagerie en étudiant le potentiel du FL pour transformer la manière dont les chercheurs collaborent et analysent les données, ouvrant ainsi la voie à des avancées plus efficaces et efficientes en neuroimagerie.

Nous commençons par aborder la question de l'hétérogénéité des données dans les configurations d'apprentissage fédéré, en introduisant deux méthodes, à savoir "Fed-ComBat" et "modélisation à effets mixtes fédérée", qui visent à réaliser respectivement l'harmonisation des données et la modélisation sur des données hétérogènes.

Deuxièmement, nous introduisons un schéma d'optimisation boîte noire pour le FL visant à améliorer le processus d'optimisation dans les configurations fédérées. Cette méthode est basée sur l'optimisation sans gradient d'un modèle global, grâce à l'affinement itératif collaboratif de la fonction de coût associée au problème d'optimisation distribué entre les clients. Cette approche vise à centraliser les coûts de calcul et à atténuer les problèmes de surajustement liés aux approches basées sur la descente de gradient, tout en permettant aux institutions et aux hôpitaux disposant de ressources informatiques limitées de participer aux configurations d'apprentissage fédéré tout en obtenant des modèles précis et généralisables.

Enfin, pour permettre et renforcer les applications fédérées dans la vie réelle, nous présentons Fed-BioMed, un cadre open source pour l'apprentissage fédéré dans le domaine de la santé, visant à répondre au besoin d'un cadre collaboratif commun qui est également conforme aux normes de confidentialité et d'éthique.

Dans l'ensemble, cette thèse comprend des contributions méthodologiques et techniques qui abordent les défis de l'hétérogénéité des données, de l'optimisation et de l'infrastructure dans les configurations d'apprentissage fédéré pour la recherche en neuroimagerie, avec pour objectif ultime de faciliter des avancées plus efficaces et efficientes dans les soins de santé tout en préservant la confidentialité des patients et la gouvernance des données.

Mots-clés: apprentissage fédéré, santé, protection des données, RGPD, CCPA, imagerie médicale, harmonisation des données, méta-analyse, méganalyse, Fed-BioMed, optimisation bayésienne, modèles à effets aléatoires, FedComBat.

Financial Support

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 847579 (Marie Skłodowska-Curie Actions), and by the Centre INRIA d'Université Côte d'Azur "NEF" computation cluster.

Data Use Agreements

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (https://adni.loni.usc.edu/).

Data used in this article were provided by the Australian Imaging Biomarkers and Lifestyle flagship study of ageing (AIBL) funded by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) which was made available at the ADNI database (https://adni.loni.usc.edu/aibl-australian-imaging-biomarkers-and-lifestyle-study-of-ageing-18-month-data-now-released/).

Data used in the preparation of this article were obtained from the Autism Brain Imaging Data Exchange (ABIDE) I database.

Data used in this article were obtained from the Autism Brain Imaging Data Exchange (ABIDE) II database (http://fcon_1000.projects.nitrc.org/indi/abide/abide_II.html).

The AIBL researchers contributed data but did not participate in the analysis or writing of this report. AIBL researchers are listed at https://www.aibl.csiro.au.

Data used in the preparation of this article were obtained from the MIRIAD database.

Data used in the preparation of this article were obtained from the Open Access Series of Imaging Studies (OASIS) database.

Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database (http://www.ppmi-info.org/).

Data used in the preparation of this article were obtained from the Pre-symptomatic Evaluation of Experimental or Novel Treatments for Alzheimer's Disease (A4) study.

Contents

List of abbreviations

Federated Learning and Collaborative Computing

FL	Federated Learning
TFF	TensorFlow Federated
FC	Federated Core
MPC	Multi-Party Computation
FedAvg	Federated Averaging

Research Initiatives and Datasets

ADNI	Alzheimer's Disease Neuroimaging Initiative
AIBL	Australian Imaging Biomarkers and Lifestyle
CSIRO	Commonwealth Scientific and Industrial Research Organisation
ABIDE	Autism Brain Imaging Data Exchange
OASIS	Open Access Series of Imaging Studies
PPMI	Parkinson's Progression Markers Initiative
MIRIAD	Minimal Interval Resonance Imaging in Alzheimer's Disease
COINSTAC	The Collaborative Informatics and Neuroimaging Suite Toolkit for Anonymous Computation

Machine Learning and Optimization

ML	Machine Learning
BO	Bayesian Optimization
GP	Gaussian Process
BBMM	Black-Box Matrix Multiplication
BBO	Black Box Optimization
PI	Probability of Improvement
EI	Expected Improvement
UCB	Upper Confidence Bound
SGD	Stochastic Gradient Descent

Data Privacy and Regulations

CCPA	California Consumer Privacy Act
GDPR	General Data Protection Regulation
VPN	Virtual Private Network

Neurological Disorders and Conditions

PD	Parkinson's Disease
AD	Alzheimer's Disease
ASD	Autism Spectrum Disorder
MCI	Mild Cognitive Impairment

Imaging and Analysis Techniques

MRI	Magnetic Resonance Imaging
EEG	Electroencephalography
PET	Positron Emission Tomography
CT	Computed Tomography
DTI	Diffusion Tensor Imaging

Statistical Analysis and Evaluation

RMSE	Root Mean Square Error
MAE	Mean Absolute Error
AIC	Akaike Information Criterion
ANOVA	Analysis of Variance

Others

EB	Empirical Bayes
GAM	Generalized Additive Models
IID	Independent and Identically Distributed
PCA	Principal Component Analysis
VAE	Variational Autoencoder
ICV	Intracranial Volume

1 Introduction

1.1 Clinical context

Rapid advancements in information technology and the digitization of healthcare services have led to a remarkable growth in the volume and variety of data generated by hospitals and healthcare institutions. Electronic health records (EHRs), medical imaging, wearable devices, and genomic data are just a few examples of the diverse sources contributing to this data explosion [Undavia et al., 2020; Miotto et al., 2018]. The increasing adoption of telemedicine, remote monitoring systems, and mobile health applications further accelerates the production of such digital health information [Wosik et al., 2020; Torous et al., 2018]. As a result: massive and complex datasets, often referred to as "big data," which hold the potential to revolutionize medical research, improve patient care, and streamline healthcare operations [Murdoch et al., 2013].

The wealth of information contained in these datasets offers numerous opportunities for data-driven research, including the identification of new biomarkers, optimization of treatment strategies, and development of personalized medicine approaches [Schneeweiss,

2014; Jensen et al., 2012]. Additionally, the analysis of this data can provide valuable insights into healthcare service delivery and quality of care [Rumsfeld et al., 2016]. Given the geographical distribution and immense volume of such data, storage often takes place across multiple, geographically dispersed facilities, each with its own unique patient population and data characteristics. To leverage the advantages of this geographical and demographic diversity, multicentric studies have emerged as a key initiative in the scientific community. These studies pool data from numerous sources, thus creating larger and more heterogeneous cohorts. In turn, this allows for a more robust and generalized understanding of biological phenomena, such as diseases.

In the field of neurosciences, particularly when dealing with neuroimaging data, heterogeneity plays a pivotal role in medical studies. The success of these studies often depends on population diversity, including demographic differences and other factors that may be outside of primary scientific interest [Benkarim et al., 2022]. The understanding of neurodegenerative diseases, such as Alzheimer's disease or Parkinson's, is also highly impacted by such heterogeneity [Maito et al., 2023; Devignes et al., 2022; Ibanez et al., 2021]. This implies that larger and more diverse datasets are crucial for uncovering relevant findings and advancing our knowledge of these complex disorders, ultimately contributing to improved healthcare outcomes and patient care.

However, the vast volumes and complexity of neuroimaging data also pose several challenges, including data storage, management, integration, and analysis [Sun et al., 2013; Webb-Vargas et al., 2017]. The sensitive nature of health information necessitates strict adherence to data protection regulations such as the General Data Protection Regulation (GDPR) [Voigt et al., 2017] and the California Consumer Privacy Act (CCPA) [Solove et al., 2020], which can further complicate data sharing and collaboration efforts. The growing need to harness the full potential of healthcare data while overcoming these challenges has motivated significant interest in the development of novel computational methods and infrastructures tailored to the unique requirements of medical research.

1.2 Data Protection Regulations: GDPR and CCPA

As the volume of healthcare data continues to grow, so does the need for robust data protection and privacy regulations. Health information is often sensitive, and its unauthorized disclosure can have serious consequences for individuals and organizations alike. To address these concerns, governments have enacted strict data protection laws, such as the European Union's General Data Protection Regulation (GDPR) [Voigt et al., 2017] and the United States' California Consumer Privacy Act (CCPA) [Solove et al., 2020].

The GDPR came into effect in 2018 as a comprehensive data protection regulation applicable to all EU member states, aims to unify data privacy laws across Europe and empower individuals with greater control over their personal information. The GDPR imposes stringent requirements on organizations that process personal data, including the need to obtain explicit consent from individuals, the right to data portability, and the right to be forgotten [Voigt et al., 2017]. Additionally, the GDPR mandates that organizations implement appropriate technical and organizational measures to ensure data security and privacy, with significant penalties for non-compliance.

The CCPA, a data protection regulation similar to GDPR, was enacted in 2020 by the state of California in the United States. It grants Californian residents several rights concerning their personal information, such as the right to know what personal data is collected, the right to delete personal information held by businesses, and the right to opt-out of the sale of personal data [Solove et al., 2020]. Although the CCPA is not as extensive as the GDPR, it serves as a significant step toward strengthening data protection and privacy in the United States.

Compliance with data protection regulations such as the GDPR and CCPA is a critical consideration for medical researchers working with healthcare data. It can be particularly challenging in the context of multi-institutional collaborations and large-scale data sharing efforts, where data must be anonymized and de-identified to protect patient privacy while preserving its utility for research purposes [El Emam et al., 2014]. Moreover, these regulations may limit data accessibility, potentially hindering the development and deployment of innovative data-driven solutions in healthcare.

As a result, there is a growing interest in exploring alternative approaches, such as federated learning, that enable collaborative research and model development without the need to share raw patient data [Brisimi et al., 2018; Higgins et al., 2019]. In the next section, we discuss the most common methods used for multicentric data analysis, and how federated learning may offer a better solution.

1.3 Multi-centric studies in neuroimaging

To address the challenges and maximize the potential of large scale neuroimaging data sets, researchers often rely on multicentric studies involving the collaboration of multiple institutions and research centers. These studies enable the aggregation of larger and more diverse datasets, thereby increasing the statistical power and generalizability of the findings. Within the context of multicentric studies, conventional approaches are established such as meta-analysis [Glass, 1976] and mega-analysis [Costafreda, 2009] in order to facilitate collaborative research efforts. Meta-analysis and mega-

analysis focus on pooling summary statistics or raw data, respectively, from individual studies. Nevertheless, these methods come with their own set of challenges, particularly concerning data privacy and governance concerns.

1.3.1 Meta-analysis and Mega-analysis

Meta-analysis and mega-analysis are two well-established techniques for combining and synthesizing information from multiple independent studies. Both methods aim to increase statistical power, improve effect size estimates, and provide more generalizable results by pooling data from individual studies [Hedges, 1992; Higgins et al., 2019]. However, they differ in their respective approaches and underlying assumptions.

Meta-analysis is a statistical technique that combines the effect size estimates from multiple studies to derive an overall effect size estimate. The method generally involves two steps. First, an effect size measure (e.g., odds ratio, risk ratio, or standardized mean difference) is calculated for each study included in the analysis. Then, a weighted average of these effect size measures is computed, with weights typically assigned based on the inverse of the variance of each study's effect size estimate. This process can be represented mathematically as:

$$\hat{\theta}_{meta} = \frac{\sum_{i=1}^{k} w_i \hat{\theta}_i}{\sum_{i=1}^{k} w_i}, \tag{1.1}$$

where $\hat{\theta}_{meta}$ is the overall effect size estimate, k is the number of studies, $\hat{\theta}_i$ is the effect size estimate for study i, and w_i is the weight assigned to study i [Hedges, 1992]. Meta-analysis can employ either fixed-effects or random-effects models, depending on the assumptions made about the true effect sizes and the between-study heterogeneity [Higgins et al., 2019].

Mega-analysis, on the other hand, is a large-scale, collaborative research approach that involves the analysis of individual participant data (IPD) from multiple studies. This approach is particularly useful when the original studies use different outcome measures, as the individual-level data can be harmonized and analyzed in a consistent manner. Unlike meta-analysis, which involves compiling results from published studies, mega-analysis focuses on directly analyzing IPD using an agreed-upon processing strategy. This approach offers several advantages over traditional meta-analysis, such as improved consistency in inclusion criteria across sites, better treatment of confounds and missing data, verification of statistical model assumptions, standardized procedures, increased statistical power, and reduced biases.

Both meta and mega-analysis require the transfer of data to coordinating facilities. In the case of aggregated data, such as histograms of quality metrics, effect sizes, confidence intervals, and standard errors, the risk of reidentification is minimal, as the data is not identifiable at the individual level. However, it is important to note that without proper precautions, repeated computation of aggregate results using slightly varying subsets of participants could expose information about individuals [Dwork, 2006]. This risk can be minimized by establishing agreements among researchers regarding the nature and amount of aggregated data to be transferred.

For mega-analyses, where individual participant data (IPD) is transferred, further attention is required due to differences in regulations across sites that protect the confidentiality, integrity, and security of the IPD and their use in human research. Although mega-analyses with IPD have been shown to provide superior results compared to meta-analyses in terms of higher statistical power and acceptable false-positive rates, the major challenge of mega-analysis is the need for at least one site to possess the necessary resources and expertise to handle large datasets. Additionally, this approach is only possible when IPD can be shared with a central facility, which is often limited due to varying data protection regulations among research projects, consortia, and countries [Mathew et al., 1999; Eisenhauer, 2021].

An exemplary illustration in the utilization of meta-analyses and mega-analyses is the Enhancing Neuro Imaging Genetics through Meta-Analysis. The ENIGMA consortium is a global network of researchers that aims to identify genetic and environmental factors that affect brain structure and function using imaging and other measures. The consortium was established in 2009 and has since grown to include over 70 research groups worldwide. Figure 1.1 provides a visual comparison highlighting the distinctions between meta-analyses and mega-analyses, specifically within the context of the ENIGMA consortium [Zugman et al., 2022].

The ENIGMA consortium has conducted many influential studies that have shed light on the genetic and environmental factors that underlie brain development, aging, and disease. For example, a meta-analysis of 94 studies conducted by ENIGMA found that the volume of the hippocampus, a brain region associated with memory and learning, is reduced in individuals with major depressive disorder [Schmaal et al., 2016]. ENIGMA has also investigated the genetic architecture of brain structure and identified novel genetic variants associated with brain volume [Satizabal et al., 2019], underscoring the importance and potential of such collective efforts in the field.

While both meta-analysis and mega-analysis are valuable techniques for synthesizing information from multiple studies, they each have their limitations. Meta-analysis relies on summary statistics from individual studies, which may not capture the full complexity and heterogeneity of the underlying data. Mega-analysis, in contrast, requires access

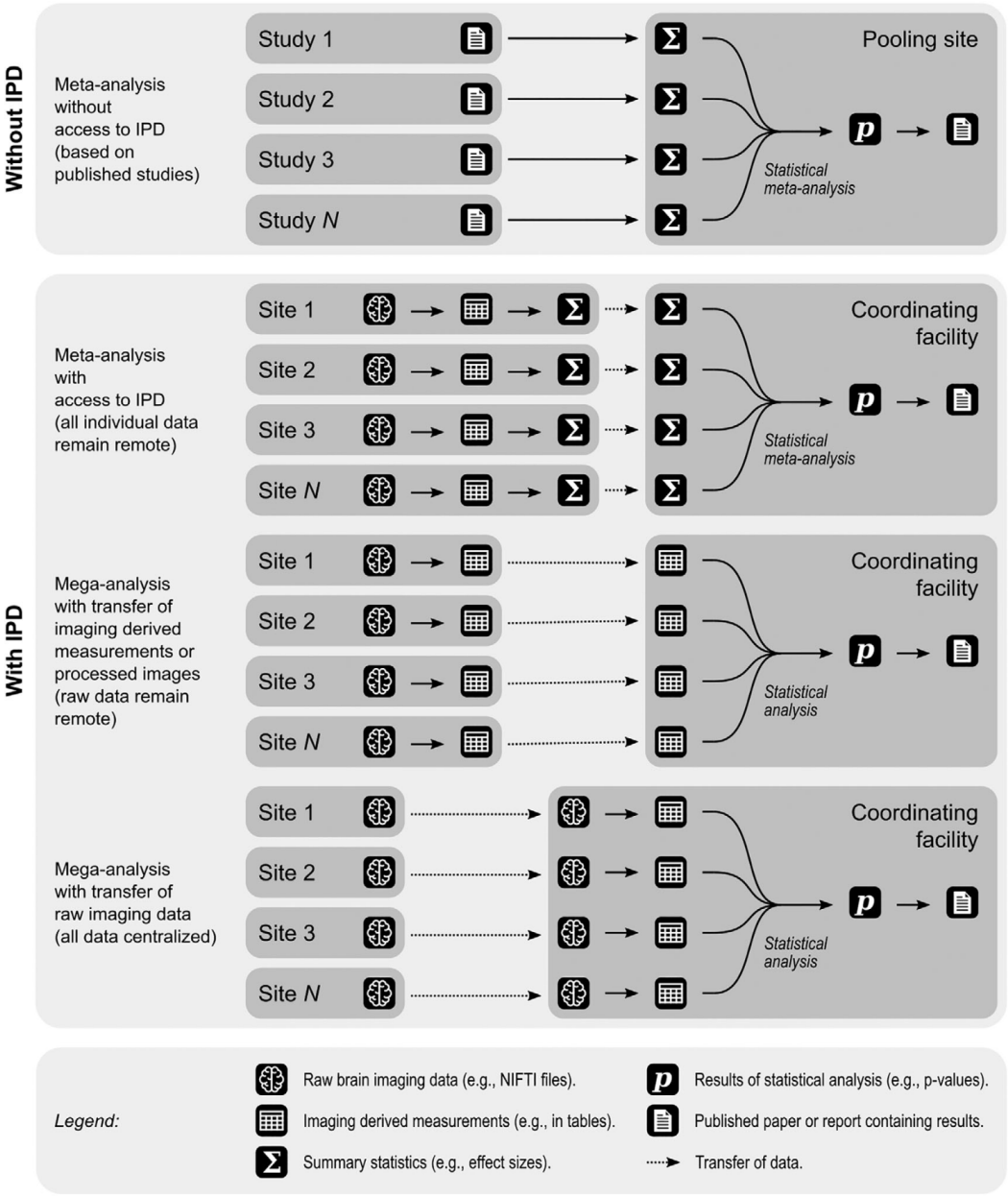


Fig. 1.1.: Comparison between classical literature-based meta-analyses conducted without access to individual participant data (IPD) (upper panel) and approaches used by different ENIGMA working groups with access to IPD (lower panel). The lower panel shows three main approaches: (top) data processed using common methods at each site, summary statistics computed and sent to a coordinating facility for meta-analysis; (middle) data processed using common methods at each site, sent to the coordinating facility for mega-analysis; and (bottom) raw data sent to the coordinating facility for batch processing and mega-analysis while accounting for site-specific effects. **Reprinted with permission from Andre Zugman, Mega-analysis methods in ENIGMA: The experience of the generalized anxiety disorder working group, 2022.**

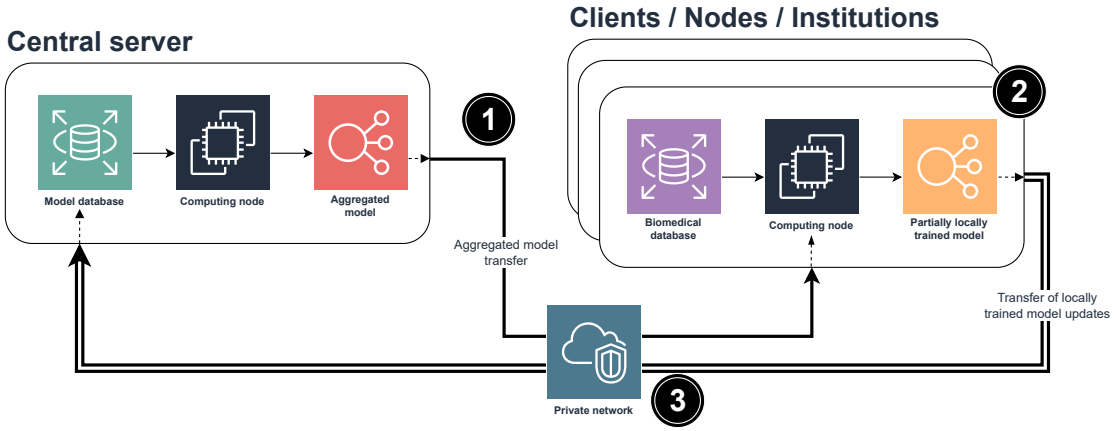


Fig. 1.2.: Illustration of the actors and the federated learning process in a collaborative setting. The diagram depicts two main actors: the central server and data owner institutions. It showcases the step-by-step process involved in federated learning: 1) The central server initiates the training by sending a common initialization to all participating institutions. 2) Each institution, serving as a data owner for a specific data source, performs local training using its own data without sharing or compromising data privacy. 3) The locally trained model (local updates) are securely transferred to the central server's model database. The central server's computing node then aggregates these updates to create a global model. 4) The aggregated model is sent back to the institutions, serving as a reinitialization point for the next round of training. This iterative process continues until convergence criteria are met or a predetermined communication or computing budget is reached. To ensure secure communication and data privacy, it is recommended to employ techniques such as a virtual private network (VPN) or cloud (VPC), which offer additional encryption layers.

to individual-level data, which can be challenging to obtain and share due to privacy concerns and data protection regulations, such as GDPR and CCPA [Voigt et al., 2017; Solove et al., 2020]. These challenges have directed the attention of researchers towards new paradigms that align with privacy regulations without compromising the robustness of the analysis. In this context, federated learning emerges as a viable alternative approach that meets these criteria.

1.4 Federated Learning

Federated learning (FL) enables privacy-aware collaborative model training on distributed data, while addressing governance concerns. It holds significant promise for multi-centric studies in healthcare, empowering researchers to conduct analyses across multiple institutions while prioritizing the privacy and confidentiality of sensitive data [Konečnỳ et al., 2016; McMahan et al., 2017].

Instead of centralizing data or querying it from all participants, Federated Learning (FL) performs decentralized optimization where only parameters of a model are allowed to

be shared without compromising the privacy of sensitive information. In FL, the model training process typically involves the following steps, as illustrated in Algorithm 1:

1. **Initialization:** A central server initializes the global model parameters $\boldsymbol{\theta}$ and shares them with all participating clients (e.g., healthcare institutions). This is represented in the algorithm as the initialization of $\boldsymbol{\theta}$.

2. **Local Training:** Each client k trains the model on their local dataset $\mathcal{D}_k$ using the current global model parameters $\boldsymbol{\theta}$, producing local model updates $\boldsymbol{\theta}_k$. This is performed in parallel across all clients, and each client performs E iterations of Stochastic Gradient Descent (SGD) on their local dataset, minimizing the local loss function L_k.

3. **Aggregation:** After local training, clients send their local model updates $\boldsymbol{\theta}_k$ to the central server. The server aggregates these updates to produce a new global model. The aggregation is a weighted average, where the weights are the sizes of the local datasets. This is represented in the algorithm as the update of $\boldsymbol{\theta}$.

4. **Global Update and Penalization:** The central server shares the updated global model parameters $\boldsymbol{\theta}$ with all clients, and the process repeats from step 2 until convergence. Each repetition of this cycle is considered as a round of communication. Additionally, the global model parameters $\boldsymbol{\theta}$ can be penalized based on some criteria (e.g., model complexity, divergence from prior round, etc.) to discourage overfitting and encourage stability.

This process, which is iterated for T communication rounds, allows the model to learn from all clients' data without the need to directly access or centralize the data, thus preserving privacy.

Two main actors are to be identified in a FL setup as shown in Figure 1.2, two main actors are to be identified in a FL setup: the *clients* and the *server*. Clients, also known as data owners or nodes, are individual entities, such as healthcare institutions, that possess their local datasets. These clients are connected to a network and participate in the federated learning process by training their local models and communicating with the central server. The central server, sometimes referred to as a third-party aggregator or coordinator, facilitates the federated learning process by initializing the global model, aggregating local model updates from clients, and sharing the updated global model back with the clients. The central server plays a crucial role in managing the communication and coordination among the clients, ensuring that the federated learning process converges to a global model that benefits from the collective knowledge of all participating clients.

Algorithm 1: Federated Learning

Data: Global model parameters $\boldsymbol{\theta}$, local datasets $\mathcal{D}_k$, learning rate η, number of local iterations E, number of communication rounds T
Result: Trained global model parameters $\boldsymbol{\theta}$
for $t \leftarrow 1$ **to** T **do**
  foreach *client* k *in* $\{1, 2, ..., K\}$ ***in parallel*** **do**
    $\boldsymbol{\theta}_k \leftarrow \boldsymbol{\theta}$ (initialize local model parameters)
    for $i \leftarrow 1$ **to** E **do**
      $\boldsymbol{\theta}_k \leftarrow \boldsymbol{\theta}_k - \eta \nabla L_k(\boldsymbol{\theta}_k) + \mu(\boldsymbol{\theta}_k)$ (perform local SGD on $\mathcal{D}_k$)
      Penalize $\boldsymbol{\theta}$ using μ based on some criteria (e.g., model complexity, divergence from prior round, etc.)
    end
  end
  $\boldsymbol{\theta} \leftarrow$ Aggregate$(\boldsymbol{\theta}_k)$ (aggregate local model parameters)
end

Mathematically, federated learning can be formulated as follows. As previously defined, $\mathcal{D}_k$ denote the local dataset of the k-th participant, and $f_k(\theta)$ denote the local objective function that measures the quality of the model parameters θ on the local dataset $\mathcal{D}_k$ (e.g. loss function). The global objective function is the average of local objective functions:

$$F(\theta) = \frac{1}{K} \sum_{k=1}^{K} f_k(\theta), \tag{1.2}$$

where K is the total number of participants. In FL, the goal is to minimize $F(\theta)$ by updating θ through a collaborative optimization process.

A commonly used approach for aggregation federated learning is Federated Averaging (FedAvg) [McMahan et al., 2017], which is a variant of the stochastic gradient descent (SGD) algorithm adapted for the federated setting. In each round of FedAvg, participants compute gradients on their local datasets and send the gradients or model updates to the central server. The server aggregates the updates and computes the global model as follows:

$$\theta^{(t+1)} = \sum_{k=1}^{K} \frac{|\mathcal{D}_k|}{|\mathcal{D}|} \Delta\theta_k^{(t)}, \tag{1.3}$$

where $\theta^{(t)}$ is the global model at round t, $\Delta\theta_k^{(t)}$ denote the gradients of the model from the k-th participant, and $\mathcal{D}$ is the union of all local datasets where $|\mathcal{D}_k|$ defines the number of observations for participant k.

Despite FedAvg's guarantees and practicality, there have been additional efforts to tackle some of the challenges discussed in Section 1.5, which also depend on the particular use case. Some of these aggregation schemes simply extend the advantages of existing centralized optimizers, such as FedAdam, FedYogi, and FedAdagrad [Reddi et al., 2020], which extend Adam [Kingma et al., 2014], Yogi [Zaheer et al., 2018], and Adagrad [Lydia et al., 2019], respectively. Other methods, like FedProx [Li et al., 2020a], focus on mitigating the effects of heterogeneous data on the generalizability of the models. Finally, some aggregation schemes aim to reduce communication rounds by scheduling learning rates under different strategies [Chang et al., 2018; Sheller et al., 2019].

1.5 Challenges in Federated Learning and impact on Applications to Healthcare

Despite the considerable potential of federated learning in advancing multi-centric studies in healthcare, FL also faces a unique set of challenges. These challenges stem from the need to maintain data privacy and security, the inherent heterogeneity of healthcare data, communication constraints, and the demands of model convergence and interpretability. Addressing these challenges is critical for the successful application of federated learning to healthcare and the development of reliable, generalizable, and privacy-preserving models.

1.5.1 Framework and Infrastructure

Developing a robust and scalable federated learning framework for healthcare applications is a significant challenge. The framework should accommodate a wide range of data types, models, and optimization methods while ensuring that data privacy and security are maintained. Additionally, the infrastructure should be adaptable to different computational resources, hardware configurations, and network conditions, allowing for seamless collaboration between institutions with varying technical capabilities [Brisimi et al., 2018; Chang et al., 2018].

1.5.2 Communication Efficiency and Resource Utilization

In federated learning, communication between the central server and participating clients (e.g., healthcare institutions) is crucial for model updates and coordination. The communication efficiency directly impacts the overall performance of the federated learning process, particularly in settings with limited bandwidth or unreliable network connections. It is essential to develop communication-efficient federated learning algorithms

and techniques that minimize the amount of transmitted information while maintaining model accuracy [Konečnỳ et al., 2016; Li et al., 2019c].

Resource utilization is another critical aspect of federated learning. The computational resources of participating clients can be heterogeneous, ranging from high-performance computing clusters to resource-constrained devices. Efficient resource utilization strategies should be devised to balance the workload and ensure that clients can participate in the federated learning process without overburdening their computational resources [Smith et al., 2017; Lim et al., 2020].

1.5.3 Heterogeneity

The inherent heterogeneity among participating institutions in federated learning can result in biased model updates and suboptimal learning performance, limiting the accuracy and generalizability of the resulting models. Such heterogeneity can come from different data distributions, demographic compositions, and data collection protocols. To address this issue, two main approaches can be taken: developing robust models that can generalize well despite the heterogeneity [Li et al., 2020a; Zhao et al., 2018], or, as is typically suggested in medical studies, correcting biases via harmonization methods [Orlhac et al., 2022].

Harmonization techniques aim to account for variations in the data that can result from differences in data collection, preprocessing, and other factors that could introduce biases. By adjusting the data to minimize these biases, harmonization methods can improve the performance of models trained on multi-center data. This approach is particularly relevant in medical imaging studies, where differences in imaging protocols and scanner hardware can lead to significant variations in the data.

Despite the widespread use of harmonization techniques in medical studies, their adoption in federated learning has been relatively limited. This is because the main objective of harmonization is to remove biases that could lead to misleading findings while preserving the effect of interest. In contrast, machine learning approaches may prioritize better classification or regression performance on unseen data which could be one reason why harmonization techniques are not as extensively utilized in federated learning compared to traditional medical studies. Currently, the proposals for harmonization remain valid mostly for multi-centric studies centralizing the data. These proposals range from the widely adopted ComBat model in multi-centric medical studies [Johnson et al., 2007a], to more complex solutions such as nonlinear versions of it such as ComBat-GAM [Pomponio et al., 2020a]. In addition, differently driven approaches like domain adaptation integration using generative adversarial networks (GANs) have also been proposed, as suggested by Wachinger et al. [Wachinger et al., 2021]. Incorporating harmonization

techniques in federated learning is a promising area of research that could improve the performance and generalizability of models trained on heterogeneous data.

1.5.4 Privacy and Security

While federated learning is inherently designed to preserve privacy, ensuring the confidentiality and security of sensitive patient data remains a challenge. Potential risks include model inversion attacks, membership inference attacks, and adversarial machine learning attacks. Thus, it is crucial to integrate advanced privacy-preserving mechanisms, such as differential privacy and secure multi-party computation, into the federated learning framework to protect against potential threats [Geyer et al., 2017; Bonawitz et al., 2017].

1.5.5 Model Convergence and Stability

The distributed nature of federated learning can make model convergence and stability more challenging compared to traditional centralized machine learning approaches. Communication delays, resource constraints, and data heterogeneity can lead to slow or inconsistent model convergence. Developing techniques to monitor and improve convergence rates, as well as ensuring model stability under varying conditions, is essential for the practical application of federated learning in healthcare [Li et al., 2018; Sattler et al., 2019].

1.6 Aims and Structure of the Thesis

The primary aim of this thesis is to address the challenges and explore the potential of federated learning in healthcare. We investigated the development of a robust federated learning framework, optimize resources required for effective federated learning, and improve data harmonization techniques within a federated learning framework, particularly in the context of medical imaging. This work may potentially enable researchers to access and analyze previously inaccessible data in a more robust and generalizable manner while adhering to privacy and ethical standards.

The thesis is structured as follows:

- Chapter 2 introduces Fed-ComBat, a federated approach for harmonization on decentralized data that preserves nonlinear covariate effects.

- In Chapter 3, we present another effort in bias modeling based on our proposed federated mixed-effects variational autoencoders.

- Chapter 4 focuses on optimizing the utilization of communication and computation resources in federated learning through black-box optimization.

- Chapter 5 highlights the need for a common framework for federated learning in healthcare and proposes: Fed-BioMed, an open-source infrastructure for federated learning that fulfilling such need.

By addressing these key aspects, this thesis will contribute to the ongoing development and adoption of federated learning in healthcare, enabling more accurate and personalized medicine.

1.7 Publications

As a result of the previously described work, following publications have been achieved:

- **(Under review)** Silva, Santiago, et al. "Fed-ComBat: A Generalized Federated Framework for Batch Effect Harmonization in Collaborative Studies." Humman Brain Mapping, 2023.

- Silva, Santiago, et al. "Fed-BioMed: A general open-source frontend framework for federated learning in healthcare." Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning: Second MICCAI Workshop, DART 2020, and First MICCAI Workshop, DCL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 2. Springer International Publishing, 2020.

1.7.1 Collaborations

During the conception and course of this thesis, various collaborations were pursued, resulting in the following publications:

- Terrail, J. O. D., Ayed, S. S., Cyffers, E., Grimberg, F., He, C., Loeb, R., ... & Andreux, M. (2022). FLamby: Datasets and Benchmarks for Cross-Silo Federated Learning in Realistic Healthcare Settings. NeurIPS, 2022.
  CRediT Statement: Data Curation, Formal Analysis, Software, Writing - Original Draft.

- Balelli, I., Silva, S., Lorenzi, M., & Alzheimer's Disease Neuroimaging Initiative. (2021). A probabilistic framework for modeling the variability across federated datasets. In Information Processing in Medical Imaging: 27th International Conference, IPMI 2021, Virtual Event, June 28–June 30, 2021, Proceedings 27 (pp. 701-714). Springer International Publishing.
  CRediT Statement: Data curation, Software, Writing - Original Draft.

- S. Silva, B. A. Gutman, E. Romero, P. M. Thompson, A. Altmann and M. Lorenzi, "Federated Learning in Distributed Medical Databases: Meta-Analysis of Large-Scale Subcortical Brain Data," 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), Venice, Italy, 2019, pp. 270-274.
  CRediT Statement: Conceptualization, Software, Validation, Formal Analysis, Investigation, Writing - Original Draft, Visualization.

2

Fed-ComBat: Secure Data Harmonization via Federated Learning

Abstract: In neuroimaging research, the use of multi-centric analyses is crucial for obtaining sufficient sample sizes and representative clinical populations. Data harmonization techniques are typically employed in the pipeline of multi-centric studies to address systematic biases and ensure the interoperability of the data. However, most multicentric studies require data centralization at some point during the analysis pipeline, thus presenting the risk to expose individual patient information. This poses a significant challenge in data governance. To mitigate the risk of exposing patient information, various privacy regulations such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA), have been introduced. While these regulations address such governance concerns, they also hinder data access for researchers. Federated learning (FL) offers a privacy-preserving alternative approach in machine learning, enabling models to be collaboratively

trained on decentralized data without the need for data centralization or sharing. FL provides a solution to the problem of data centralization. However, for FL models to effectively work with decentralized data, it is crucial to ensure data harmonization.

In this paper, we present Fed-ComBat, a federated framework for batch effect harmonization on decentralized data. Fed-ComBat extends existing centralized linear methods, such as ComBat and distributed as d-ComBat, and nonlinear approaches like ComBat-GAM in accounting for potentially nonlinear and multivariate covariate effects. By doing so, Fed-ComBat enables the preservation of nonlinear covariate effects without requiring centralization of data and without prior knowledge of which variables should be considered nonlinear or their interactions, differentiating it from ComBat-GAM. We assessed Fed-ComBat and existing approaches on simulated data and multiple cohorts comprising healthy controls (CN) and subjects with various disorders such as Parkinson's disease (PD), Alzheimer's disease (AD), and autism spectrum disorder (ASD).

Results indicate that our nonlinear version of Fed-ComBat outperforms centralized ComBat in the presence of nonlinear effects and is comparable to centralized methods such as ComBat-GAM. Using synthetic data, Fed-ComBat is able to better reconstruct the target unbiased function by 35% (RMSE = 0.5952) with respect to d-ComBat (RMSE = 0.9162) and 12% with respect to our proposed federated d-ComBat-GAM (RMSE= 0.6751) and exhibits comparable results on MRI-derived phenotypes to centralized methods as ComBat-GAM without the need of prior knowledge on potential nonlinearities.

2.1 Introduction

Neuroimaging studies, especially those incorporating machine learning (ML) techniques, greatly benefit from large sample sizes. However, the current approach of centralizing data for analysis faces significant challenges due to data protection regulations such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) [European Commission, 2016; Bukaty, 2019; Calzada, 2022]. Despite efforts in anonymization, such as defacing, it is important to note that these techniques may introduce alterations in certain image-derived phenotypes, and some report re-identification rates of around 30%. [Schwarz et al., 2021]. This underscores the need for a reevaluation of the traditional paradigms for multicentric clinical studies, which heavily rely on data centralization.

Data harmonization is known to be a crucial factor in tackling certain challenges of multicentric data analysis due to systematic errors or *batch effects* commonly present

in such studies [Wachinger et al., 2021]. Harmonization aims to address biases and variations among data collected from different sources (e.g., scanner brands or institutions), ensuring that the data is comparable and can be combined effectively for analysis. [Johnson et al., 2007a] proposed ComBat in the field of gene expression analysis, as a way to correct these batch effects while preserving the desired covariate effects (e.g., sex, diagnosis, age). Later, this method was adapted to brain magnetic resonance imaging (MRI) phenotypes [Fortin et al., 2017; Fortin et al., 2018]. However, the typical use of ComBat in multicentric studies still requires data to be centralized, thus limiting the application to real-life collaborative analysis scenarios. To overcome this problem, [Chen et al., 2022] proposed d-ComBat to harmonize data within a distributed setup by extending the optimization of ComBat to allow bias correction without the need to exchange the data across centers. To achieve this, d-ComBat shares full local covariance matrices computed at each site, and the subsequently aggregates these covariance matrices to estimate the effect parameters at a global level. This method was shown to achieve almost an exact solution with respect to its centralized version requiring only a few rounds of communication. However, sharing full covariance matrices represents a potential risk of sensitive data leakage [Kesteren et al., 2019], and like ComBat, d-ComBat considers that covariates have a linear influence on phenotypes, which may not always be the case as reported by [Bethlehem et al., 2022] where age shows a clear nonlinear influence on MRI-derived phenotypes. To handle nonlinear covariate effects, [Pomponio et al., 2020a] introduced ComBat-GAM, which relies on generalized additive models (GAMs) for the preservation of such nonlinear effects. By using smoothing functions, such as polynomial or splines, to decompose covariate effects into basis functions, ComBat-GAM approximates the aforementioned covariate influences. Although ComBat-GAM showed better performance than ComBat in the LIFESPAN dataset [Pomponio et al., 2020a], it remains limited by design as it has been conceived only for centralized settings. Moreover, ComBat-GAM requires an explicit definition of covariate interactions, typically limited to additive or multiplicative terms (e.g., $\mathrm{Sex} \times \mathrm{Diagnosis}$). Despite efforts, a gap still exists in the availability of methods for nonlinear covariate effect preservation without requiring data centralization. Furthermore, as highlighted by [Gebre et al., 2023] in a study conducted at the Mayo Clinic, the harmonization problem remains unsolved, calling for further research and collaborative efforts in the field to address this challenge.

Federated learning (FL) is a machine learning paradigm that allows for collaborative model optimization while maintaining data privacy and governance [Konečnỳ et al., 2016]. In FL, models are trained on data that remains decentralized across multiple institutions or devices, ensuring that sensitive data stays securely within the respective entities [Konečnỳ et al., 2016]. FL operates by sharing only model updates or parameters, aggregating these updates on a central server, and distributing the updated global model back to each institution. This iterative process continues until convergence is achieved. Federated Averaging (FedAvg) is a popular approach for aggregation in federated learning. It adapts the stochastic gradient descent (SGD) optimization algorithm for the

federated setting providing similar convergence guarantees [McMahan et al., 2017]. FL facilitates access to larger cohorts and it has successfully been applied in neuroimaging for tumor segmentation and diagnosis [Li et al., 2019b; Mahlool et al., 2022], disease studies in functional magnetic resonance imaging (fMRI) [Li et al., 2020b], intracranial hemorrhage (ICH) detection[Cheung et al., 2023] and continues to be increasingly adopted in neuroimaging applications.

An opportunity opens for the development of flexible and effective harmonization approaches that can be easily integrated into FL pipelines, enabling harmonization of neuroimaging data across sites while accounting for complex covariate effects on heterogeneously distributed data. And with this approaches, to have the potential to fill the current gap and allow for more comprehensive analysis of distributed cohorts using FL. We contribute to tackling these challenges by first introducing a generalized formulation of ComBat that extends to multivariate and nonlinear covariate effect modeling by allowing the incorporation of more complex functions. This enables a more nuanced representation of covariate effects during the harmonization process while encompassing ComBat and ComBat-GAM as specific cases. Secondly, we propose Fed-ComBat, a generalized federated ComBat framework designed specifically for data batch effect harmonization in federated settings. This framework enables the harmonization of data while preserving privacy and security within distributed environments, aligning with the principles of federated learning.

We benchmarked and compared our proposed methods with existing centralized approaches (ComBat, and ComBat-GAM) and distributed (d-ComBat) [Johnson et al., 2007a; Pomponio et al., 2020a; Chen et al., 2022]. Moreover, for the sake of comparison we also formulated and implemented a distributed version of ComBat-GAM (d-ComBat-GAM), which relies on distributed covariance estimation, and extend the method of [Chen et al., 2022] by allowing basis decomposition to interpolate nonlinear covariate effects using smoothing functions.

The different methods were compared for their ability to harmonize batch effects and preserve the quality of covariate effects on simulated data (Section 2.4). Furthermore, we performed an evaluation on derived phenotypes from MRI-brain images from nine cohorts: A4 [Sperling et al., 2014], ABIDE-I [Di Martino et al., 2014], ABIDE-II [Di Martino et al., 2017], ADNI [Weiner et al., 2013], AIBL [Ellis et al., 2009], MIRIAD [Malone et al., 2013], OASIS3 [LaMontagne et al., 2019], PDPB [Rosenthal et al., 2016], PPMI [Marek et al., 2018]. Comprising controls and people with different brain disorders: patients with different subtypes of Parkinson's disease (PD), Alzheimer's disease (AD) and Autism spectrum disorder (ASD).

Using synthetic data, we find that Fed-ComBat achieves an RMSE of 0.9162(i.e., 35%) improvement in reconstructing the target unbiased function compared to d-ComBat,

and a 12% improvement compared to our federated extension of ComBat-GAM, d-ComBat-GAM (RMSE = 0.6751). In addition, our application on real data shows that Fed-ComBat on MRI-derived phenotypes is comparable to centralized ComBat-GAM. This implies that the studied cohorts may not have exhibited complex nonlinear effects or interactions that could not be captured by a GAM. However, Fed-ComBat requires no prior assumptions of nonlinearities unlike ComBat-GAM and automatically captures such effects and interactions through the model. Moreover, Fed-ComBat can be considered a better approach towards privacy preserving methods as it only shares model parameters, while d-ComBat and our included extension d-ComBat-GAM share full covariance and cross-covariance matrices [Kesteren et al., 2019].

2.2 Methods

2.2.1 Generalized ComBat model

Following the original ComBat formulation proposed by [Johnson et al., 2007a], let us denote a batch (represented by different scanners protocols, machines, or institutions) indexed by $i \in \{1, 2, ..., S\}$ on a particular phenotype (e.g., a brain region) indexed by $g \in \{1, 2, ..., G\}$. Each batch contains n_i number of observations, and the total number of observations is $N = \sum_i^S n_i$. S can denote for simplicity the number of sites in the study, but it can also be extended to the total number of scanners between sites or any other number of batch effects. We can model a specific phenotype g observed in the j-th patient who belongs to the i-th site denoted by y_{ijg} as follows:

$$y_{ijg} = \alpha_g + \phi(\mathbf{x}_{ij}, \boldsymbol{\theta}_g) + \gamma_{ig} + \delta_{ig}\varepsilon_{ijg}, \tag{2.1}$$

where $\mathbf{x}_{ij}$ denotes the covariate effects expected to be preserved after removing the batch effects (e.g., sex and age), α_g acts as a global fixed intercept (i.e., the mean), while γ_{ig} indicates a random intercept that accounts for the site-specific shift. ε_{ijg} is a noise model that captures the variability of each phenotype $\varepsilon_{ijg} \sim \mathcal{N}(0, \sigma_g^2)$, and δ_{ig} is a multiplicative effect that scales the "unbiased" phenotype variability to fit the one at each site.

This formulation generalizes the original linear model proposed by [Johnson et al., 2007a] and the nonlinear covariate effect model proposed by [Pomponio et al., 2020a]. While the original model considered $\phi(\mathbf{x}; \boldsymbol{\theta}_g)$ to be a linear function of $\mathbf{x}$, and the nonlinear model considered $\phi(\mathbf{x}; \boldsymbol{\theta}_g)$ to be an univariate spline, this formulation treats $\phi(\mathbf{x}; \boldsymbol{\theta}_g)$ as a general function that can potentially be multivariable and nonlinear and parametrized by $\boldsymbol{\theta}_g$. This generalization allows for the use of more complex approximating functions such as kernel-based models or neural networks as the multilayer perceptron (MLP),

which is known to be an universal function approximator. However, it is essential to use cross-validation techniques to avoid overfitting and ensure the model generalizes well to unseen data.

Some constraints are defined during the fixed effect parameter estimation ($\hat{\alpha}_g, \hat{\boldsymbol{\theta}}_g, \hat{\gamma}_{ig}, \hat{\sigma}_g$):

$$\arg \max_{\hat{\alpha}_g, \hat{\boldsymbol{\theta}}_g, \hat{\gamma}_{ig}, \hat{\sigma}_g} P(y_{ijg} | \hat{\alpha}_g, \hat{\boldsymbol{\theta}}_g, \hat{\gamma}_{ig}, \hat{\sigma}_g) \tag{2.2}$$

$$\text{subject to} \quad \mathbb{E}_g[\hat{\gamma}_i] = \sum_i^S \frac{n_i}{N} \hat{\gamma}_{ig} = 0, \, \forall g \in \{1, ..., G\} \tag{2.3}$$

$$\text{and} \quad \phi(\mathbf{x}, \boldsymbol{\theta}_g)|_{x=0} = 0 \tag{2.4}$$

A first constraint in Equation (2.3) is set to allow identifiability of the intercept parameters as explained by [Johnson et al., 2007a] (explained in practice in Appendix A.4). However, in the same regard, the constraint in Equation (2.4) ensures that no batch effects are captured by the covariate function $\phi(\cdot)$. This second constraint, despite not being mentioned, is also fulfilled by [Johnson et al., 2007a] and [Pomponio et al., 2020a]; making ComBat and ComBat-GAM a particular case of the proposed formulation in this work.

For a centralized setup, the estimation of all these parameters is performed in three steps: i) maximum likelihood estimation (MLE) for parameters $\hat{\alpha}_g, \hat{\boldsymbol{\theta}}_g, \hat{\gamma}_{ig}$ (see Equation (2.2)) and of the phenotype variance $\hat{\sigma}_g^2 = \frac{1}{N} \sum_{ij} \left(y_{ijg} - \hat{\alpha}_g - \phi(\mathbf{x}_{ij}; \hat{\boldsymbol{\theta}}_g) - \hat{\gamma}_{ig} \right)^2$, ii) residual standardization mapping the residuals to satisfy the form $y_{ijg} \rightarrow z_{ijg} \sim \mathcal{N}(\gamma_{ig}, \delta_{ig}^2)$ as follows:

$$z_{ijg} = \frac{y_{ijg} - \hat{\alpha}_g - \phi(\mathbf{x}_{ij}; \hat{\boldsymbol{\theta}}_g)}{\hat{\sigma}_g} \tag{2.5}$$

and iii) estimation of the additive and multiplicative batch effects $\hat{\gamma}_{ig}^*$ and $\hat{\delta}_{ig}^*$ as in Equation (2.6), using empirical Bayes (EB) with priors on γ_{ig} and δ_{ig}^2 to iteratively estimate these parameters as in Equation (2.7) [Johnson et al., 2007a, Sec. 3.2].

$$\gamma_{ig} \sim \mathcal{N}(Y_i, \tau_i^2) \quad \text{and} \quad \delta_{ig}^2 \sim \text{Inverse Gamma}(\lambda_i, \vartheta_i) \tag{2.6}$$

$$\gamma_{ig}^* = \frac{n_i \bar{\tau}_i^2 \hat{\gamma}_{ig} + \delta_{ig}^{2*} \bar{\gamma}_{ig}}{n_i \bar{\tau}_i^2 + \delta_{ig}^{2*}}, \; \delta_{ig}^* = \frac{\bar{\theta}_i + \frac{1}{2} \sum_{j=1}^{n_i} (z_{ijg} - \gamma_{ig}^*)^2}{\frac{n_i}{2} \bar{\lambda}_i - 1} \tag{2.7}$$

Lastly, phenotypes can be harmonized while preserving the covariate effects of interest as follows:

$$y_{ijg}^{\text{ComBat}} = \frac{\hat{\sigma}_g}{\hat{\delta}_{ig}^*}(z_{ijg} - \hat{\gamma}_{ig}^*) + \hat{\alpha}_g + \phi(\mathbf{x}_{ij}; \hat{\boldsymbol{\theta}}_g) \tag{2.8}$$

In the following section, we will discuss how this formulation facilitates the incorporation of harmonization within the federated learning framework.

2.2.2 Federated ComBat: Fed-ComBat

Considering the formulation previously presented in Equation (2.1), the parameters $\alpha_g, \boldsymbol{\theta}_g$ and γ_{ig} can be optimized by minimizing an objective function F. As data is now siloed, it is only possible to have an evaluation of the cost function at each site F_i, thus defining the federated optimization problem as:

$$\arg \min_{\alpha_g, \boldsymbol{\theta}_g, \gamma_{ig}} F(\alpha_g, \boldsymbol{\theta}_g, \gamma_{ig}), \tag{2.9}$$

$$\text{where, } F(\alpha_g, \boldsymbol{\theta}_g, \gamma_{ig}) := \sum_{i=1}^{S} \frac{n_i}{N} F_i(\alpha_g, \boldsymbol{\theta}_g, \gamma_{ig}) \tag{2.10}$$

We rely on *Federated averaging* (FedAvg) to tackle this optimization problem [McMahan et al., 2017] by allowing each site to conduct partial optimization using stochastic gradient descent (SGD) locally on their data, followed by an aggregation step that combines the shared parameters of each site. The convergence of the resulting iterative process has been demonstrated for both IID and non-IID data distribution across clients [Li et al., 2019c].

The Fed-ComBat framework requires the following steps for harmonization: i) a federated standardization step to avoid scaling issues in gradient descent as proposed by [Silva et al., 2019], ii) a federated estimation of $\alpha_g, \boldsymbol{\theta}_g$ and γ_{ig} as presented in Equation (2.10), and iii) a local estimation of the random effects following using EB as in Equations (2.6) and (2.7). A description of the steps followed in federated harmonization using Fed-ComBat is described in Algorithm 2 including standardization as preprocessing, local updates relying on SGD, global updates across sites of shared parameters using FedAvg and random effect estimation using Empirical Bayes.

2.2.3 Related works

d-ComBat

[Chen et al., 2022] proposed a distributed version of ComBat that allows correction

Algorithm 2: Fed-ComBat

Data: Non-harmonized phenotypes (y_{ijg}) and covariates ($\mathbf{x}_{ij}$).
Result: Harmonized phenotypes with siloed data.
$\mathbf{x} \leftarrow$ FEDERATEDSTANDARDIZATION($\mathbf{x}$);

```
// Estimation of fixed effects and random intercept
```
initialization of parameter space $\mathbf{\Omega} := \{\boldsymbol{\theta}_g, \alpha_g, \gamma_{ig}\}_{g=1}^{G}$;
while *not converged* **do**
  foreach *site* i **do**
    $\mathbf{\Omega}_i^t \leftarrow \mathbf{\Omega}^{(t+1)}$; `// Initialization.`
    `// Partial local optimization using SGD.`
    foreach *local gradient step* t **do**
      $\mathbf{\Omega}_i^{(t+1)} = \mathbf{\Omega}_i^{(t)} - \eta \nabla_{\mathbf{\Omega_i}} F(\mathbf{\Omega_i}^{(t)})$;
  `// Aggregate and update every parameter using FedAvg.`
  $\mathbf{\Omega}^{(t+1)} \leftarrow \sum_i \frac{n_i}{N} \mathbf{\Omega}_i^t$
forall *site* i **do**
  `// Standardize using estimated parameters`
  $z_{ijg} \leftarrow \frac{y_{ijg} - \hat{\alpha}_g - \phi(\mathbf{x}_{ij}; \hat{\boldsymbol{\theta}}_g)}{\hat{\sigma}_g}$
  `// Estimate` γ_{ig}^* `and` δ_{ig}^* `using EB`
  $\gamma_{ig}^*, \delta_{ig}^* \leftarrow$ EMPIRICALBAYES($\mathbf{x}$) ; `// Equations (2.6) and (2.7)`
  `// Correct data`
  return $y_{ijg}^{\text{ComBat}} \leftarrow \frac{\hat{\sigma}_g}{\hat{\delta}_{ig}^*}(z_{ijg} - \hat{\gamma}_{ig}^*) + \hat{\alpha}_g + \phi(\mathbf{x}_{ij}; \hat{\boldsymbol{\theta}}_g)$

for batch effects in a distributed manner. This is achieved by taking advantage of the multivariate linear regression problem solved in ComBat, requiring to compute the product between the inverse of a covariance matrix and the data-targets cross-covariance. Both can be computed by aggregating the individual matrices provided by each site. This allows us to reformulate Equation (2.1) as done by [Chen et al., 2022]:

$$\underbrace{\begin{bmatrix} y_{11g} \\ \vdots \\ y_{ijg} \\ \vdots \\ y_{Sn_sg} \end{bmatrix}}_{\mathbf{y}_g} = \underbrace{\left[\begin{array}{ccc|ccc|c} x_{111} & \dots & x_{11C} & 1 & \cdots & 0 & 1 \\ \vdots & x_{ijc} & \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{Sn_S1} & \dots & x_{Sn_sC} & 0 & \dots & 1 & 1 \end{array}\right]}_{\mathbf{X}} \underbrace{\left[\begin{array}{c} \hat{\theta}_{g1} \\ \vdots \\ \hat{\theta}_{gc} \\ \hline \hat{\gamma}_{1g} \\ \vdots \\ \hline \hat{\alpha}_g \end{array}\right]^{\intercal}}_{\hat{\mathbf{\Theta}}_g} + \boldsymbol{\delta}_g \boldsymbol{\varepsilon}_g \tag{2.11}$$

Starting from the panel formulation where $\mathbf{y_g}$ is the stacked vector of all subjects across all sites i, and $\mathbf{X}$ is a design matrix containing the covariate effects to be preserved

indexed by c in subject j from site i (x_{ijc}), an indicator matrix encoding the site, and a column of ones to capture α_g, we can estimate the augmented parameter matrix $\hat{\mathbf{\Theta}}_g$ using maximum likelihood. The estimator can be decomposed as a sum of the covariance and the cross-covariance matrices at each site, as shown in Equation 2.12. This formulation has also been used to distribute latent variable models, such as PCA and PLS, among multiple sites [Silva et al., 2019; Lorenzi et al., 2018]. The authors have also suggested sharing a partial-eigen decomposition to mitigate the risk of data leakage, as shown in Equation 2.13 where the least squares problem is decomposed by $\mathbf{X}^\intercal\mathbf{X}$ which denotes the covariance matrix, $\mathbf{X}^\intercal\mathbf{y}_g$ the cross-covariance matrix for phenotype g, $\mathbf{Q}_i, \mathbf{\Lambda}_i$ the eigen decomposition for the covariance matrix in site i and, $\mathbf{U}_i\mathbf{\Sigma}_i\mathbf{V}_i$ the SVD decomposition for the cross-covariance matrix in site i.

$$\hat{\mathbf{\Theta}}_g = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}_g = \left(\sum_{i=1}^{S}\mathbf{X}_i^\intercal\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{S}\mathbf{X}_i^\intercal\mathbf{y}_{ig}\right) \tag{2.12}$$

$$= \left(\sum_{i=1}^{S}\mathbf{Q}_i\mathbf{\Lambda}_i\mathbf{Q}_i^{-1}\right)\left(\sum_{i=1}^{S}\mathbf{U}_i\mathbf{\Sigma}_i\mathbf{V}_i^\intercal\right) \tag{2.13}$$

d-ComBat-GAM

Seeking a federated nonlinear approach, we here propose d-ComBat-GAM as an adaptation for d-ComBat for non-linear covariate effect preservation via generalized additive models (GAM) as proposed in its centralized version by [Pomponio et al., 2020a]. This method, as d-ComBat, allows a fast estimation in a closed form by sharing either full or partial decomposition of the covariance and cross covariance matrices but transforming covariates in a design matrix using smooth functions such as polynomials or splines.

We can consider d-ComBat-GAM as a particular case of Fed-ComBat where $\phi(\cdot)$ becomes a linear combination parametrized by $\boldsymbol{\theta}$. Let $b(\cdot)$ be an arbitrary basis representation, the covariate function approximate is then defined as follows:

$$\phi(\mathbf{x}_{ij};\boldsymbol{\theta}_g) = \sum_{c=1}^{C}\sum_{k=1}^{K_c}\theta_{gk}b_{gk}(x_{ijc}) \quad \text{where} \quad \boldsymbol{\theta}_g = [\theta_{g1},\ldots,\theta_{gK_c}] \tag{2.14}$$

Note that Equation (2.14) expresses a linear combination of parameters, hence the same procedure as in Section 2.2.3 can be followed to estimate the fixed effects. Although polynomial decomposition is straightforward as it is the equivalent of transforming and appending the variable to a design matrix, splines require a set of defined control points based on the covariate range. We rely on federated standardization to map the nonlinear covariate to a normalized distribution where the range is known and thus the control points of the splines can be predefined.

2.3 Materials

We evaluated Fed-ComBat on synthetic data accounting for different sources of batch- and covariate-wise heterogeneity. In addition we benchmarked this approach on a collection of nine cohorts corresponding to different studies in neurodegenerative disorders including participants diagnosed with or at risk of developing Parkinson's and Alzheimer's disease, Autism spectrum disorder as well as cognitively normal.

2.3.1 Synthetic data

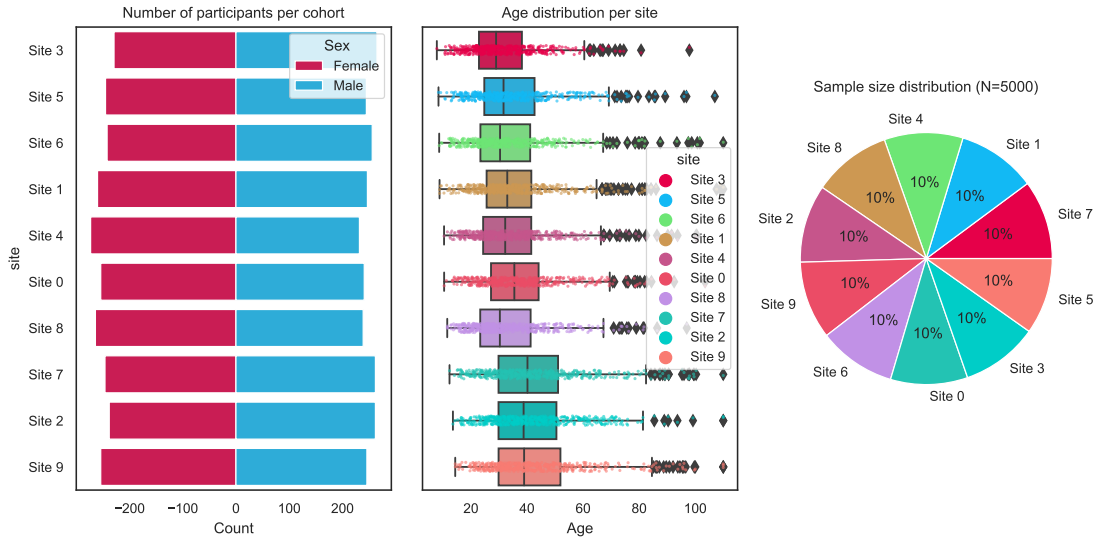
We created simulated datasets under three main scenarios: i) when cohorts have similar distributions for covariates and sample size effects (IID), ii) when covariate distributions slightly overlap but follow different patterns, and when there are differences in sample sizes (mid-non-IID), and iii) when there is no overlap between covariate distributions between cohorts and there are evident differences in sample sizes (non-IID). An example of IID and non-IID covariate distributions on synthetic data is shown in Figure 2.1. We simulated ComBat parameters following the methodology outlined in [Reynolds et al., 2022], using the graphical model depicted in Figure 2.2. The primary difference between our approach and that of [Reynolds et al., 2022] was the model used to relate the covariates and phenotypes ϕ. We used a nonlinear function in this step, which was employed as the unbiased target function to be learned after the harmonization process for the synthetically biased phenotypes.

Sex was generated by sampling from a $\mathrm{Bernoulli}(p_i^{(\mathrm{sex})})$ distribution, which modulated sex proportion distribution across sites. Age was drawn from a $\mathcal{N}(\mu_i^{(\mathrm{age})}, {\sigma_i^{(\mathrm{age})}}^2)$ distribution, where $\mu_i^{(\mathrm{age})}$ and $\sigma_i^{(\mathrm{age})}$ were sampled from a $\mathrm{Uniform}(a, b)$ distribution, with heterogeneity modulated by the width of the distribution $(b - a)$. Sample size heterogeneity was simulated using a concentration model $p_i \sim \mathrm{Dir}(\boldsymbol{\alpha})$, as previously suggested for measuring the effects of heterogeneity in Federated Learning by [Hsu et al., 2019].

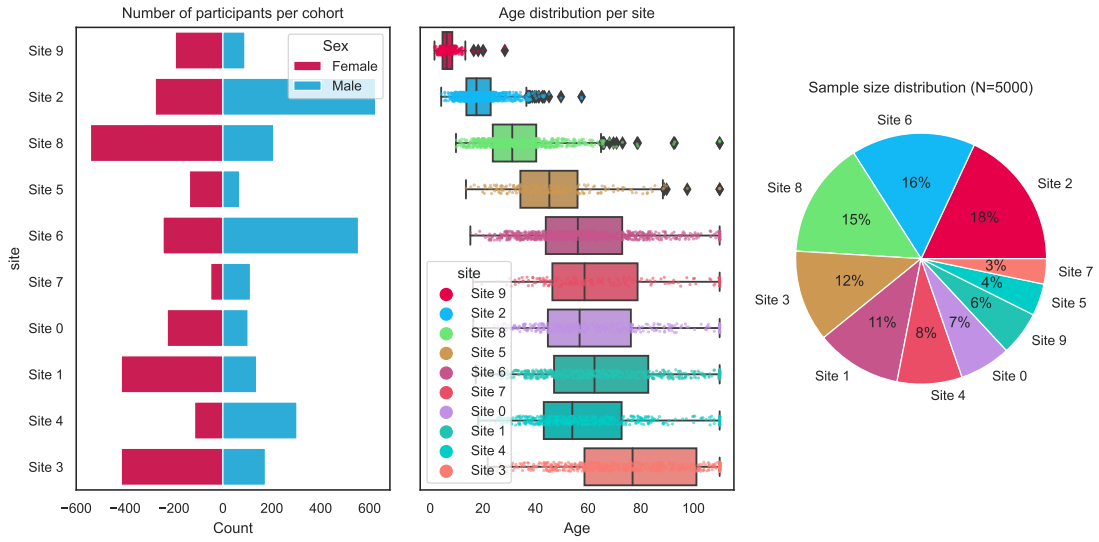
As shown in Figure 2.3 and explained by multiple authors in the review by [Zhu et al., 2021], non-IID scenarios consistently pose the greatest challenges for federated approaches. Therefore, we chose to focus on non-IID setups to provide a more realistic and comprehensive evaluation.

2.3.2 Brain MRI-Data

We evaluated our method on a cross-sectional cohort of 7265 participants from nine public studies, utilizing MRI-derived phenotypes from baseline structural T1-weighted magnetic resonance imaging (MRI).



(a)



(b)

Fig. 2.1.: Population distribution of synthetic data across cohorts in two different scenarios. The left panel shows the sex distribution, the middle panel shows the age distribution, and the right panel shows the sample size distribution. Panel 2.1a shows a homogeneous and IID population, while panel 2.1b shows a non-IID distribution of covariates, which is more commonly observed and used in this work.

Participants were categorized by diagnosis as follows: 3992 cognitively normal (CN), of which 731 of them were considered at risk of cognitive impairment from the A4 study [Sperling et al., 2014]; 1065 were identified as having mild cognitive impairment (MCI) from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [Weiner et al., 2013] and the Australian Imaging, Biomarker and Lifestyle Flagship Study of Aging (AIBL) [Ellis et al., 2009]; 998 had been diagnosed with Autism Spectrum Disorder (ASD) from ABIDE-I and ABIDE II [Di Martino et al., 2014]; and 538 were patients diagnosed with Alzheimer's disease (AD) from ADNI and Open Access Series of Imaging Studies (OASIS)

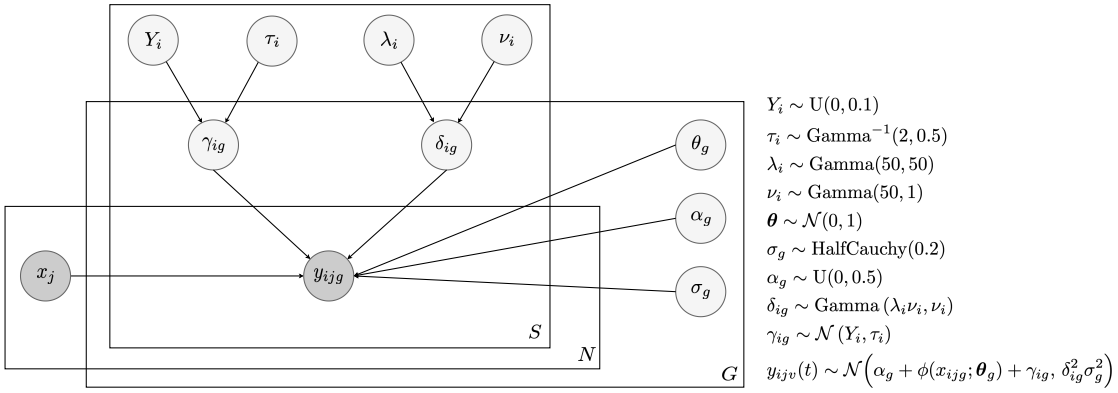


Fig. 2.2.: Graphical model used to generate synthetic data. The shaded circles indicate observed measurements, including covariates and imaging feature values, while unshaded circles represent latent parameters.

dataset [LaMontagne et al., 2019] and, 686 participants diagnosed in multiple phases of Parkinson's disease from the Parkinson's Disease Biomarkers Program (PDPB) [Rosenthal et al., 2016] and the Parkinson's Progression Markers Initiative (PPMI) [Marek et al., 2018]. More detailed demographic information can be found in Table 2.1. Covariate distribution across the multiple cohorts used in this work can be better observed in Figure 2.3, showing that when using multiple cohorts, it's very likely to have non-IID distributions.

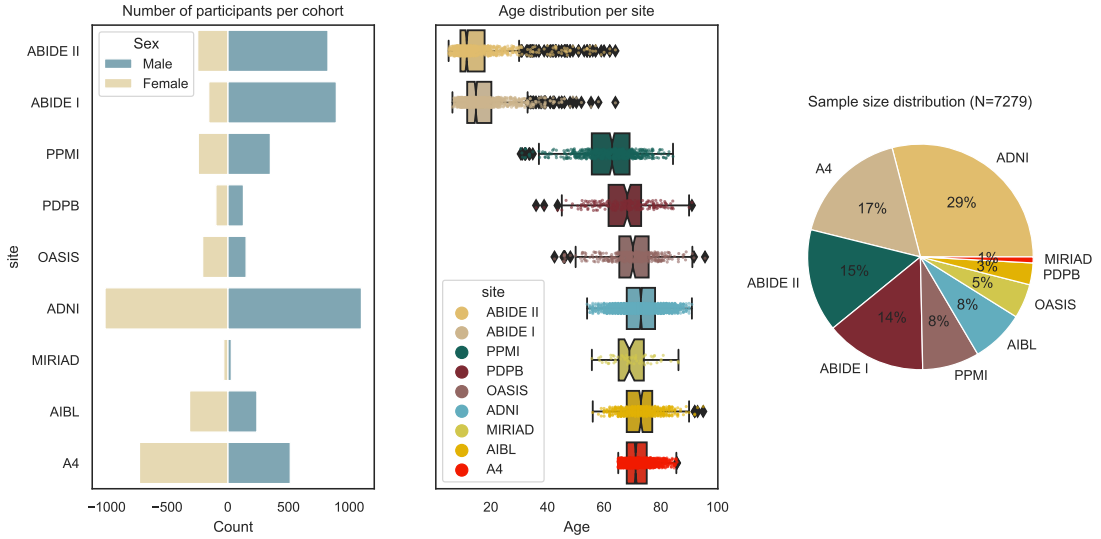


Fig. 2.3.: Population pyramid representing the demographics of the real data used in this study. The left panel shows the distribution of sex, the middle panel shows the distribution of age, and the right panel shows the sample size distribution. Cohorts were sorted by ascending median age. The demographics for the population are described in Table 2.1. The figure shows a clear non-IID distribution, similar to the distribution observed in Figure 2.1b.

Image processing

Cortical thickness and subcortical volume were extracted from MRI scans using FreeSurfer v7.1.1 (documented and freely available for download online at: `http://surfer.nmr`

.mgh.harvard.edu/) [Dale et al., 1999; Fischl et al., 2004]. Steps for phenotypical measures extraction included skull stripping [Ashburner et al., 2005], Non-uniform intensity Normalization (N3) [Sled et al., 1998], segmentation using the Desikan–Killiany atlas [Desikan et al., 2006] and extraction of the MRI-derived phenotypes.

Harmonization was carried out preserving age, sex, diagnosis and intracranial volume (ICV / eTIV), and each study was considered a site whose systematic biases are expected to be corrected [Reynolds et al., 2022].

		N		Age in years ± SD [age range]	
site	Group \ Sex	Female	Male	Female	Male
A4	CN	728	515	71.3 ± 4.5 [65.0 - 85.7]	72.9 ± 5.1 [65.0 - 85.7]
ABIDE I	Autism	59	434	16.3 ± 8.1 [8.1 - 45.0]	17.4 ± 8.6 [7.0 - 64.0]
	CN	98	446	15.4 ± 6.6 [7.8 - 46.0]	17.4 ± 7.6 [6.5 - 48.0]
ABIDE II	Autism	74	416	13.1 ± 8.1 [5.2 - 54.0]	15.0 ± 9.3 [5.1 - 62.0]
	CN	175	397	13.5 ± 7.4 [5.9 - 46.6]	15.6 ± 10.0 [5.9 - 64.0]
ADNI	AD	156	190	73.6 ± 7.8 [55.0 - 91.0]	75.0 ± 7.6 [55.0 - 90.0]
	CN	444	345	71.8 ± 6.2 [55.0 - 90.0]	74.0 ± 6.1 [60.0 - 90.0]
	MCI	406	555	71.5 ± 7.8 [55.0 - 88.0]	73.6 ± 7.2 [54.0 - 90.0]
AIBL	AD	34	25	75.6 ± 8.1 [58.0 - 93.0]	72.7 ± 6.9 [60.0 - 83.0]
	CN	237	169	71.9 ± 6.0 [60.0 - 89.0]	73.0 ± 6.3 [60.0 - 90.0]
	MCI	44	46	75.8 ± 8.4 [56.0 - 95.0]	74.1 ± 5.6 [64.0 - 87.0]
MIRIAD	AD	22	17	69.5 ± 6.3 [58.1 - 80.2]	69.4 ± 7.5 [55.7 - 86.1]
	CN	11	12	66.3 ± 4.9 [58.8 - 73.8]	73.6 ± 7.2 [63.5 - 86.3]
OASIS	AD	45	41	75.6 ± 6.8 [62.9 - 95.6]	75.2 ± 7.3 [60.5 - 91.7]
	CN	162	109	67.6 ± 8.6 [45.7 - 88.8]	69.1 ± 8.3 [42.5 - 86.2]
PDPB	PD	97	129	67.8 ± 9.1 [50.7 - 90.0]	67.2 ± 9.7 [36.0 - 91.0]
PPMI	CN	51	84	59.1 ± 11.6 [31.0 - 81.9]	60.8 ± 11.7 [30.6 - 82.8]
	GenCohort PD	29	27	65.2 ± 8.0 [51.7 - 81.2]	63.9 ± 9.3 [32.2 - 78.6]
	GenCohort Unaff	52	35	61.1 ± 8.2 [33.7 - 84.3]	61.3 ± 7.6 [46.8 - 75.0]
	PD	94	164	61.2 ± 10.2 [33.5 - 81.7]	62.2 ± 9.4 [34.8 - 82.9]
	Prodromal	2	14	70.1 ± 4.3 [67.1 - 73.2]	69.0 ± 6.7 [61.7 - 82.5]
	SWEDD	15	26	61.7 ± 8.9 [46.8 - 77.6]	61.9 ± 9.9 [39.2 - 77.0]

Tab. 2.1.: Subject demographics of real data used in this work.

2.4 Results

Two versions of Fed-ComBat were used for comparison: a first one defining a linear model ($\phi(\mathbf{x};\boldsymbol{\theta}) = \mathbf{x}^\intercal\boldsymbol{\theta}$) to show equivalence with respect to ComBat and d-ComBat, and a second one defining ϕ as a multi-layer perceptron model (MLP) for nonlinear effect preservation. These two versions of Fed-ComBat were tested jointly with our extended proposal d-ComBat-GAM against their two centralized versions: ComBat[1] (a.k.a. NeuroComBat), ComBat-GAM[2] and the only reported distributed version d-ComBat-GAM[3].

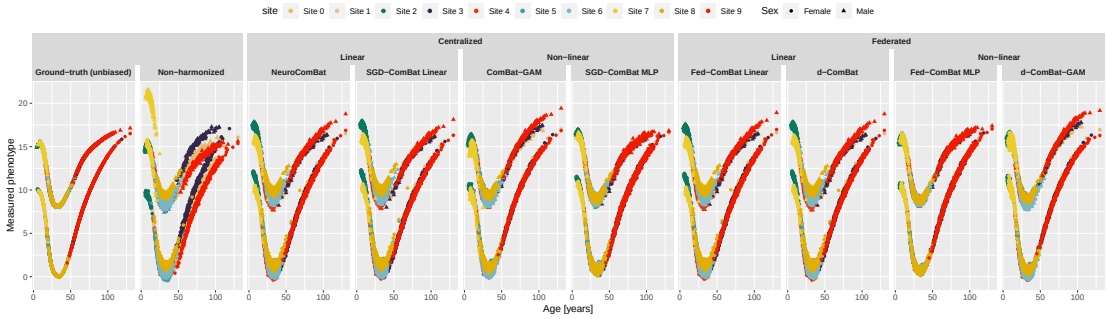
[1] https://github.com/Jfortin1/neuroCombat
[2] https://github.com/rpomponio/neuroHarmonize
[3] https://github.com/andy1764/Distributed-ComBat

2.4.1 Synthetic data

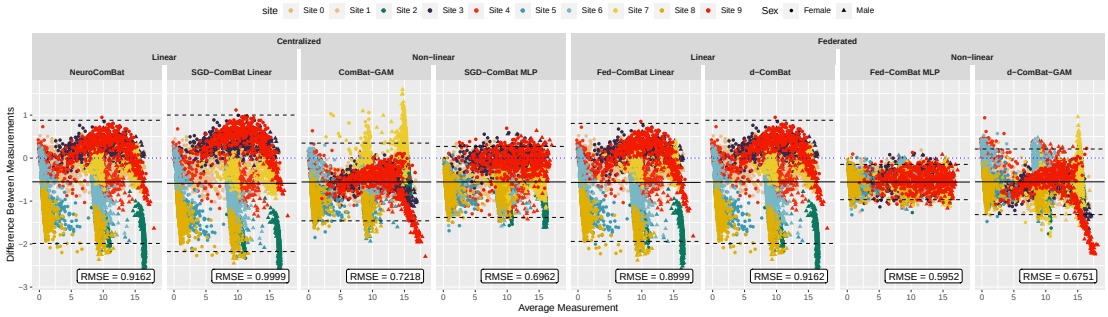
Reconstruction of an unbiased phenotype function depending on age and sex after harmonizing using different methods is shown in Figure 2.4a. As observed, centralized ComBat, d-ComBat, and Fed-ComBat in its linear version follow similar trends. It can also be observed that for nonlinear covariate effects linear correction can even induce biases due to its limitation to capture complex relationships between covariates and phenotypes (see Non-harmonized vs. NeuroCombat in Figure 2.4a).

While both ComBat-GAM and our distributed adaptation d-ComBat-GAM demonstrate improved correction for nonlinear covariate effects compared to ComBat (as shown in Figure 2.4a), it is worth noting that d-ComBat-GAM requires a priori definition of covariate interactions. In cases where these interactions are not defined, the performance of d-ComBat-GAM is reduced to that of ComBat. A better reconstruction of the target unbiased trajectories is achieved with Fed-ComBat using a two-layer perceptron neural network with 100 hidden units and with no hyperparameter tuning. Hyperparameter tuning was out of the scope of this work. Root mean square errors (RMSE) evaluating the reconstruction quality is presented in Table 2.2 for all centralized and federated methods.

Bland-Altman plots contrasting the difference between the different harmonization methods and the ground truth are presented in Figure 2.4b highlighting that there is less error variability (σ_ε) in estimating the unbiased function after harmonization for nonlinear methods and, particularly for Fed-ComBat in terms of the Root mean square error (RMSE = 0.5952) compared to the second federated best d-ComBat-GAM (RMSE = 0.6751).



(a) Reconstruction of an unbiased groundtruth function.



(b) Bland-Altmann plots and root mean square errors (RMSE) contrasting the different centralized and federated harmonization methods against the groundtruth.

Fig. 2.4.: Qualitative comparison across harmonization methods evaluating the quality of harmonization for a simulated phenotype with a nonlinear relationship with age and following a different trajectory per group (males and females). Results are shown for the extreme non-IID covariate distribution is depicted in Figure 2.1b. This can be seen as certain centers contain extremely young or old cohorts (tails). (a) Expected result after harmonization: groundtruth (unbiased), worst case scenario: biased data (non-harmonized), and harmonized phenotypes using the different ComBat methods considered and proposed in this work. As shown, ComBat could even insert biases when data is too heterogeneous. Fed-ComBat preserves better the real trajectories than the linear approaches and ComBat-GAM without the need of defining where the nonlinearities may be. Fed-ComBat using a multi-layer perceptron (MLP) shows the best reconstruction with an improvement of 35% in RMSE with respect to d-ComBat and 12% with respect to d-ComBat-GAM.

Setup	Approach	Method	$\lvert\bar{\varepsilon}\rvert$ (MAE)	σ_ε	RMSE
Federated	Non-linear	Fed-ComBat MLP	0.55710	**0.20960**	**0.59520**
Federated	Non-linear	d-ComBat-GAM	**0.55160**	0.38930	0.67510
Centralized	Non-linear	SGD-ComBat MLP	0.55420	0.42150	0.69620
Centralized	Non-linear	ComBat-GAM	0.55530	0.46120	0.72180
Federated	Linear	Fed-ComBat Linear	0.56540	0.70020	0.89990
Centralized	Linear	NeuroComBat	0.55230	0.73110	0.91620
Federated	Linear	d-ComBat	0.55230	0.73110	0.91620
Centralized	Linear	SGD-ComBat Linear	0.58640	0.81000	0.99990

Tab. 2.2.: Quantitative results for Figure 2.4a in terms of mean absolute error between methods and the target hidden unbiased trajectories ($\lvert\bar{\varepsilon}\rvert$), its standard deviation (σ_ε) and root mean square error (RMSE). Highlighted values correspond to the best metric across different methods.

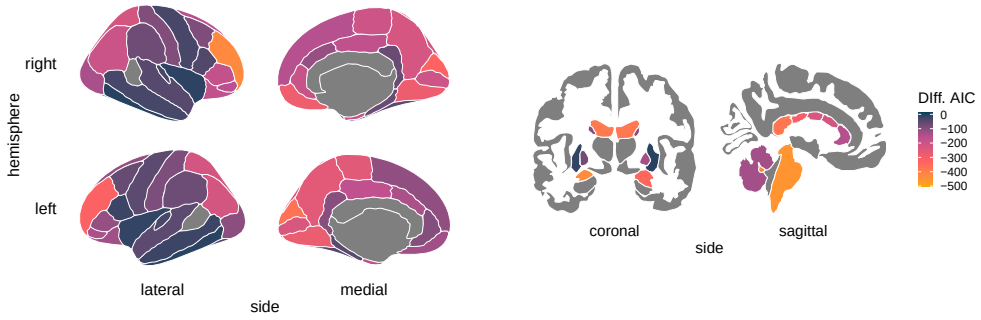
2.4.2 Brain MRI-data

To provide evidence for nonlinear covariate effects of age on brain phenotypes, we compared the goodness of fit of two models: a linear model and a generalized additive model (GAM). Data was firstly centralizd and harmonized accounting for study as the batch source using ComBat. The criterion used to evaluate the presence of nonlinearities was the difference in the Akaike Information Criterion (AIC) between the GAM and the linear model. When the difference is negative, it indicates that the GAM is a better fit, and the magnitude of the difference indicates by how much. Figure 2.5b shows the AIC metric across the regions of the brain using the Desikan-Killiany parcellation and cortical thickness across regions and the ASEG atlas for subcortical volumes as the dependent variables. The data was controlled for sex, diagnosis, ICV, and diagnosis group, and age was considered an independent variable. Figure 2.5a shows the top three regions that are better explained by a nonlinear model of age in terms of AIC, as well as the residuals after the regression illustrating remaining effects or trends. More detailed results on the AICs differences is presented in Table A.1.

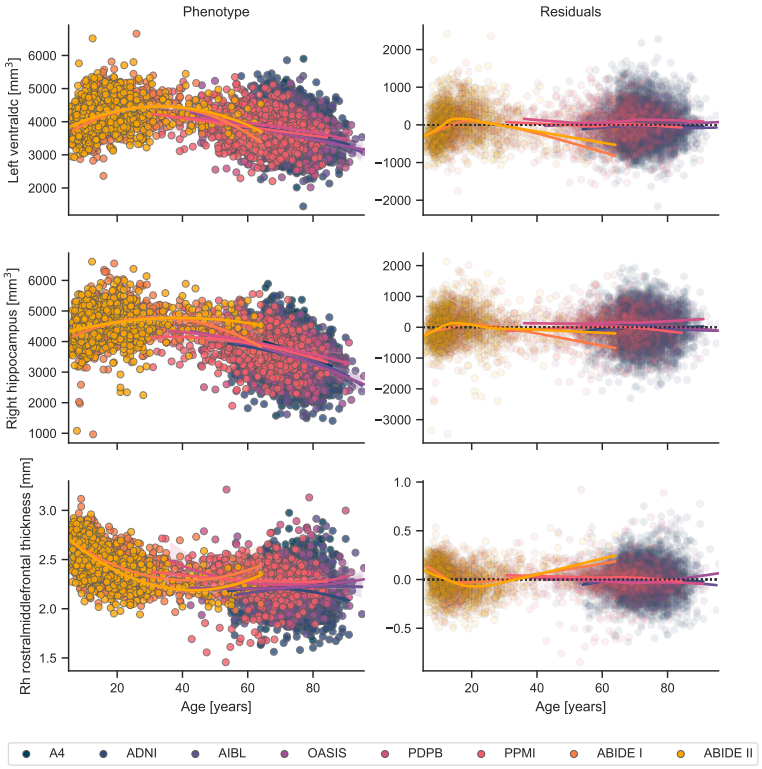
Then data was decentralized and we illustrate the ages trajectories by regrouping participants in two trajectories: cognitively normal were labeled as “Healthy controls“ while those diagnosed with a particular neurological disorder were labeled as “Atypical”. The harmonized phenotypes were obtained using as linear centralized method: ComBat, nonlinear centralized: d-ComBat, linear federated: d-ComBat and our adaptation Fed-ComBat using a linear model, and nonlinear federated: our extension d-ComBat-GAM and Fed-ComBat using an MLP with two layers and 100 hidden units.

A GAM was used to estimate these trajectories after each method controlling for Sex and ICV. Resulting trajectories are illustrated in Figure 2.7. The results show that our proposals Fed-ComBat Linear and d-ComBat-GAM produce consistent results with respect to ComBat, d-ComBat, and ComBat-GAM. After the integration of the MLP as covariate effect approximating function, both centralized and federated show similar trajectory derivation. The structure of the trajectories is preserved with the main difference being the exacerbation or attenuation of certain effects like the higher rate of deterioration in thickness for the right rostral middle frontal cortex in the young population presented in Figure 2.7 bottom right panel using Fed-ComBat MLP, and a less attenuated effect in the elder cohort on the right hippocampal volume confirming not only the viability of global harmonization without sharing data, but also doing it with a more privacy-aware approach using FL and integrating more complex models such as networks.

Trajectories traced after harmonizing with ComBat-GAM and d-ComBat-GAM suggest an increase in right hippocampal volume which does not seem to be supported by the literature and seems more like a residual site effect.



(a) Brain regions where a generalized additive model (GAM) provides a better fit than a linear model, as measured by the difference in Akaike Information Criterion (AIC) between the two models. The parameters for the GAM are set as in [Pomponio et al., 2020a], with age as a smoothing term, B-splines with 10 degrees of freedom or control points uniformly distributed between the minimum and maximum values, and a maximum polynomial of 3rd degree. A negative difference indicates a better fit with the GAM, while a positive difference indicates a better fit with the linear model.



(b) Results of regression using a GAM on the top 3 regions with the most discrepant AIC values on the left and residual plots illustrating goodness of fit on the right. Top: Ventral diencephalon, middle: Right hippocampus, bottom: Right hemisphere rostral middle frontal gyrus.

Fig. 2.5.: Evidence of nonlinear effects present in used cohorts. Subjects were harmonized by cohort using ComBat [Fortin et al., 2018].

Another important point to remark is the inflection points in age where right hippocampal volume reduction begins. While the non-harmonized data shows that reduction starts to occur in the late 50s, linear approaches exacerbate this effect in this cohorts suggesting a declination after 25 years old in cognitively normal populations, which can be questionable with respect to what has been reported so far in this region [Nobis et al., 2019; De Francesco et al., 2021; Liu et al., 2021b].

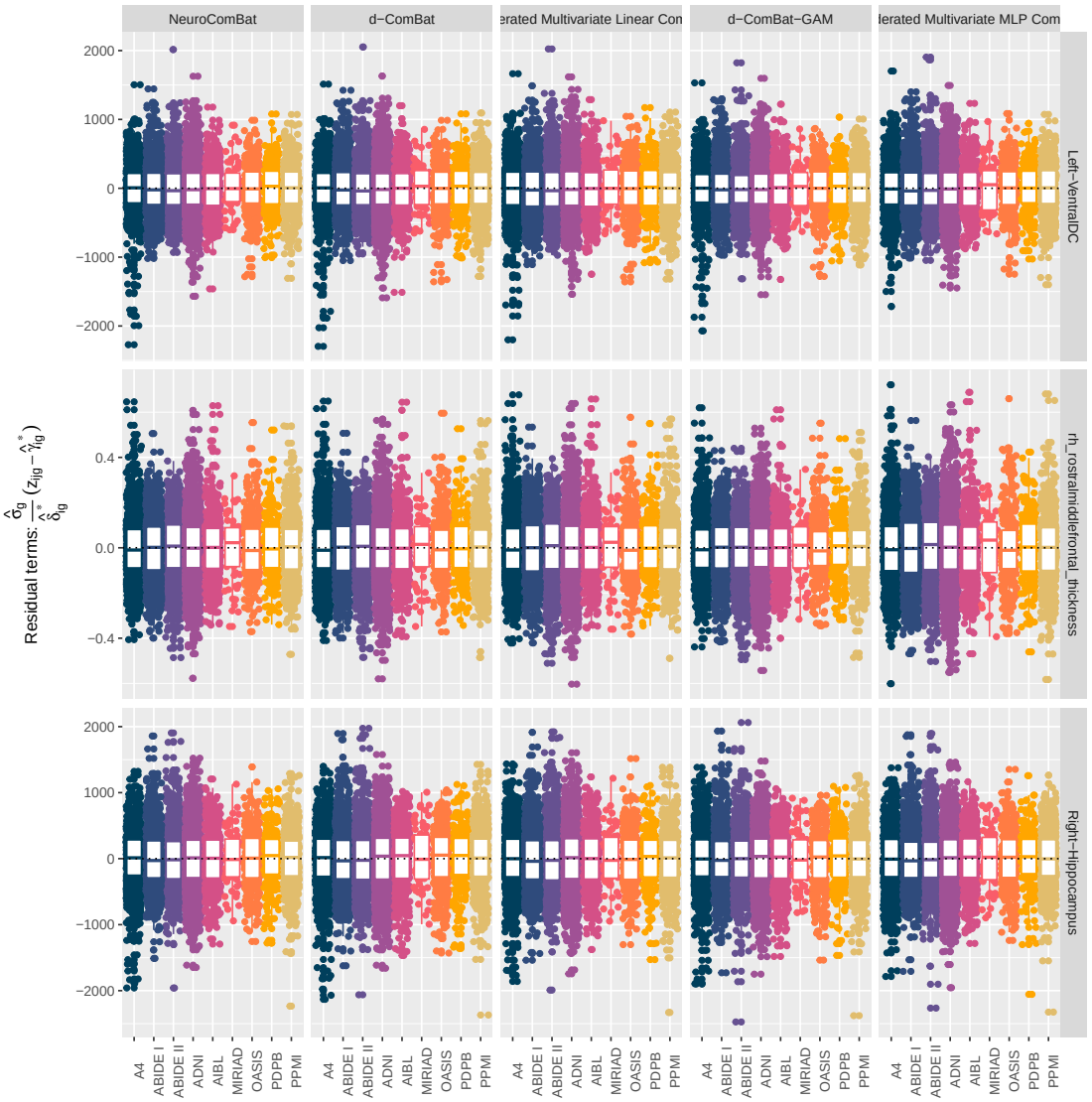


Fig. 2.6.: Residual terms after harmonization for top three regions discuses in Figure 2.5a for federated methods. Additional comparisons using all the methods are presented in Appendix A.3.

Figure 2.6 shows the residuals of each distributed model comprehended in this work after harmonization. The residual harmonized term $\frac{\hat{\sigma}_g}{\hat{\delta}^*_{ig}}\left(z_{ijg}-\hat{\gamma}^*_{ig}\right)$ should follow a normal standard noise distribution by definition and should not have center residual effects. [Pomponio et al., 2020a] proposed fitting a GAM and then extract the residuals. However, by definition, fitting a GAM will work better with a ComBat model that is GAM-based.

Instead, we use the corresponding model for each method (e.g., linear for ComBat, GAM for ComBat-GAM and MLP for Fed-ComBat MLP). This results are almost identical and do not show residual batch effects after harmonization. This is due to the nature of the formulation of each ComBat model, suggesting this visualization not to be the best for qualitative evaluation.

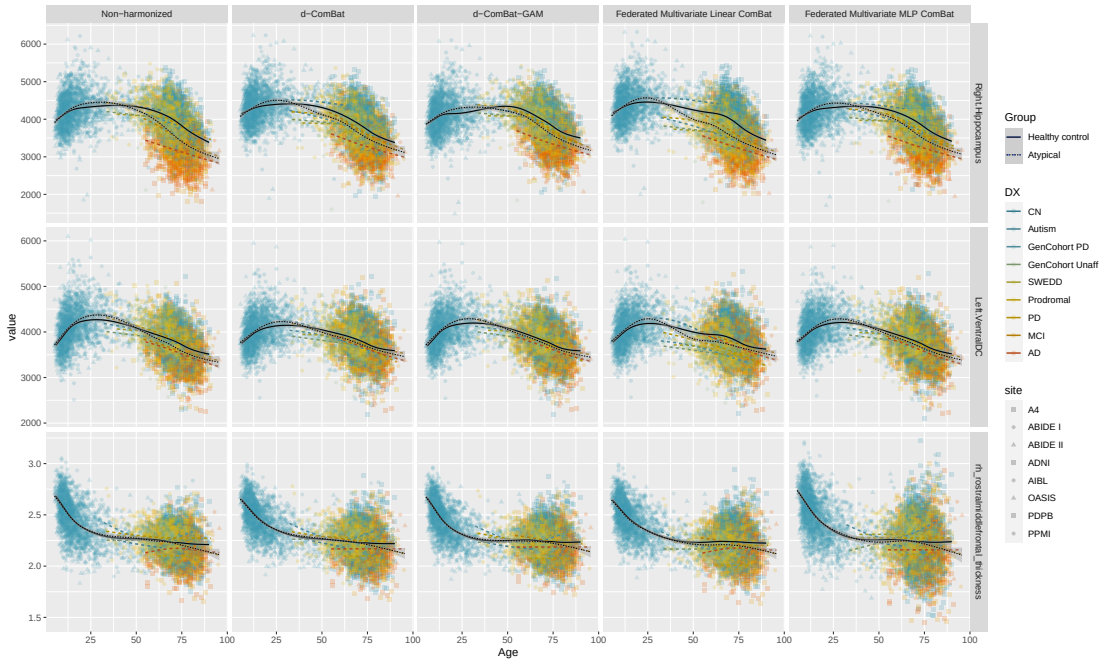


Fig. 2.7.: Harmonized phenotypes mentioned in Figure 2.5a in each row. Non-harmonized, d-ComBat and all the proposed methods in this work in each column. Trajectories are drawn after accounting for sex, and ICV after harmonization preserving age, sex, diagnosis and ICV.

Further analysis and evaluation on datasets and cohorts specially where data sharing is prohibited is required. As pointed by [Gebre et al., 2023] further exploration is needed in harmonization methods as some of the centralized methods evaluated in this study may result in non-satisfactory results.

2.5 Limitations and Future Directions

While Fed-ComBat has demonstrated promising results for data harmonization in a federated learning setting, there are potential limitations and challenges that should be considered, as well as opportunities for future research directions.

Model architecture: The choice of model architecture, such as the two-layer perceptron function employed in the current implementation of Fed-ComBat, can influence the performance of the harmonization process. In order to better capture complex nonlinearities in various harmonization problems, it may be useful to explore alternative architec-

tures and conduct hyperparameter tuning if necessary, while respecting the constraints required for integration within the Fed-ComBat framework (see Equation (2.4)).

Heterogeneity: The performance of Fed-ComBat may be affected by the level of heterogeneity across participating institutions. In cases where there is extreme disparity in data distributions, demographic compositions, or data collection protocols, the method might struggle to correct biases effectively. Therefore, the suitability of Fed-ComBat or any ComBat method in such cases should be carefully evaluated. Alternative directions could include exploring optimization methods such as FedProx[Li et al., 2020a].

Computational resources and connectivity: The current implementation of Fed-ComBat relies on the assumption that the participating institutions have enough computational resources and connectivity to perform federated learning. In practice, resource constraints or network issues might limit the applicability of this method, particularly in low-resource settings or in cases where institutions have unreliable or slow network connections.

Application to imaging data: The current implementation of Fed-ComBat focuses on derived phenotypes just like the first version of ComBat [Johnson et al., 2007a], but future work could explore the application of this method directly to imaging data. By making $\phi(\mathbf{x}_{ij}, \boldsymbol{\theta}_g)$ a convolutional neural network (CNN), Fed-ComBat could be extended to handle image data directly, which would open up new opportunities for harmonization in multi-centric imaging studies in federated setups.

By addressing these limitations and challenges, and exploring potential future directions, Fed-ComBat could be further improved and extended to a wider range of applications, ultimately enhancing its utility in federated learning and data harmonization for healthcare research.

2.6 Discussion and Conclusion

In conclusion, we have shown that it is possible to achieve acceptable results compared to the linear case without the need to share full covariance matrices or assume nonlinearities. The proposed Fed-ComBat method offers a novel, generalized approach for data harmonization, inspired by ComBat, and is suitable for scenarios where data sharing is restricted and nonlinear covariate effects might be present. By relying on federated learning, Fed-ComBat can capture these nonlinearities and interactions through its fully connected architecture, allowing for comparable results with centralized approaches and d-ComBat while keeping data decentralized and potentially avoiding data leakage.

In contrast to ComBat-GAM, Fed-ComBat does not require predetermined definitions of variables to be decomposed or possible interactions across variables (i.e., smoothing factor). This is because complex approximation functions like multi-layer perceptrons can effectively capture nonlinearities and interactions without explicit definitions.

As a future direction, it would be beneficial to explore the performance of Fed-ComBat on larger cohorts, as in [Bethlehem et al., 2022], where nonlinear effects are better exacerbated. This would provide further evidence of the method's effectiveness in handling complex harmonization problems in neuroimaging research.

2.7 Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 847579 (Marie Skłodowska-Curie Actions), and by the Centre INRIA d'Université Côte d'Azur "NEF" computation cluster. This study was partially supported by the Early Detection of Alzheimer's Disease Subtypes (E-DADS) project, an EU Joint Programme - Neurodegenerative Disease Research (JPND) project (see www.jpnd.eu). The project is supported under the aegis of JPND through the following funding organizations: United Kingdom, Medical Research Council (MR/T046422/1); Netherlands, ZonMW (733051106); France, Agence Nationale de la Recherche (ANR-19-JPW2-000); Italy, Italian Ministry of Health (MoH); Australia, National Health & Medical Research Council (1191535); Hungary, National Research, Development and Innovation Office (2019-2.1.7-ERA-NET-2020-00008).

2.8 Acknowledgments

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`https://adni.loni.usc.edu/`). The investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at `https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`. Data used in the preparation of this article was obtained from the Australian Imaging Biomarkers and Lifestyle flagship study of ageing (AIBL) funded by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) which was made available at the ADNI database (`https://adni.loni.usc.edu/aibl-australian-imaging-biomarkers-and-lifestyle-study-of-ageing-18-month-data-now-released/`). Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`https://adni.loni.usc.edu/`). Data used in this article were

provided by the Australian Imaging Biomarkers and Lifestyle flagship study of ageing (AIBL) funded by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) which was made available at the ADNI database (`https://adni.loni.usc.edu/aibl-australian-imaging-biomarkers-and-lifestyle-study-of-ageing-18-month-data-now-released/`). Data used in the preparation of this article were obtained from the Autism Brain Imaging Data Exchange (ABIDE) I database. Data used in this article were obtained from the Autism Brain Imaging Data Exchange (ABIDE) II database (`http://fcon_1000.projects.nitrc.org/indi/abide/abide_II.html`). The AIBL researchers contributed data but did not participate in the analysis or writing of this report. AIBL researchers are listed at `https://www.aibl.csiro.au`. Data used in the preparation of this article were obtained from the MIRIAD database. Data used in the preparation of this article were obtained from the Open Access Series of Imaging Studies (OASIS) database. Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database (`http://www.ppmi-info.org/`). Data used in the preparation of this article were obtained from the Pre-symptomatic Evaluation of Experimental or Novel Treatments for Alzheimer's Disease (A4) study.

2.9 Supplementary Material

For a more detailed exploration of the topics discussed in this section, including the relationship between the indicator matrix and its elements as composites of α_g and γ_{ig}, please refer to the supplementary material (specifically, Appendix A). The supplementary material provides additional insights and information on the optimization problem and the generalized approach proposed for comparing centralized and Fed-ComBat's formulation. It also includes harmonization results on synthetic and real data, evidence of non-linear effects on brain phenotypes, residual ComBat effects on real data, identifiability of ComBat parameters, and the CRediT author statement.

3 Federated Data Harmonization in Biomedical Research using Mixed Effects Models: A Focus on Conditional Variational Autoencoders

Data harmonization plays a crucial role in integrating and reconciling information from various sources and formats to navigate the inherent heterogeneity and extract meaningful insights. In multicentric neuroimaging studies, bias is introduced due to variations in acquisition protocols, scanner brands, and other factors. Current harmonization techniques, such as surrogate variable analysis and ComBat, are limited in capturing complex relationships and nonlinearities. Additionally, these methods rely on centralized data, conflicting with data protection regulations. Federated learning (FL) emerges as a promising paradigm for multicentric studies, ensuring data governance and security while complying with privacy regulations. However, existing FL methods have their limitations and may require retraining when new institutions join the study. This chapter explores the use of conditional generative models, specifically variational autoencoders (VAEs), for data harmonization in federated learning. The proposed approach aims to address the limitations of existing techniques by capturing complex relationships, accounting for mixed effects, and ensuring data privacy. This work-in-progress chapter presents the motivation,

methodology, results, and challenges of using CVAEs for data harmonization in the context of federated learning. We present some preliminary results on synthetic data and we expect to move forward to imaging aplications.

3.1 Introduction

In multicentric neuroimaging studies, bias is introduced due to different acquisition protocols, magnetic resonance imaging (MRI) scanner brands, and other factors [Orlhac et al., 2022]. This bias can significantly affect the generalizability of the derived models and limit the potential of collaborative research. Several harmonization techniques have been developed to address this bias, including surrogate variable analysis [Leek et al., 2012], ComBat [Johnson et al., 2007a] and random effect models [Kim et al., 2022]. However such methods remain univariate and are not able to properly capture more complex relationships like nonlinearities. Also, these methods currently rely on data being centralized which as discussed, is conflicting with the current data protection regulations.

Federated learning (FL) is an emerging paradigm in the field of machine learning that enables multicentric studies in brain imaging data to remain compliant with data protection regulations such as the General Data Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA). By distributing the learning process across multiple institutions while keeping the data locally, FL ensures better data governance and security [Li et al., 2020a].

In the field of neuroimaging data harmonization, several methods have been developed to address the challenge of scanner-related biases and inter-site variability. Notable among these are ComBat [Johnson et al., 2007a], NeuroHarmony [Garcia-Dias et al., 2020], and domain adaptation methods on raw imaging data [Zuo et al., 2021]. These methods have significantly contributed to the field, but it's crucial to understand their inherent limitations and the contexts where they excel.

ComBat, one of the most widely used methods, is known for adjusting batch effects in high-throughput data. It employs empirical Bayes frameworks to estimate batch effect magnitudes and adjust for them. Although ComBat is effective and commonly used, it is designed for simpler linear adjustments and may fall short when dealing with more complex, non-linear data transformations that are often encountered in neuroimaging data.

NeuroHarmony, a specialized tool, was developed to harmonize neuroimaging data and mitigate scanner-related biases in MRI datasets. It uses a machine learning approach and

a set of Image Quality Metrics (IQMs) to harmonize individual images from previously unseen scanners. Despite its novel approach, NeuroHarmony is primarily designed for linear shifts and scaling adjustments. Also, when data from a new scanner is introduced, the model may require retraining, posing challenges in dynamic research environments.

Zuo's method, on the other hand, uses convolutional layers for domain adaptation on raw imaging data, offering flexibility and the ability to handle more complex data transformations. However, this method is not without its limitations. Improper training of the convolutional layers may introduce artifacts, potentially distorting the data and leading to inaccurate analyses.

The aim of this chapter is to explore conditional generative models such as variational autoencoders (VAEs) for data harmonization in federated learning and their potential to address the limitations of existing techniques by capturing complex relationships and interactions while accounting for mixed effects introduced by different sites or institutions.

The chapter is organized as follows: Section 3.2 provides a background on the importance of data harmonization in biomedical research, while Section 3.3 introduces the concept of CVAEs and their potential for addressing data harmonization challenges and a detailed methodology for implementing CVAEs in this context. Section 3.4 discusses the results and performance evaluation of CVAEs in data harmonization, followed by an exploration of the challenges and future directions in Section 3.5. The chapter concludes with a summary of the key findings in Section 3.6.

3.2 Challenges of Data Harmonization

Data harmonization is a process that aims to create consistency and compatibility between datasets originating from different sources, formats, and measurement units, thereby enabling researchers to analyze and draw meaningful conclusions from the combined data. The need for data harmonization arises from the growing volume and complexity of biomedical data, which often exhibit heterogeneity due to variations in data collection methodologies, experimental conditions, and measurement units. In this section, we discuss the significance of data harmonization in biomedical research, as well as the common challenges and issues associated with harmonizing heterogeneous datasets.

Data harmonization plays a critical role in biomedical research, as it allows for the integration and comparison of data from different sources, ultimately enhancing the reproducibility and generalizability of research findings [Fortin et al., 2014]. Furthermore, data harmonization can facilitate the discovery of novel associations and patterns in data,

leading to new insights and hypotheses in biomedical research. By enabling researchers to leverage the full potential of the available data, data harmonization contributes to the acceleration of scientific discovery and the improvement of patient care.

While existing methods such as ComBat, NeuroHarmony, and Zuo's method provide valuable solutions for harmonizing neuroimaging data, they all rely on data centralization, are either univariate, or are tailored to specific types of data. These characteristics limit their applicability in diverse, multi-site studies that require a multivariate approach and mandate decentralized data handling for privacy preservation.

In response to these limitations, our proposed method aims to approximate the underlying data distribution, capturing the batch effects in a multivariate manner in a federated setup. This approach allows us to account for covariate effects comprehensively. Our technique leverages a regularized generative model, specifically a conditional variational autoencoder (CVAE), trained in a Bayesian fashion [Sohn et al., 2015].

The Bayesian nature of our method applies not only to the latent space of the CVAE but also to its parameters. This dual application of Bayesian principles helps avoid overfitting while preserving the essential covariate effects. The CVAE model is trained using the Federated Averaging (FedAvg) algorithm [McMahan et al., 2017], ensuring that the training process respects a decentralized data setup where data is not exchanged.

Once the model is trained, we exploit the generative abilities of the CVAE to mitigate the batch effects. We achieve this by conditioned sampling, effectively eliminating the bias in expectation. Consequently, our proposed federated, multivariate harmonization method offers a promising solution to tackle the challenges in neuroimaging data harmonization, prioritizing both effectiveness and privacy preservation.

3.3 Methods

In what follows, we outline the methods used for data harmonization in a federated setting using Conditional Variational Autoencoders (CVAEs). We introduce the concept of Federated Harmonization with CVAEs and discuss the derivation of the Evidence Lower Bound (ELBO) as the objective function for the harmonization process.

3.3.1 Federated Harmonization with Conditional Variational Autoencoders (CVAEs)

To address the challenges of data harmonization in a federated setting, we propose a federated harmonization framework based on Conditional Variational Autoencoders (CVAEs) [Sohn et al., 2015]. CVAEs have shown promise in capturing complex relationships and generating harmonized data samples conditioned on specific attributes or features [Zuo et al., 2021; Russkikh et al., 2020].

Let $\mathbf{y}$ denote the site indicator variable, $\mathbf{z}$ represent the latent space of the autoencoder, $\mathbf{x}$ is an observation of a design matrix containing the measured phenotypes and the covariates to be preserved, and θ and ϕ parametrize the autoencoder. We model θ and ϕ as random effects, centered at zero and with a variance σ^2. We begin by defining the notation and deriving the evidence lower bound (ELBO).

3.3.2 Derivation of the Variational Lower Bound

In this subsection, we will focus on the derivation of the variational lower bound for our proposed model, based on the work of [Sohn et al., 2015]. This derivation will provide a detailed explanation of the mathematical framework underlying our model's training process.

Consider a set of observed phenotypes and covariates denoted by $\mathbf{x}$ coming from a design observation matrix. Additionally, let $\mathbf{y}$ represent the site or batch indicator variable, which indicates the source or origin of the data points. This variable allows the model to account for any variations or discrepancies that may arise from different sources. As proposed by [Sohn et al., 2015], we aim to approximate the joint distribution to gain valuable insights into the dependence of the observed data with respect to the batch effects by defining a set of latent variables or factors and site indicators as part of a generative process.

Such joint distribution over the observed variables, latent variables, and parameters, can be defined as:

$$p(\mathbf{x}, \mathbf{y}, \mathbf{z}, \theta, \phi) = p(\mathbf{x}|\mathbf{y}, \mathbf{z}, \theta)p(\mathbf{z}|\mathbf{y}, \phi)p(\mathbf{y})p(\theta)p(\phi) \tag{3.1}$$

Here, $p(\mathbf{x}|\mathbf{z}, \mathbf{y}, \theta)$ is the likelihood of the observed phenotypes and the covariates $\mathbf{x}$ given the latent variable $\mathbf{z}$, the batch indicator variable $\mathbf{y}$, and the parameters θ of the decoder.

Similarly, $p(\mathbf{z}|\mathbf{x}, \mathbf{y}, \phi)$ is the prior over the latent variable $\mathbf{z}$ given the observed variables $\mathbf{x}$, the site indicator variable $\mathbf{y}$, and the parameters ϕ of the encoder. Finally, $p(\theta)$ and $p(\phi)$ are the priors over the parameters θ and ϕ, respectively. These priors represent our initial beliefs about the distribution of the parameters before observing any data. During training, these parameters are updated to maximize the likelihood of the observed variables.

From Equation (3.1), we can derive the log marginal likelihood as:

$$\log p(\mathbf{x} \mid \mathbf{y}) = \log \int_{\mathbf{z}} \int_{\theta} \int_{\phi} p(\mathbf{x}, \mathbf{y}, \mathbf{z}, \theta, \phi) d\phi d\theta d\mathbf{z} \tag{3.2}$$

Now, we introduce an approximate posterior distribution $q(\mathbf{z}, \theta, \phi|\mathbf{x}, \mathbf{y})$ to make the problem tractable. By multiply and divide the integrand by this distribution:

$$\log p(\mathbf{x} \mid \mathbf{y}) = \log \int_{\mathbf{z}} \int_{\theta} \int_{\phi} \frac{p(\mathbf{x}, \mathbf{y}, \mathbf{z}, \theta, \phi)}{q(\mathbf{z}, \theta, \phi \mid \mathbf{x}, \mathbf{y})} q(\mathbf{z}, \theta, \phi \mid \mathbf{x}, \mathbf{y}) d\phi d\theta d\mathbf{z} \tag{3.3}$$

Then we use Jensen's inequality, which states that the log of an expectation is greater than or equal to the expectation of the log, hence:

$$\log p(\mathbf{x} \mid \mathbf{y}) \geq \int_{\mathbf{z}} \int_{\theta} \int_{\phi} q(\mathbf{z}, \theta, \phi \mid \mathbf{x}, \mathbf{y}) \log \frac{p(\mathbf{x}, \mathbf{y}, \mathbf{z}, \theta, \phi)}{q(\mathbf{z}, \theta, \phi \mid \mathbf{x}, \mathbf{y})} d\phi d\theta d\mathbf{z} \tag{3.4}$$

Finally we can rewrite Equation (3.4) in terms of expectations:

$$\begin{aligned} \mathcal{L}(q) = & \mathbb{E}_q[\log p(\mathbf{x} \mid \mathbf{y}, \mathbf{z}, \theta)] + \mathbb{E}_q[\log p(\mathbf{z} \mid \mathbf{y}, \phi)] + \mathbb{E}_q[\log p(\theta)] + \mathbb{E}_q[\log p(\phi)] \\ & - \mathbb{E}_q[\log q(\mathbf{z} \mid \mathbf{x}, \mathbf{y}, \phi)] - \mathbb{E}_q[\log q(\theta)] - \mathbb{E}_q[\log q(\phi)] \end{aligned} \tag{3.5}$$

Equation (3.5) presents the Evidence Lower Bound (ELBO) denoted by $\mathcal{L}$. The ELBO is a measure used to approximate the likelihood of observed data in probabilistic models, especially valuable when the likelihood function is intractable, as it provides a surrogate function that serves as an approximation of the true likelihood. Here q is short for $q(\mathbf{z}, \theta, \phi|\mathbf{x}, \mathbf{y})$. The expectation is taken over $\mathbf{z}$, θ and ϕ.

In the transition from the expanded expectation to the ELBO, we recognize certain terms as Kullback-Leibler (KL) divergences. Allowing us to rewrite the ELBO as:

$$\begin{aligned}\mathcal{L}(q) =& \mathbb{E}_q[\log p(\mathbf{x} \mid \mathbf{y}, \mathbf{z}, \theta)] \\ & - KL(q(\mathbf{z} \mid \mathbf{x}, \mathbf{y}) | p(\mathbf{z} \mid \mathbf{y})) - KL(q(\theta) | p(\theta)) - KL(q(\phi) | p(\phi))\end{aligned} \tag{3.6}$$

Here, $KL(p||q) = \mathbb{E}_p[\log p - \log q]$ is the Kullback-Leibler divergence.

In Equation (3.6), we categorize each term into two main components. The *reconstruction term*, represented by $\mathbb{E}_q[\log p(\mathbf{x} \mid \mathbf{y}, \mathbf{z}, \theta)]$, is the expected log-likelihood of the observed data under the model. This term ensures the model explains the observed data well.

Finally, the last three terms collectively represent the *Kullback-Leibler (KL) divergence* term, which measures the difference between the approximate posterior q and the prior p. Specifically, $KL(q(\mathbf{z} \mid \mathbf{x}, \mathbf{y}) | p(\mathbf{z} \mid \mathbf{y}))$, $KL(q(\theta) | p(\theta))$ and $KL(q(\phi) | p(\phi))$ account for the discrepancy between the approximate posterior distributions and the prior distributions of $\mathbf{z}$, θ and ϕ respectively. These terms act as a regularization that encourages the learned parameter distributions to be close to their priors. By doing so, it helps to prevent overfitting and ensures that the learned parameters are not too far from their initial values, providing a form of prior knowledge or constraint on the model.

We rely on stochastic variational inference (SVI) for the optimization scheme, which approximates the true posterior by minimizing the negative ELBO [Hoffman et al., 2013]. In the federated setting, we employ federated averaging (FedAvg) to optimize the global parameters [McMahan et al., 2017]. This involves aggregating the local updates of the variational parameters θ and ϕ across sites, weighted by the number of samples at each site. This approach allows for the harmonization of multicentric data while addressing the privacy and regulatory concerns associated with sharing raw data between institutions.

3.3.3 Harmonization

Once the CVAE model has been trained using a standard federated learning approach such as FedAvg, we proceed to the harmonization step, which aims to align and reconcile the distributions of the data across different institutions. This step is essential for ensuring data consistency and comparability, enabling meaningful analysis and modeling within the federated learning framework.

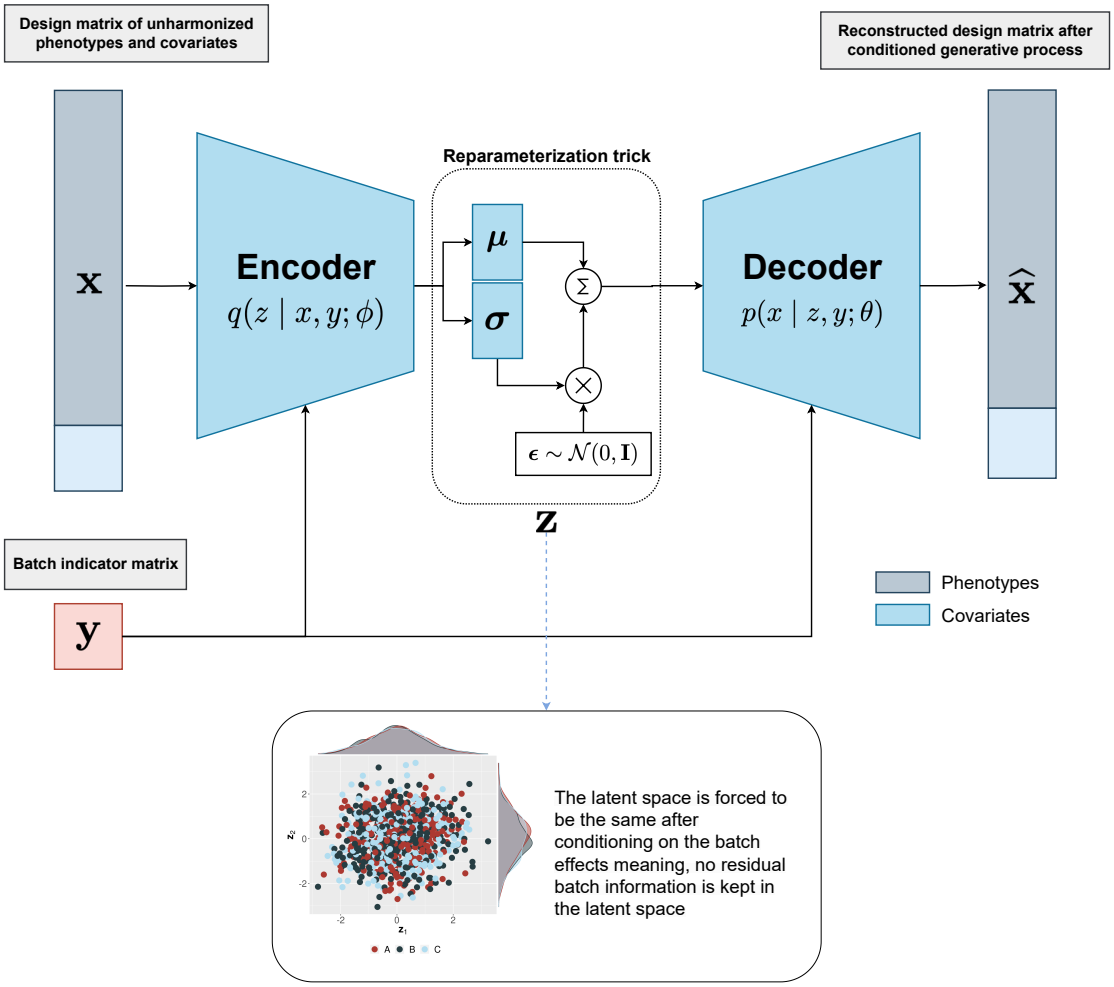


Fig. 3.1.: Architecture of the Conditional Variational Autoencoder (CVAE) for data harmonization. The input data $\mathbf{x}$ consists of a design matrix containing both phenotypes and covariates, with the goal of preserving covariate effects. The conditioning variable $\mathbf{y}$ influences the generation process, while the latent variables $\mathbf{z}$ capture underlying representations. The decoder, parametrized by θ, reconstructs the input data based on the latent variables and conditioning variable. The encoder, parametrized by ϕ, captures the approximate posterior or inference model $q(\mathbf{z}|\mathbf{x}, \mathbf{y}; \phi)$. In the bottom part, an example illustrates the latent space of the CVAE. Conditioning on the batch effects ensures that the latent space remains consistent, removing residual batch information.

We aim to mitigate any remaining batch effects captured by the parameters of the CVAE by incorporating the site indicator variable $\mathbf{y}$. This variable represents the source institution or batch information for each data point. By sampling $\mathbf{y}$ from a categorical distribution, denoted as $\mathbf{y} \sim \text{Cat}(p_1, p_2, \ldots, p_k)$, we can effectively average out the batch effects and diminish their impact on the reconstruction process. The probabilities $p_1, p_2, \ldots, p_k$ determine the likelihood of selecting a specific institution or batch.

Sampling $\mathbf{y}$ ensures that we obtain in expectation a diverse set of site indicator values for each observation $\mathbf{x}$. By considering multiple samples of $\mathbf{y}$ for each data point, we capture the variability and distribution of the site indicator information. This averaging process

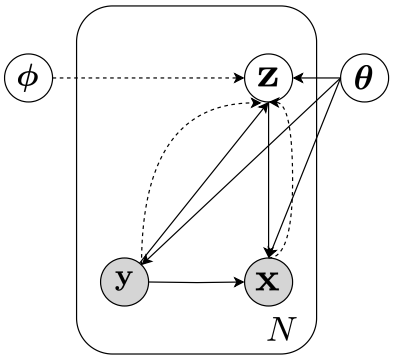


Fig. 3.2.: Generative process (graph model) for the proposed CVAE. Latent variables $\mathbf{z}$ and parameters $\boldsymbol{\theta}$ and ϕ are drawn from their respective priors. Observed data $\mathbf{x}$ and $\mathbf{y}$ are generated from the distributions $p(\mathbf{x}|\mathbf{z}, \mathbf{y}; \boldsymbol{\theta})$ and $p(\mathbf{y})$ respectively.

helps to minimize the impact of any specific batch effect on the reconstruction and leads to a less biased representation of the underlying features.

By incorporating the expected values of $\mathbf{y}$ during the reconstruction step, we mitigate the influence of batch effects, allowing for a more accurate and unbiased reconstruction of the original data $\mathbf{x}$. This approach improves the harmonization process by reducing the residual variability associated with different institutions or batches, ultimately leading to a more consistent and reliable representation of the underlying features across the dataset.

We illustrate our proposed method in Algorithm 3, in which we propose a workflow for federated data harmonization across multiple institutions. The algorithm begins by constructing a design matrix, preprocessed through federated standardization of numerical variables, where pooled means and variance deviations are computed for each institution. The design matrix is then standardized accordingly. Next, the algorithm performs federated training of the CVAE using FedAvg to update global parameters. Local updates from each institution are aggregated to refine the global parameters until convergence or budget achieved. Finally, the algorithm removes batch effects through a harmonization step, generating reconstructions for each sample using CVAE and averaging them to obtain the harmonized data. With this workflow we expect to provide a practical solution for harmonizing data in a federated setting, facilitating collaboration among institutions and enhancing the reliability of multi-institutional studies.

3.4 Results

We designed a synthetic dataset to evaluate the performance of the proposed harmonization method. The dataset incorporates covariates such as Sex, Age, and Scanner effects, which are commonly encountered in neuroimaging studies. These covariates capture

Algorithm 3: Federated Harmonization Framework Using CVAEs

Result: Harmonized data $\hat{\mathbf{x}}$
Input: Design matrix $\mathbf{x}$, batch indicator matrix $\mathbf{y}$
Output: Harmonized data $\hat{\mathbf{x}}$

Step 1: Design matrix construction
1.1 Federated standardization of numerical variables
Compute pooled mean $\overline{\mathbf{x}}$ and standard deviation $\mathbf{s}$ of $\mathbf{x}$:
foreach *site* k **do**
  $\overline{\mathbf{x}}_k \leftarrow \frac{\sum_{i=1}^{N_k} \mathbf{x}_{ki}}{N_k}$, $\mathbf{s}_k^2 \leftarrow \frac{\sum_{i=1}^{N_k} (\mathbf{x}_{ki} - \overline{\mathbf{x}}_k)^2}{N_k - 1}$
end
Compute pooled mean and variance deviation:
$\overline{\mathbf{x}} \leftarrow \frac{\sum_{k=1}^{K} N_k \cdot \overline{\mathbf{x}}_k}{N}$, $\mathbf{s}^2 \leftarrow \frac{\sum_{k=1}^{K} (N_k - 1) \cdot \mathbf{s}_k^2}{\sum_{k=1}^{K} (N_k - 1)}$
Standardize $\mathbf{x}$: $\mathbf{x} \leftarrow \frac{\mathbf{x} - \overline{\mathbf{x}}}{\mathbf{s}}$

Step 2: Federated training of CVAE
2.1 Federated Averaging
Initialize global parameters θ^0, ϕ^0
for *each round* $t = 1, 2, \ldots, T$ **do**
  for *each local node (institution)* k **do**
    Compute local updates $\Delta\theta_k^t$, $\Delta\phi_k^t$ by optimizing local objective (Equation (3.6))
  end
  Update global parameters: $\theta^t \leftarrow \theta^{t-1} + \frac{1}{K} \sum_{k=1}^{K} \Delta\theta_k^t$, $\phi^t \leftarrow \phi^{t-1} + \frac{1}{K} \sum_{k=1}^{K} \Delta\phi_k^t$
end
Define optimal parameters θ and ϕ

Step 3: Harmonization (batch effect removal)
for *each sample* $i = 1, 2, \ldots, N_s$ **do**
  3.1 Fix $\mathbf{x}$ and sample $\mathbf{y}$ from a Categotical distribution
  $\mathbf{y}_i \sim \text{Cat}(p_1, p_2, ..., p_k)$
  3.2 Compute reconstruction: $\hat{\mathbf{x}}_i \leftarrow \text{CVAE}(\mathbf{x}, \mathbf{y}_i; \theta, \phi)$
end
Return average reconstructions: $\hat{\mathbf{x}} \leftarrow \frac{1}{N_s} \sum_{i=1}^{N_s} \hat{\mathbf{x}}_i$

important factors that can introduce variations in brain measures across different data sources or institutions. The graphical model on how data was generated is illustrated in Figure 2.2.

By using principal component analysis (PCA) and analysis of variance (ANOVA), we gain insights into the effectiveness of the data harmonization method. Figure 3.3 showcases the results obtained through PCA. The top plot depicts the PCA of the non-harmonized features, where distinct clusters corresponding to different sites or institutions are evident.

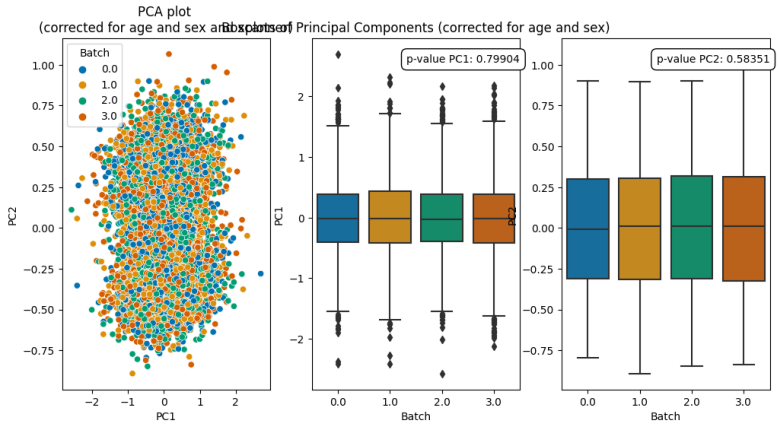
These clusters highlight the presence of batch effects, indicating potential variations and biases in the data.

After applying the data harmonization method, as illustrated in the middle plot, the feature distributions become more aligned, reducing the variations between sites and leading to a more integrated representation of the data. The clusters become less distinct, suggesting that the harmonization process successfully mitigates the impact of batch effects and promotes a more unified view of the dataset.

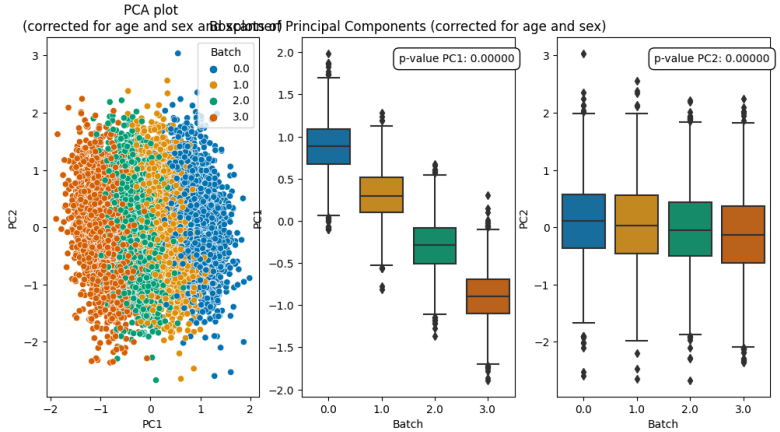
To further assess the impact of harmonization, ANOVA is conducted on the harmonized data. The results reveal a significant reduction in the variability attributed to the site indicator variable, indicating a successful harmonization outcome. This reduction in site-specific variability supports the notion that the data harmonization approach effectively mitigates batch effects and enables a more reliable and unbiased analysis of the dataset.

Figure 3.4 provides an overview of the feature distributions across multiple sites or institutions. The top plot reveals distinct clusters corresponding to different sites, indicating the presence of batch effects and potential biases in the data. In the middle plot, the feature distribution after harmonization is shown, demonstrating a reduction in variations between sites. This alignment of feature distributions aims to achieve a more consistent representation of the underlying features, effectively mitigating the impact of batch effects. The bottom plot represents the desired outcome: an unbiased distribution where batch effects have been entirely eliminated. This homogeneous feature distribution enables the collective analysis of data from different institutions, free from the confounding effects of batch variations.

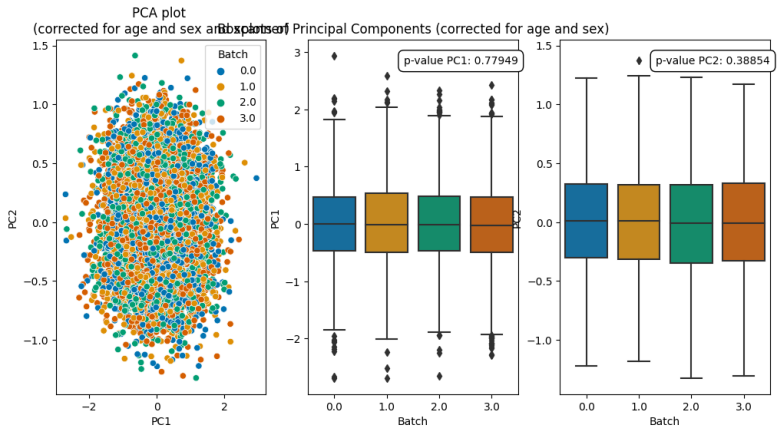
We evaluated the effectiveness of the harmonization method using Bland-Altman plots. Figure 3.5 presents the feature differences between the unbiased dataset and the unharmonized dataset (Figure 3.5a), as well as between the unbiased dataset and the harmonized dataset using the CVAE method (Figure 3.5b), stratified by batch effects. In these plots, the x-axis represents the mean of the feature values, while the y-axis shows the difference between the corresponding feature values. The colored dots indicate different batches, enabling the assessment of batch effects on feature differences. Comparing the two plots, we observe a substantial reduction in feature differences in the harmonized dataset compared to the unharmonized dataset. This finding highlights the efficacy of the CVAE-based harmonization approach in mitigating the confounding effects of batch variations, thereby achieving a more consistent representation of the underlying features across diverse institutions or sites.



(a) PCA of the feature space of the test dataset after harmonization. The harmonization process effectively aligns the distributions of the features from different institutions, resulting in reduced batch effects.



(b) PCA of the feature space of the non-harmonized features. Without harmonization, the features exhibit distinct clusters corresponding to different institutions, indicating the presence of significant batch effects.



(c) PCA of the feature space of the unbiased target features (ground truth). These features represent the ideal scenario where batch effects are completely eliminated, resulting in a homogeneous distribution of data points across different institutions.

Fig. 3.3.: Principal Component Analysis (PCA) of the feature space illustrating the effects of harmonization. (a) Shows the harmonized features, (b) shows the non-harmonized features, and (c) shows the unbiased target features. Comparing these plots provides insights into the effectiveness of the harmonization method in reducing batch effects and achieving a more consistent representation of the underlying features.

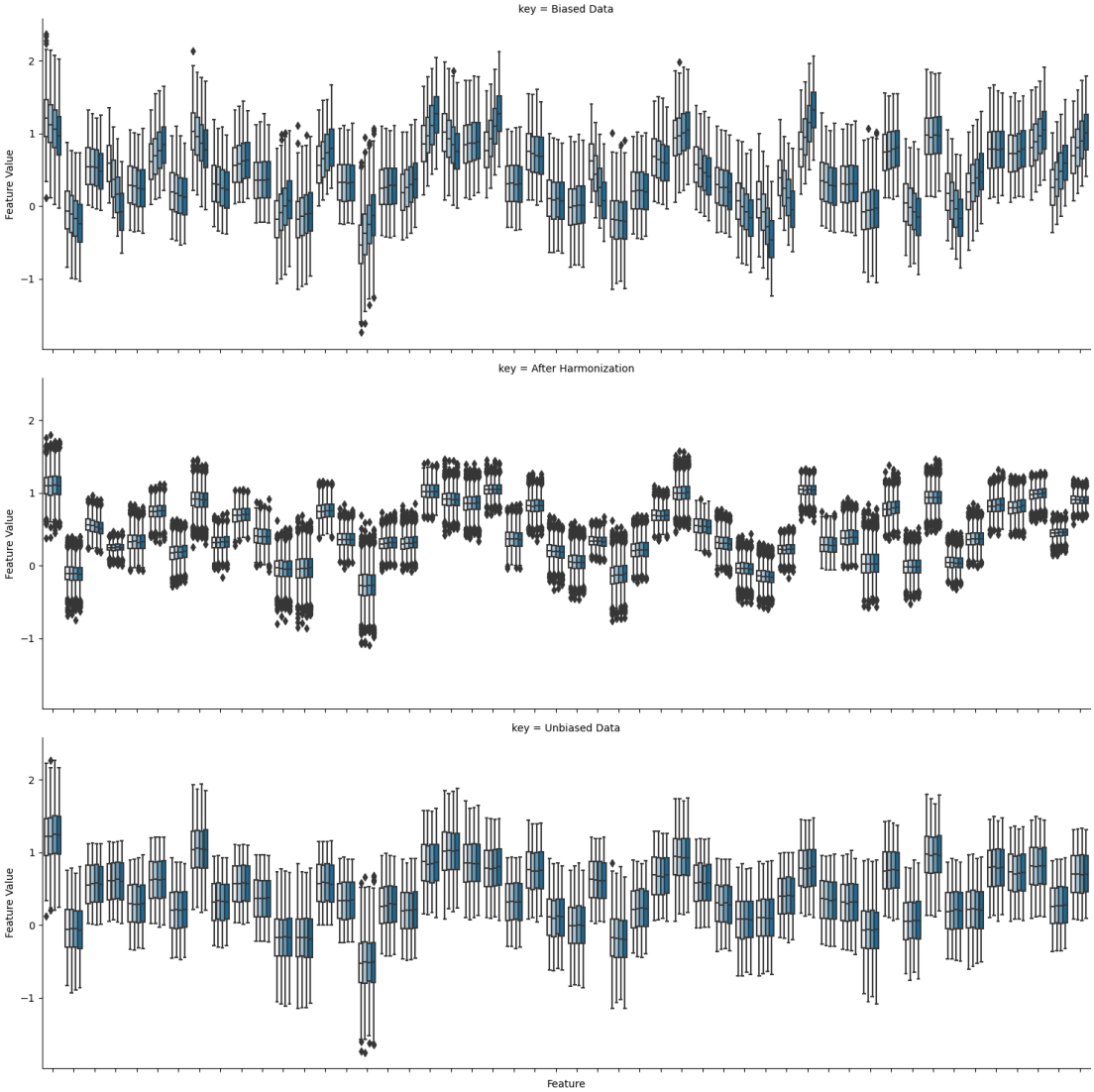
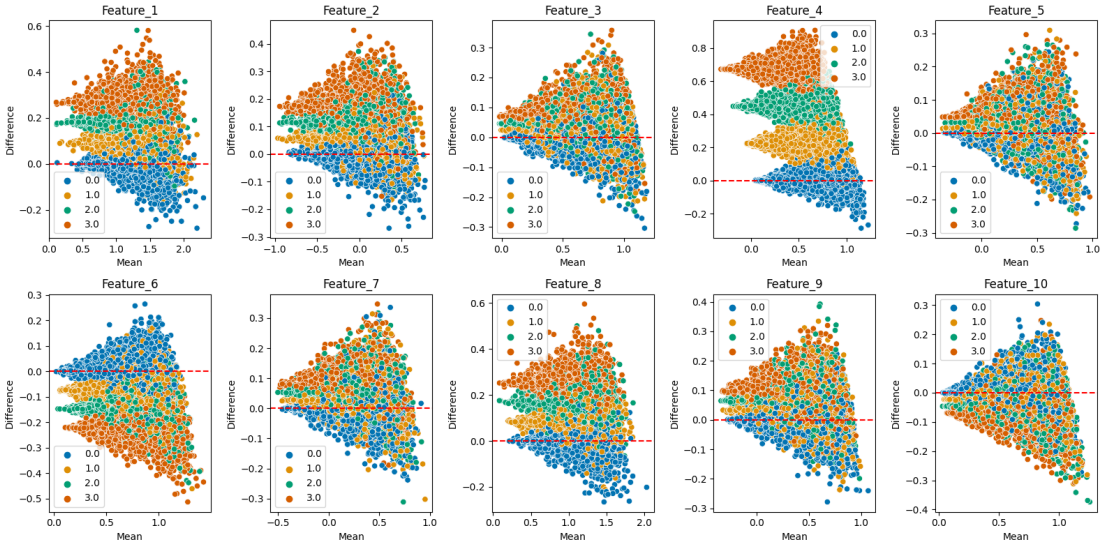
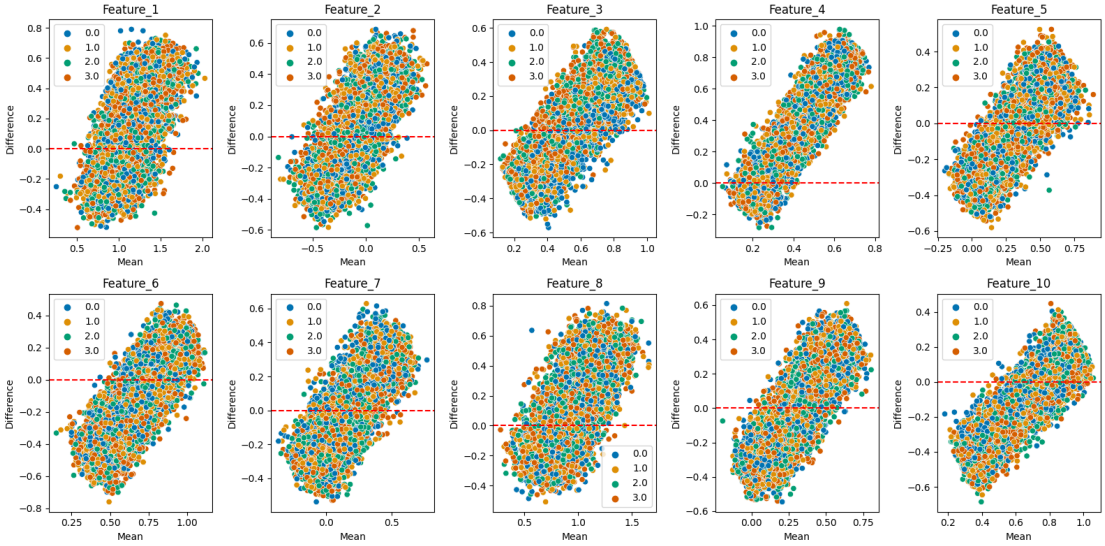


Fig. 3.4.: Distribution of features across sites: (top) Before harmonization, (middle) After harmonization, and (bottom) Unbiased distribution. The top plot illustrates the initial feature distribution, where distinct clusters corresponding to different sites are observed. In the middle plot, after applying the harmonization method, the feature distributions are aligned, reducing the variations between sites. The bottom plot represents the ideal scenario with an unbiased distribution, where batch effects have been completely eliminated.



(a) Bland-Altman plot comparing the feature differences between the unbiased dataset and the unharmonized dataset, stratified by batch effects.



(b) Bland-Altman plot comparing the feature differences between the unbiased dataset and the harmonized dataset using the CVAE method, stratified by batch effects.

Fig. 3.5.: Bland-Altman plots illustrating the feature differences between the unbiased dataset and the unharmonized dataset (a) and the harmonized dataset (b), stratified by batch effects.

3.5 Challenges and Future Directions

The application of CVAEs for data harmonization in the context of federated learning presents several challenges and offers avenues for future research. While our study demonstrates promising results using synthetic data, further exploration using real-world datasets is necessary to validate the effectiveness of CVAEs in capturing the complex variations present in heterogeneous neuroimaging data.

Furthermore, it is important to investigate the impact of different data characteristics and sources of heterogeneity on the performance of CVAEs in harmonization. This includes exploring the influence of imbalanced data distributions, missing data, and outliers, as well as understanding how different neuroimaging modalities and preprocessing techniques affect the harmonization process.

In addition to these technical challenges, ethical and privacy considerations are of utmost importance in federated learning approaches. It is crucial to establish robust privacy-preserving mechanisms, data governance frameworks, and regulatory guidelines to ensure the responsible and secure sharing of sensitive neuroimaging data across institutions while maintaining patient privacy and data protection.

3.6 Conclusions

In this chapter, we have explored the potential of Conditional Variational Autoencoders (CVAEs) for data harmonization in the context of federated learning for neuroimaging. By leveraging the power of CVAEs, we have demonstrated the ability to align and harmonize heterogeneous data from multiple institutions while preserving the underlying features and mitigating the impact of batch effects. Our results on synthetic data have shown promising outcomes, highlighting the potential of CVAEs as a valuable tool for data harmonization.

Data harmonization plays a critical role in biomedical research, enabling the integration and comparison of data from diverse sources. The application of CVAEs in this process offers several advantages. By utilizing the site indicator variable as a control mechanism, CVAEs can effectively capture the variations present in the latent space and align the distributions of the data points from different institutions. This leads to a more consistent representation of the underlying features, enhancing the reproducibility and generalizability of research findings.

While our preliminary exploration has provided insights into the use of CVAEs for data harmonization, there are important challenges and future directions to consider. Real-world validation using large-scale, diverse datasets is necessary to assess the performance and generalizability of CVAEs in capturing the complex variations present in neuroimaging data. Furthermore, addressing computational scalability, exploring the impact of different data characteristics, and addressing ethical and privacy concerns are crucial areas for further research.

In addition to its applications in phenotypical data, it is worth noting that the use of Conditional Variational Autoencoders (CVAEs) for data harmonization in federated

learning extends beyond numerical and tabular data. CVAEs can also be effectively applied to image data, leveraging the power of convolutional layers to capture complex spatial patterns and structures.

In conclusion, we explore a potential tool for data harmonization in federated learning, enabling the integration and analysis of multi-centric neuroimaging data while addressing challenges related to batch effects and data heterogeneity. The potential impact of this approach on healthcare outcomes and scientific discoveries is significant, and ongoing research and collaboration will drive further advancements in the field, leading to improved understanding, diagnosis, and treatment of neurological disorders.

4 Federated black-box Bayesian optimization

Abstract: One of the most critical steps in federated learning involves the iterative aggregation of models from multiple clients while optimizing the communication channel to reduce communication rounds and maintain generalization. In this work, we propose a black-box federated Bayesian optimization approach for model aggregation in federated learning, considering the learning of search space (bounds) for the parameters. Our method was evaluated on synthetic data and morphological measures from multi-centric databases for Alzheimer's disease. We assessed the applicability of Bayesian optimization as part of the optimization scheme in federated learning on synthetic data and on multicentric brain data from four different studies. Results show that despite the promising potential of Bayesian optimization, standard approaches such as FedAvg are more suitable and scalable than Bayesian optimization, especially in complex problems where the number of parameters increases.

4.1 Introduction

Federated learning (FL) has emerged as a field that empowers institutions to participate in multicentric studies while preserving data governance. This collaborative learning

paradigm allows multiple clients or data owners to share only the parameters of locally trained models with a central node, enabling larger sample access and encouraging consensus between clinical centers and researchers without exchanging or transferring data [Rieke et al., 2020].

Despite its potential, FL faces multiple challenges, including communication usage, hardware heterogeneity, and resource availability across institutions, which may prevent them from harnessing the power of their data. Additionally, ensuring the generalizability of models in heterogeneous setups, such as those found in collaborative studies, is crucial.

Existing federated optimization methods, like FedAvg [McMahan et al., 2017], FedProx [Li et al., 2018], and FedDANE [Li et al., 2019a], focus on aggregating locally partially trained models to increase generalizability. However, they often require multiple rounds of communication and depend on institutions having the computational power and resources to perform local gradient-based optimizations. This can be particularly challenging for institutions with limited resources or when facing hardware heterogeneity across different participants.

In contrast to the current federated optimization methods, centralized approaches like black-box optimization (BBO) and, more specifically, Bayesian optimization (BO) offer alternative solutions by building a response surface to optimize complex functions. BBO and BO optimize expensive-to-evaluate functions without relying on gradient descent, focusing on evaluating the target function instead. BO is particularly well-suited for tasks involving expensive-to-evaluate functions, as it offers a data-efficient and robust approach to global optimization.

Furthermore, BO provides convergence guarantees [Bull, 2011] and is especially suitable for problems with multiple minima due to its ability to construct a surrogate of the response surface. By leveraging the probabilistic representation of the objective function, BO not only accounts for the current best observation but also considers the uncertainty in the predictive distribution. Despite its use in centralized setups, BBO approaches have not been yet considered as an alternative for federated learning optimization.

In this work, we propose a federated optimization approach relying on black-box optimization that aims to centralize the optimization process in FL, reduce communication, and increase or be comparable in terms of generalizability by limiting participants of collaborative studies to only evaluate the objective function without requiring local gradient optimization. We approximate a response surface with the results collected from the institutions, leveraging BO's probabilistic representation of the objective function, which not only accounts for the current best observation but also considers the uncertainty in the predictive distribution.

Our proposed method delegates the majority of the computational load to the central server, resulting in a centralized optimization process that is not affected by variations in available computing resources across different institutions. This solution addresses the challenges of communication usage, hardware heterogeneity, and resource availability while also ensuring the generalizability of models in heterogeneous setups. By doing so, we hope to pave the way for future research in federated learning, ultimately advancing the field and enabling more effective utilization of decentralized data in healthcare.

4.2 Overview of Black Box Optimization

Black box optimization (BBO) is a powerful approach for optimizing complex functions without any knowledge of their internal workings, relying solely on their evaluation at different input points [Bajaj et al., 2021]. This technique is particularly useful when the function is computationally expensive to evaluate, has an unknown analytical form, or has many parameters that need to be tuned. Black box optimization has found numerous applications in machine learning, where it is used for hyperparameter tuning [Bergstra et al., 2012] and neural architecture search [Real et al., 2019]. It is also widely used in engineering and scientific fields for optimizing expensive simulations and experiments [Sobester et al., 2008].

In black box optimization, the goal is to find the optimal set of input values that minimize or maximize the function of interest. This can be formulated as an optimization problem of the form:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Omega}} f(\boldsymbol{\theta}), \tag{4.1}$$

where $\boldsymbol{\theta}$ is a vector of input values, $\boldsymbol{\Omega}$ is the set of feasible input values (search space), and $f(\boldsymbol{\theta})$ is the function to be optimized. The optimization problem in Equation 4.1 is often challenging to solve because the function $f(\boldsymbol{\theta})$ is typically complex, expensive to evaluate, and may have multiple local minima. Black box optimization methods aim to efficiently search the input space to find the global minimum or maximum of the function, without relying on any knowledge of its internal workings.

Several strategies exist for solving the optimization problem in Equation 4.1. For instance, Grid search exhaustively evaluates the function at all combinations of parameters in the search space, which is computationally expensive and may not scale well to high-dimensional search spaces. Random search samples parameters uniformly at random from the search space and evaluates the target function at these points, which is more

efficient than grid search but may still require a large number of function evaluations [Bergstra et al., 2012].

Another popular approach to black box optimization is Bayesian optimization (BO), which uses probabilistic models to guide the search towards promising regions of the search space. This approach is especially useful when the objective function may not only be complex but also has a limited number of evaluations (i.e. a communication budget).

4.2.1 Bayesian optimization

Bayesian optimization is an optimization technique for expensive black-box functions, where the objective function is treated as a random function and modeled using Gaussian Processes (GPs) [MacKay et al., 2003]. It is particularly useful when evaluating the function is time-consuming or costly, and the number of evaluations is limited. Now the target function is consider as:

$$f(\boldsymbol{\theta}) \sim \mathcal{GP}(\mu(\boldsymbol{\theta}), k(\boldsymbol{\theta}, \boldsymbol{\theta}')), \tag{4.2}$$

where $f(\boldsymbol{\theta})$ represents the surrogate model of the objective function, approximated by a GP with mean function $\mu(\boldsymbol{\theta})$ and covariance function $k(\boldsymbol{\theta}, \boldsymbol{\theta}')$. With its goal being to optimize the problem in Equation (4.1), the use of a GP as a model of the target function requires a guide for searching the optimal candidate parameters at each iteration. This guide is a function known as an "acquisition function".

Acquisition functions guide the search for the optimal input by balancing exploration and exploitation. They quantify the utility of evaluating the objective function at a particular input point, given the current model. Probability of Improvement (PI), Expected Improvement (EI), and Upper Confidence Bound (UCB) are some of the popular acquisition functions used in Bayesian optimization [Brochu et al., 2010]. These are described more in detail in Section 4.2.2.

The optimization process using BO takes place iteratively as the response surface (target function) is approximated by the surrogate model (GP) and new samples are proposed as follows:

1. Optimize the acquisition function over the GP to find the next sampling point in the search space: $\boldsymbol{\theta}^t = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Omega}} a(f(\boldsymbol{\theta}))$. Where a denotes the acquisition function.

2. Compute a noisy evaluation of the black-box function using the suggested set of input values from step 1, $\boldsymbol{\theta}^t$: $y^t = f(\boldsymbol{\theta}^t) + \epsilon_t$.

3. Update the GP with the new known sample from the search space and its measured response (e.g. loss value): $\boldsymbol{\Omega} = \boldsymbol{\Omega} \cup \{\boldsymbol{\theta}^t\}$ and $\mathcal{Y} = \mathcal{Y} \cup \{y^t\}$.

This is repeated until a stopping criterion is met, such as a maximum number of iterations or a convergence threshold. The optimal $\boldsymbol{\theta}^*$ will be finally defined as the set of input values that minimize the GP model.

4.2.2 Acquisition functions

The acquisition function serves as a guide to balance *exploration* and *exploitation* in BO. Its main objective is to guide the search for the optimum by directing it towards regions where the objective function is expected to be minimized/optimized. This can be accomplished either by selecting points with low predictions based on the probabilistic surrogate model or by targeting regions with high uncertainty in the model's predictions (or both).

Exploration refers to the act of seeking out and gathering information about new or unexplored regions of the search space. It involves trying out different options and sampling diverse areas to gain a broader understanding of the problem. The goal of exploration is to acquire new knowledge, uncover potential opportunities, or identify previously unknown solutions.

Exploitation, on the other hand, involves maximizing the current knowledge or exploiting known resources to achieve immediate gains. It focuses on utilizing the existing information, or exploiting well-established solutions to optimize the outcome of the objective function. The aim of exploitation is to capitalize on the available knowledge to obtain the best possible results based on the current understanding of the problem.

Among some of the most established acquisition functions there is:

Probability of Improvement (PI)

Probability of Improvement (PI) is an acquisition function that selects the next point for evaluation based on the probability that it will improve over the current best observation [Jones et al., 1998a]. Mathematically, PI can be defined as:

$$\text{PI}(\boldsymbol{\theta}) = P(f(\boldsymbol{\theta}) > f(\boldsymbol{\theta}^+) + \xi), \tag{4.3}$$

where $f(\boldsymbol{\theta})$ is the objective function, $\boldsymbol{\theta}^+$ is the current best observation, and ξ is an exploration parameter that encourages sampling in uncertain regions. The probability can be computed using the Gaussian process model's predictive distribution:

$$\text{PI}(\boldsymbol{\theta}) = \Phi\left(\frac{\mu(\boldsymbol{\theta}) - f(\boldsymbol{\theta}^+) - \xi}{\sigma(\boldsymbol{\theta})}\right), \tag{4.4}$$

where Φ is the cumulative distribution function of the standard normal distribution, $\mu(x)$ is the predictive mean, and $\sigma(\boldsymbol{\theta})$ is the predictive standard deviation at the input point x [Mockus, 1998].

Expected Improvement (EI)

Expected Improvement (EI) is another acquisition function that balances exploration and exploitation by considering the expected improvement over the current best observation [Mockus, 1998]. EI can be defined as:

$$\text{EI}(x) = E\left[\max(f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}^+), 0)\right], \tag{4.5}$$

where $f(\boldsymbol{\theta})$ is the objective function, and $\boldsymbol{\theta}^+$ is the current best observation. Using the Gaussian process model's predictive distribution, the expected improvement can be computed as:

$$\text{EI}(\boldsymbol{\theta}) = (\mu(\boldsymbol{\theta}) - f(\boldsymbol{\theta}^+) - \xi)\Phi(Z) + \sigma(\boldsymbol{\theta})\phi(Z), \tag{4.6}$$

where $Z = \frac{\mu(\boldsymbol{\theta}) - f(\boldsymbol{\theta}^+) - \xi}{\sigma(\boldsymbol{\theta})}$, Φ is the cumulative distribution function, and ϕ is the probability density function of the standard normal distribution. The parameter ξ encourages exploration, with larger values promoting greater exploration of the search space [Jones et al., 1998b].

Upper Confidence Bound (UCB)

Upper Confidence Bound (UCB) is an acquisition function that selects the next point for evaluation based on the upper confidence bound of the predictive distribution [Auer, 2002]. UCB can be defined as:

$$\text{UCB}(\boldsymbol{\theta}) = \mu(\boldsymbol{\theta}) + \kappa\sigma(\boldsymbol{\theta}), \tag{4.7}$$

where $\mu(\boldsymbol{\theta})$ is the predictive mean, $\sigma(\boldsymbol{\theta})$ is the predictive standard deviation at the input point $\boldsymbol{\theta}$, and κ is a parameter that controls the exploration-exploitation trade-off. A larger value of κ encourages more exploration.

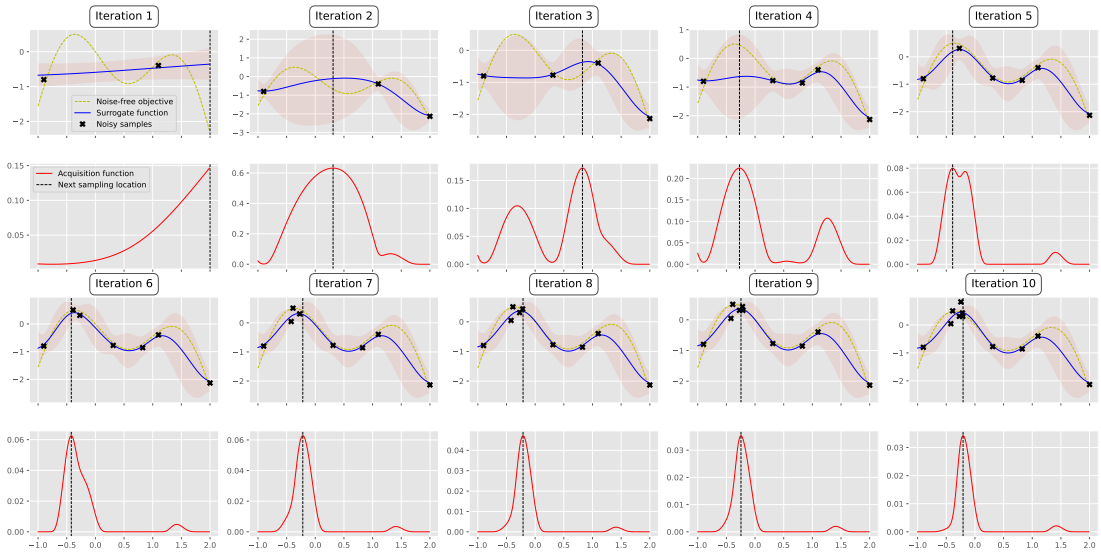


Fig. 4.1.: Bayesian optimization procedure with 10 iterations. Each row of two plots are result of an iteration. First and third row of plots depict the noise-free objective function alongside the surrogate function, represented by the Gaussian process posterior predictive mean. Additionally, the 95% confidence interval of the mean and the noisy samples obtained from the objective function are illustrated. The second and fourth row plots showcases the acquisition function, specifically the expected improvement. Notably, a vertical dashed line is included in both plots to indicate the proposed sampling point for the subsequent iteration, which corresponds to the maximum value of the acquisition function.

4.3 Methodology

In a federated scenario, we suppose there are K institutions participating in a federated learning experiment, each with a local dataset $\mathcal{D}_k$ of size N_k. The goal is to learn a global model $\boldsymbol{\theta}$ that minimizes the expected loss over all clients, subject to privacy

and communication constraints. This can be formulated as the following optimization problem:

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Omega}} f(\boldsymbol{\theta}) + \lambda\mathcal{R}(\boldsymbol{\theta}), \tag{4.8}$$

where,

$$f(\boldsymbol{\theta}) = \sum_{k=1}^{K} \frac{N_k}{N} F_k(\boldsymbol{\theta}), \tag{4.9}$$

$\boldsymbol{\Omega}$ is the set of feasible parameters for the model, $N = \sum_{k=1}^{K} N_k$ is the total number of data points, $F_k(w)$ is the local loss function for client k that measures the difference between the predicted output of the model $\boldsymbol{\theta}$ and the true output for a given input x, and $\mathcal{R}(\boldsymbol{\theta})$ is a regularization term that encourages the model to be simple or have certain desirable properties, and λ is a hyperparameter that controls the strength of the regularization.

The local average loss function value $F_k(\boldsymbol{\theta})$ for client k is computed as:

$$F_k(\boldsymbol{\theta}) = \frac{1}{N_k} \sum_{(x,y)\in D_k} \ell(\boldsymbol{\theta}; x, y)\,, \tag{4.10}$$

where, $\ell(\boldsymbol{\theta}; x, y)$ represents the specific loss function that is being optimized with respect to the model parameters $\boldsymbol{\theta}$ given the observed variables x and y.

We propose to solve this optimization problem by iteratively updating the global model $\boldsymbol{\theta}$ using local evaluations of the target function from each client without performing local updates as originally proposed by [McMahan et al., 2017]. Instead of aggregating the locally updated models at each step, a response surface is better approximated at each iteration by a GP as explained in the next section. The process is repeated until convergence or a stopping criterion is met.

In our approach, the regularization term $\mathcal{R}(\boldsymbol{\theta})$ serves a similar purpose to the exploration-exploitation trade-off terms mentioned earlier. It is a function of the acquisition function and helps balance the exploration of new regions of the input space with the exploitation of regions with high expected utility avoiding thus falling in local minima.

4.3.1 Federated Bayesian Optimization

In our approach, by considering each client's objective functions $F_k(\boldsymbol{\theta})$ as realizations of a Gaussian process as shown in Equation (4.11), then minimizing the GP with respect to the parameters of the model is equivalent to finding the best operating parameters $\boldsymbol{\theta}$ of the model in question. In order to accelerate convergence we conveniently consider that $\boldsymbol{\theta}$ lives in a subspace or parameters denominated the *search space* $\boldsymbol{\Omega}$.

$$F_k(\boldsymbol{\theta}) \sim \mathcal{GP}(\mu(\boldsymbol{\theta}), k(\boldsymbol{\theta}, \boldsymbol{\theta}')) \implies \min_{\boldsymbol{\theta}\in\boldsymbol{\Omega}} f(\boldsymbol{\theta}) \propto \min_{\boldsymbol{\theta}\in\boldsymbol{\Omega}} \mathcal{GP}(\mu(\boldsymbol{\theta}), k(\boldsymbol{\theta}, \boldsymbol{\theta}')) \tag{4.11}$$

Search space learning:
Accelerating the convergence of Bayesian optimization can be achieved by defining appropriate bounds for the search space. This, in our federated BO proposal, would also be translated as a more optimal utilization of the communication channel. To reach this goal, we propose to introduce adaptive learnable bounds as proposed in Perrone et al.'s work [Perrone et al., 2019]. Perrone's method provides a framework for learning the search space geometry from historical data, allowing to focus the optimization process on relevant regions.

In Perrone's work, they address the challenge of optimizing black-box functions by automatically designing the search space based on evaluations of previous functions. By departing from the common practice of defining arbitrary search ranges a priori, they introduce a methodology to learn search space geometries from previous ones. This approach endows Bayesian optimization methods with transfer learning properties and significantly improves the optimization process.

The key idea is to minimize the volume of the search space while ensuring that all the solutions resulting from maximizing the acquisition function are contained within the learned bounds. Perrone's method formulates the problem as minimizing the search space volume $\mathcal{Q}$, represented by the subspace bounds $\boldsymbol{\Omega}$, as shown in Equation 4.12.

$$\min_{\boldsymbol{\Omega}\in\mathbb{R}^d} \mathcal{Q}(\boldsymbol{\Omega}) \text{ such that for } t \geq 1\,, \boldsymbol{\theta}^+ \in \boldsymbol{\Omega} \tag{4.12}$$

In Equation 4.12, the variable t represents the index of the previously optimized black-box functions. It indicates that the search space should accommodate all the solutions obtained from maximizing the acquisition function for each of the functions indexed from 1 to t.

As a first instantiation, Perrone et al. consider the bounding space to be a hyperrectangle, where the search space can be described by its lower and upper bounds $\mathbf{\Omega} = (\mathbf{l}, \mathbf{u})$. The problem can then be rewritten as follows:

$$\min_{\mathbf{l}\in\mathbb{R}^d, \mathbf{u}\in\mathbb{R}^d} \frac{1}{2}||\mathbf{u}-\mathbf{l}||_2^2 \quad \text{such that,} \quad \mathbf{l} \leq \boldsymbol{\theta}^+ \leq \mathbf{u} \tag{4.13}$$

$$\tag{4.14}$$

Hence, the optimal lower and upper bounds are obtained as $\mathbf{l}^* = \min\{\boldsymbol{\theta}^+\}_{t=1}^T$ and $\mathbf{u}^* = \max\{\boldsymbol{\theta}^+\}_{t=1}^T$, representing the smallest hyperrectangle containing all the previously optimized solutions.

Optimization of the target function:
We rely on BO and the probabilistic nature of GPs to estimate not only the mean but also the uncertainty of a function. We choose Expected Improvement (EI) as the acquisition function, as it is considered less prone to falling into local minima. The fundamental idea behind EI is to maximize the distance between the current point and the candidate point, while balancing the trade-off between exploitation and exploration. EI focuses on not only identifying candidates at the optima of $\mu(\boldsymbol{\theta})$ but also penalizing the uncertainty $\sigma^2(\boldsymbol{\theta}) = k(\boldsymbol{\theta},\boldsymbol{\theta}) - \mathrm{k}\left(\boldsymbol{\theta}, \{\boldsymbol{\theta}^+\}_{t=1}^T\right)\left[\mathrm{K}\left(\{\boldsymbol{\theta}^+\}_{t=1}^T, \{\boldsymbol{\theta}^+\}_{t=1}^T\right) + \lambda^2 I\right]^{-1} \mathrm{k}\left(\{\boldsymbol{\theta}^+\}_{t=1}^T, \boldsymbol{\theta}\right)$ [Pardalos et al., 2021]. This balancing act enables a trade-off between exploitation (finding the minimum) and exploration (avoiding local optima). EI, in particular, evaluates the expected positive distance between the current point and the solution $\boldsymbol{\theta}^+$, offering an effective approach to guide the optimization process.

$$\mathrm{EI}(\boldsymbol{\theta}) = \mathbb{E}\left[\max(f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}^+), 0)\right] \tag{4.15}$$

$$\mathrm{EI}(\boldsymbol{\theta}) = \begin{cases} (\mu(\boldsymbol{\theta}) - f(\boldsymbol{\theta}^+) - \xi)\Phi(Z) + \sigma(\boldsymbol{\theta})\phi(Z) & \text{if } \sigma(w) > 0 \\ 0 & \text{if } \sigma(\boldsymbol{\theta}) = 0 \end{cases} \tag{4.16}$$

$$Z = \begin{cases} \frac{\mu(\boldsymbol{\theta}) - f(\boldsymbol{\theta}^+) - \xi}{\sigma(\boldsymbol{\theta})} & \text{if } \sigma(\boldsymbol{\theta}) > 0 \\ 0 & \text{if } \sigma(\boldsymbol{\theta}) = 0 \end{cases} \tag{4.17}$$

Where, $\mu(\boldsymbol{\theta})$ and $\sigma(\boldsymbol{\theta})$ correspond to the posterior prediction from the Gaussian process ($\mathcal{GP}$). Φ and ϕ represent the cumulative distribution function (CDF) and probability density function (PDF), respectively. The parameter ξ is a penalization term that balances exploration and exploitation in the parameter space.

In this context, each solution $\boldsymbol{\theta}^+$ serves as an exploration point within the search space, aiming to gather as much knowledge as possible from the response function f in Equation 4.9. Unlike the aggregated model, these solutions focus on exploring regions with high uncertainty (large $\sigma(\boldsymbol{\theta})$) rather than solely advancing in the direction with the steepest gradient.

Algorithm 4: Federated Bayesian optimization

Require: Define a model $\mathcal{M}_{\boldsymbol{\theta}}$, parametrized by $\boldsymbol{\theta}$, and an objective function $f(\boldsymbol{\theta}) \sim \mathcal{GP}(\mu(\boldsymbol{\theta}), k(\boldsymbol{\theta}, \boldsymbol{\theta}'))$. N_R rounds and K clients.
Ensure: Initialize set of parameters: $\boldsymbol{\theta}^+ \sim \text{Uniform}(\mathbf{l}, \mathbf{u})$.
  while $n_i \leq N_R$ **do**
    for $k \in \{1, 2, ..., K\}$ **do**
      Evaluate $F_k(\boldsymbol{\theta}^+)$
    end for
    Compute global loss $y^t \leftarrow \sum_{k=1}^{K} \frac{N_k}{N} F_k(\boldsymbol{\theta}^+)$
    Update search space $\boldsymbol{\Omega} \leftarrow \boldsymbol{\Omega} \cup \boldsymbol{\theta}^+, \mathbf{y} \leftarrow \mathbf{y} \cup y^t$
    Fit response surface $f(\boldsymbol{\theta}) \leftarrow \mathcal{GP}(\cdot, \cdot)|(\mathbf{y}, \boldsymbol{\Omega})$
    Obtain $\boldsymbol{\theta}^+ = \max_{\boldsymbol{\theta}} \text{EI}(\boldsymbol{\theta})$, where $\text{EI}(\boldsymbol{\theta})$ is the Expected Improvement acquisition function.
  end while

The aggregation scheme, which combines Bayesian optimization (BO) and the learning bounds strategy proposed by Perrone et al. [Perrone et al., 2019], is detailed in Algorithm 4. This algorithm iteratively optimizes the model by updating the search space and fitting a Gaussian process response surface to guide the exploration. The optimal model is selected based on the best-performing set of parameters $\boldsymbol{\theta}^+$ obtained during the iterations.

4.4 Results

4.4.1 Synthetic Data

Synthetic data was generated for two tasks: regression, and classification. We specifically chose these two tasks as these problems are simple and convex in the case of linear regression, allowing for a clear comparison between different approaches. These tasks also provide a baseline for evaluating the scalability and performance of our proposed federated Bayesian optimization method compared to FedAvg. By focusing on these simple problems, we can isolate the impact of dimensionality and assess the scalability of the two approaches more effectively.

Impact of Dimensionality on Scalability

The synthetic data experiments involved testing for five different levels of complexity,

corresponding to different numbers of parameters: 1, 3, 10, 30, and 100. To simulate a more realistic scenario, the data was slightly biased per site with a random shift. This test of dimensionality allows us to compare the scalability of FedAvg and our proposed Bayesian optimization method.

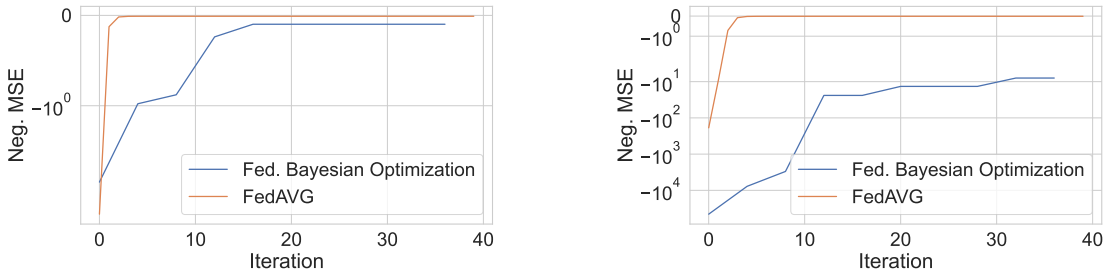
The Figure 4.3 illustrates the scalability of FedAvg and our proposed approach across different numbers of parameters. The left panel displays the cosine distance, which measures the angular dissimilarity between the groundtruth or target parameters (coefficients) optimized by each approach. The middle panel represents the Euclidean distance, capturing the overall difference between the optimized and true parameters.

The results reveal that our proposed approach falls short compared to FedAvg in terms of performance across all metrics. Both the cosine distance and Euclidean distance show higher values for our approach, indicating a larger discrepancy from the true parameters. Additionally, the execution time for our approach tends to be comparable to or even longer than that of FedAvg, indicating poorer efficiency.

These observations highlight the challenges and limitations of our proposed approach in achieving accurate parameter optimization and efficient execution. It is evident that FedAvg outperforms our method in terms of parameter approximation and execution time. These results underscore the need for further investigation and refinement of our approach to address these shortcomings and enhance its performance in the context of federated learning optimization.

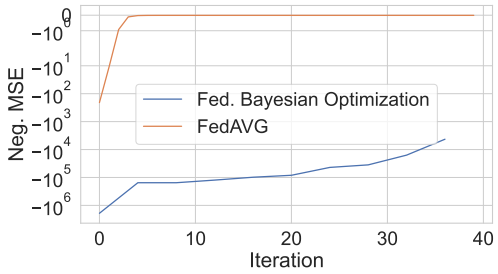
As shown in Figure 4.2, we illustrate the scalability of FedAvg and our proposed approach for linear regression with varying numbers of parameters, it is evident that FedAvg scales better as the number of parameters increases. This finding suggests that FedAvg may be more suitable for handling larger and more complex models in federated learning scenarios. Figure 4.3 illustrates the scalability of both approaches with respect to the groundtruth set of parameters and the execution time as the number of parameters increases logarithmically.

Additionally, we explored a combination of partially locally optimized parameters using stochastic gradient descent (SGD) as an additional exploration strategy in conjunction with our proposed Bayesian optimization approach. This strategy involved applying SGD for one epoch to guide the search of the parameter space towards potentially promising regions. An example of this combination is shown in Figure 4.4, illustrating the evolution of the R^2 metric for regression in a convex problem on synthetic data and brain data. The results demonstrate that the initial guided exploration by gradient descent can lead to improved convergence, resulting in better optimization performance. However, it is worth noting that the optimization progress plateaus as the exploration continues, despite using a low balance between exploration and exploitation ($\xi = 0.01$). This suggests that

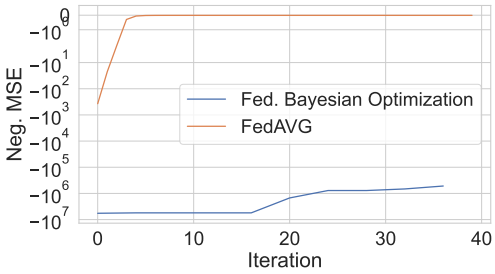


(a) 1 parameter

(b) 3 parameters



(c) 10 parameters



(d) 30 parameters

Fig. 4.2.: Negative mean squared error over iterations for different numbers of parameters on synthetic data. Model optimized: Linear regression. The x-axis represents the number of iterations or rounds of communication, while the y-axis corresponds to the negative value of the mean squared error (MSE), which serves as a surrogate of the likelihood (function to be maximized).

FedAvg techniques are better at achieving further improvement. Therefore, our proposed approach may not be as effective in this context. Same plateau observed in the previous experiments applies for the classification task on synthetic data, as shown in Figure 4.5 (left panel). The figure displays the evolution of classification metrics, specifically accuracy and F1 score, across rounds of communication.

In spite of the results, we explored our method on a more complex and widely used model in neuroimaging, the Variational Autoencoder (VAE). The VAE has gained significant popularity in the field of neuroimaging due to its ability to capture complex data distributions and generate meaningful representations. By leveraging its generative framework, the VAE can learn latent representations that capture important features and variations in the brain imaging data.

Figure 4.6 presents the results of our method applied to the VAE model. The figure illustrates the evolution of the reconstruction loss, which serves as an important metric for evaluating the quality of the VAE's generated outputs. Despite our best efforts, the optimization progress for the VAE model did not reach satisfactory levels, indicating the challenges associated with optimizing complex models in the federated learning setting.

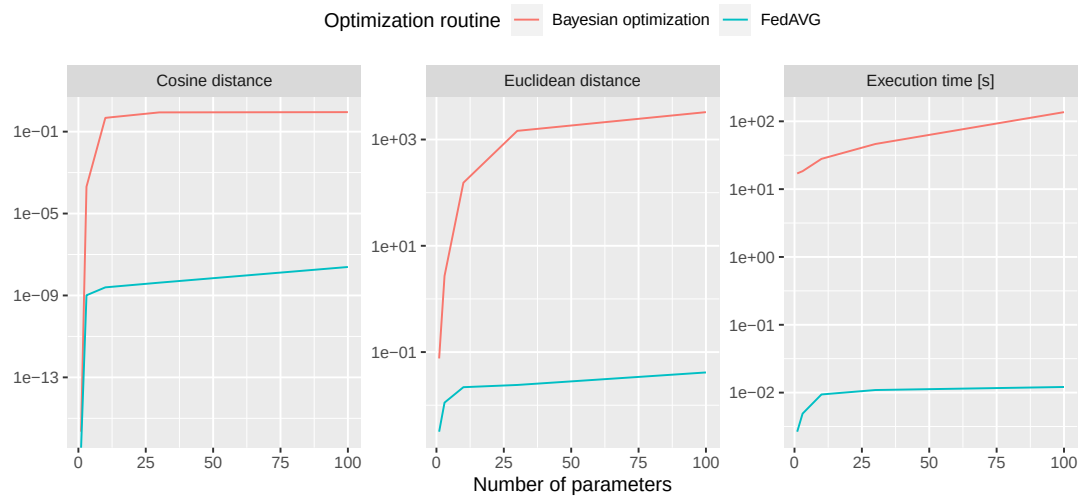


Fig. 4.3.: Comparison of scalability between FedAvg and our proposed approach. The left panel shows the cosine distance between the optimized and true parameters (lower values indicate better accuracy), the middle panel displays the Euclidean distance (lower values indicate better convergence), and the right panel presents the execution time in seconds.

4.4.2 Brain imaging data

A total of 87 measures, including subcortical volumes and cortical thicknesses derived from T1-weighted magnetic resonance images (T1w-MRI) were extracted using FreeSurfer 6.0. To create a realistic multi-center scenario with non-independent and identically distributed (non-iid) data, different datasets were incorporated. These datasets consisted of 801 baseline participants from the Alzheimer's Disease Neuroimaging Initiative (ADNI), 401 participants from the OASIS-I cross-sectional dataset, 1,117 participants from OASIS-III, and 67 participants from the Minimal Interval Resonance Imaging in Alzheimer's Disease (MIRIAD) for a total of 2369 participants.

To assess the proposed method, a binary classification task was defined, categorizing the participants into controls (NC) and cognitively impaired (CI) groups. This categorization allows for the identification of potential differences between individuals with normal cognitive functioning and those with cognitive impairments. A logistic classifier was employed for the classification task, which is a widely used algorithm for binary classification and as in the experiments with synthetic data, suitable for comparison. Classification metrics across iterations/rounds are shown in Figure 4.5.

For a more comprehensive understanding of the study's population demographics, please refer to Table 4.1.

Equivalently, the right panel of Figure 4.6 presents the results on brain imaging data using a variational autoencoder (VAE). The reconstruction errors (MSE and MAE) across

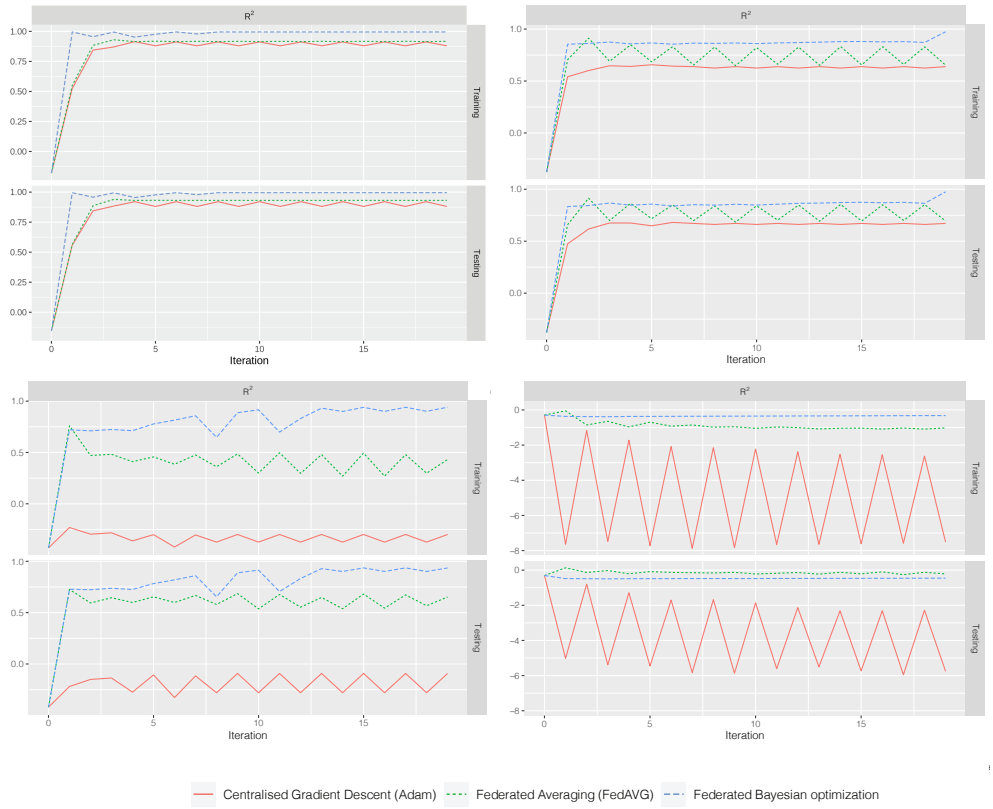


Fig. 4.4.: Evolution of R^2 for regression in a convex problem on synthetic data across different scenarios varying the number of parameters to optimize. FedAvg settings: 10 epochs locally. Each iteration corresponds to a round of communication/aggregation. Parameters being optimized: 30 (top left), 100 (top right), 300 (bottom left) and 1000 (bottom right).

communication rounds serve as a measure of the quality of the reconstructed phenotypes are illustrated. Similar to the previous experiments, the results demonstrate the presence of a plateau effect in the optimization process.

4.5 Challenges and Future Directions

The investigation into the use of Bayesian optimization in federated learning has provided important insights and lessons learned. While our approach holds potential for hyperparameter tuning and optimization in the federated learning setting, several challenges and avenues for future research and development should be considered.

Firstly, our experiments have revealed the scalability challenge associated with the number of parameters in a model. The optimization process becomes more complex and time-consuming as the number of parameters increases, even when adaptive bounds are applied. This issue can limit the applicability of our proposed method to larger and more complex models. Future research should focus on developing techniques or

		Controls (NC)	Cognitively Impaired (CI)
ADNI (N=801)	**N. Participants**	175	621
	Age at study entry (years ± std)	75.79 ± 4.99	74.86 ± 7.23
	Male proportion (%)	52%	58%
OASIS-I (N=401)	**N. Participants**	314	87
	Age at study entry (years ± std)	54.22 ± 26.20	78.86 ± 11.12
	Male proportion (%)	37%	41%
OASIS-III (N=1098)	**N. Participants**	709	306
	Age at study entry (years ± std)	67.8 ± 9.86	74.28 ± 7.87
	Male proportion (%)	41%	53%
MIRIAD (N=69)	**N. Participants**	23	46
	Age at study entry (years ± std)	69.7 ± 7.2	69.4 ± 7.1
	Male proportion (%)	52%	41%

Tab. 4.1.: Demographic characteristics of the patients and datasets included in the study.

optimizations to overcome this scalability limitation and enable efficient optimization in high-dimensional parameter spaces.

Secondly, the plateau effect observed in the optimization progress suggests that further improvement may require a more refined balance between exploration and exploitation in the parameter space. Exploring alternative exploration strategies or incorporating adaptive mechanisms to dynamically adjust the balance between exploration and exploitation could be worthwhile avenues for future investigation. Additionally, investigating the impact of different acquisition functions or surrogate models in Bayesian optimization within the federated learning context could lead to improved optimization performance.

Furthermore, the work of Dai et al. has introduced the concept of differentially private federated Bayesian optimization with distributed exploration [Dai et al., 2021]. Their approach integrates differential privacy into the federated Thompson sampling algorithm to preserve user-level privacy while improving utility through distributed exploration. This suggests that privacy-preserving optimization techniques can be further explored and incorporated into the federated learning context to enhance the privacy guarantees of our proposed Bayesian optimization method.

Additionally, the study conducted by Holly et al. on evaluating hyperparameter optimization approaches in an industrial federated learning system highlights the advantage of a

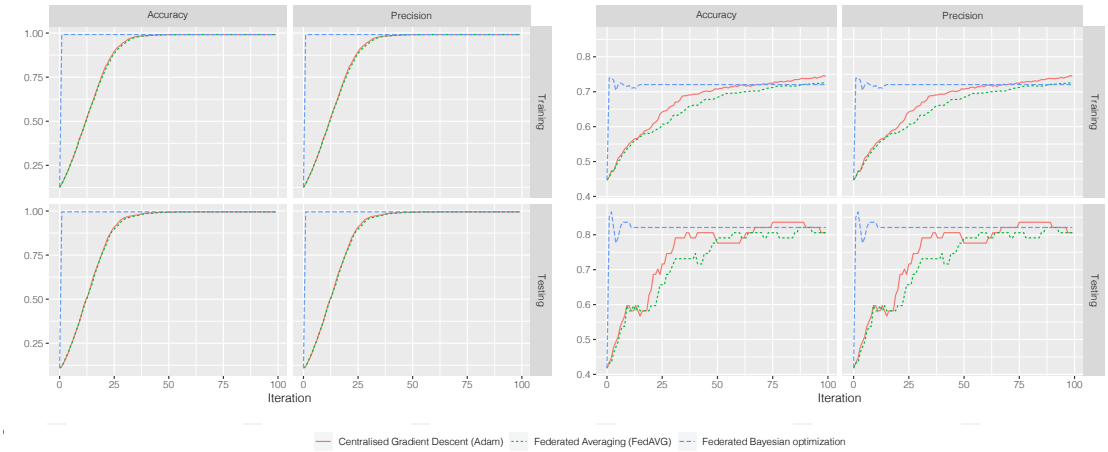


Fig. 4.5.: Evolution of classification metrics across rounds of communications (iterations). The left panel presents the results obtained on Synthetic data (iid distributed), where classification was performed on 4 classes. The right panel displays the results for brain data using MIRIAD as the testing set, addressing a control-case problem (NC vs CI). FedAvg settings included 10 locally executed epochs. Additionally, a centralized method is shown using Adam. To better estimate the response surface, a first sampling step and a single epoch of optimization were allowed in each site as a warm up step for the Bayesian optimization.

global optimization approach, specifically using a grid search algorithm, in improving the performance of federated models [Holly et al., 2022]. In their approach, the global optimization is achieved by iteratively performing the FedAvg algorithm with different learning rates on the clients, evaluating the model's performance on validation data, and selecting the learning rate that yields the highest average accuracy across clients. These findings suggest that a careful consideration of global optimization strategies, such as grid search, can yield better results in hyperparameter optimization within the federated learning context. Further research can explore the behavior and performance of different hyperparameter configurations and optimization approaches, taking into account a wider range of hyperparameters and their effects on federated learning models.

In conclusion, while our investigation has shed light on the potential of Bayesian optimization in the federated learning setting, challenges remain in terms of scalability and optimization performance. Future research should focus on addressing these challenges and further refining the proposed approach to make it more efficient, scalable, and applicable to a wider range of models. Additionally, exploring privacy-preserving optimization techniques and incorporating insights from related works can help enhance the privacy and utility trade-off in federated learning optimization. By overcoming these challenges, we may unlock the full potential of Bayesian optimization in federated learning scenarios.

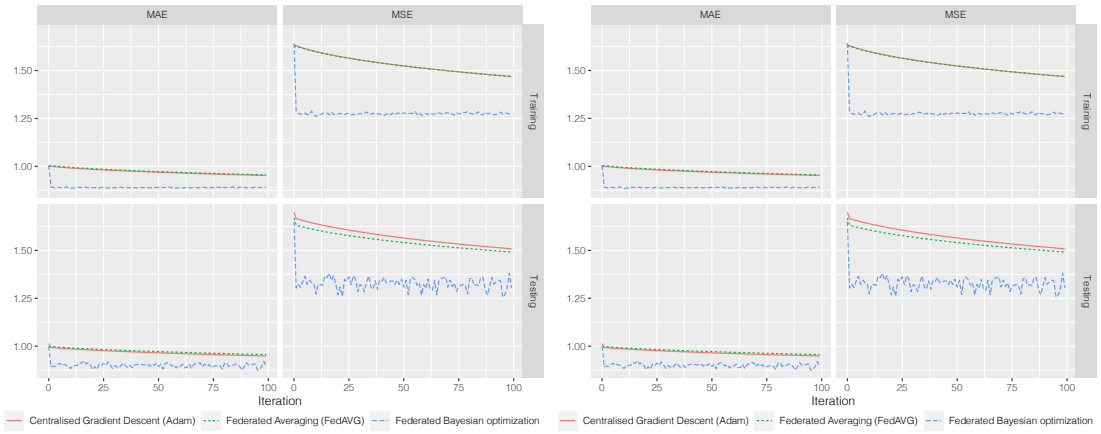


Fig. 4.6.: The left panel depicts the results on synthetic data with 4 simulated classes, while the right panel displays the results for brain data. The brain data was segregated using two classes: NC (controls) and CI (cognitively impaired). Federated Averaging (FedAvg) settings consisted of 10 locally executed epochs. Furthermore, the caption includes a comparison with the centralized approach using Adam as the optimizer. To better estimate the response surface, a first sampling step and a single epoch of optimization were allowed in each site as a warm up step for the Bayesian optimization.

4.6 Conclusions

We investigated in this chapter the use of Bayesian optimization as an optimization scheme for federated learning. While it initially showed promise in centralizing the optimization process and potentially improving the generalizability of models, scalability remains a challenge as the number of parameters in a model impacts the optimization process, even when adaptive bounds are applied. One potential solution to address this issue is to guide the initial search using gradient descent. However, further exploration is needed, as our findings indicate that the model's performance does not improve substantially after the initial search. In conclusion, Bayesian optimization offers a promising direction for federated learning, but additional research is required to address the scalability challenges and fully harness its potential in this context.

5 Fed-BioMed: A General Open-source Frontend Framework for Federated Learning in Healthcare

Abstract: As the quantity of data generated by the healthcare industry continues to grow, an array of challenges is jointly presented, particularly in terms of data access, transfer, and privacy. Federated learning emerges as a solution to these challenges, providing privacy-aware approaches to data analysis through decentralized optimization methods that maintain data distribution siloed. However, in the

field of neuroimaging, there is a lack of common frameworks that standardize or allow a seamless workflow for medical studies. Besides, existing federated learning frameworks are often constrained to specific hardware, modeling approaches or applications, and lack the innate capacity to provide a production-ready environment for deployment.

To address this gap, we proposed in 2020 Fed-BioMed, an open-source federated learning frontend framework specifically designed for healthcare applications. Our contribution lies in the development of an infrastructure that not only ensures is compliant with privacy-preserving policies such as the CCPA and the GDPR and inspired by industrial guidelines as the IEEE 3652.1-2020 [, 2021], but also brings together a versatile architecture that is model agnostic and is flexible to different optimization techniques.

This chapter outlines the pioneering work that forms the foundation of the Fed-BioMed platform. The presented work encapsulates the essential motivations and design principles that have driven the evolution of Fed-BioMed into a comprehensive collaborative development project. We explore the original 2020 software components, their deployment, and the integration of learning models into the federated learning system. Components and functionalities that are still current as to the date of this publication. We presented this work at the Workshop on Distributed And Collaborative Learning (DCL) in MICCAI 2020 and published in Springer (LNCS), this work also showcases one of the first real-world applications of federated learning in the federated analysis of multicentric brain imaging data. It is noteworthy that what makes up the Fed-BioMed framework today largely stems from this initial foundation.

5.1 Context

Fed-BioMed is an open-source initiative, dedicated to the research and development of federated learning for practical medical applications. This initiative was conceived after our work on Federated Principal Component Analysis (fPCA) [Silva et al., 2019], after realizing that the healthcare sector lacked a comprehensive framework for federated learning applications.

This chapter presents how we first filled such gap, the design choices, creation of Fed-BioMed, and a first real-life scenario benchmark using neuroimaging data. This core was initially introduced at the 2nd edition of Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning MICCAI Workshop in 2020 [Silva et al., 2020]. The presentation provided a comprehensive overview of the fundamental

aspects of Fed-BioMed, detailing its purpose, structure, and functionalities. It is worth pointing out that the Fed-BioMed framework, as it stands today, is deeply rooted in the foundational work presented in this chapter. The principles, design elements, and functionalities we outlined remain at the core of the framework, showing resilience and enduring nature in the midst of technological advancements.

Fed-BioMed, which is now continuously maintained and improved by a team of engineers at the Service d'Expérimentation et de Développement (SED) in the Centre INRIA d'Université Côte d'Azur. The team works closely with stakeholders to gather feedback, ensuring that the framework stays relevant and effective in meeting the needs of users. Fed-BioMed has evolved since our initial presentation, with new updates and functionalities continuously being added to enhance its capabilities. For the most current details on these updates and new features, we refer to the work of [Cremonesi et al., 2023] and the official website (`https://fedbiomed.gitlabpages.inria.fr/`), in which an up-to-date discussion on the state of the initiative is provided.

5.2 Introduction

The private and sensitive nature of healthcare information often hampers the use of analysis methods relying on the availability of data in a centralized location. Decentralized learning approaches, such as federated learning, represent today a key working paradigm to empower research while keeping data secure[Li et al., 2020a].

Initially conceived for mobile applications [Konečnỳ et al., 2016], federated learning allows to optimize machine learning models on datasets that are distributed across clients, such as multiple clinical centers in healthcare [Brisimi et al., 2018]. Two main actors play in the federated learning scenario: *clients*, represented for instance by clinical centers, and a *central node* that continuously communicate with the clients [Yang et al., 2019].

Federated learning methods must address three main issues: *security*, by preventing data leakages and respecting privacy policies such as the EU general data protection regulation (GDPR) [Voigt et al., 2017], *communication efficiency*, by optimizing the rounds of communication between clients and the central node, and *heterogeneity robustness*, by properly combining models while avoiding biases from the clients or adversarial attacks aiming to sabotage the models [Bhagoji et al., 2019; Bagdasaryan et al., 2020]. These issues are currently tackled through the definition of novel federated analysis paradigms, and by providing formal guarantees for the associated theoretical properties [Li et al., 2018; Li et al., 2019a].

Besides the vigorous research activity around the theory of federated learning, applications of the federated paradigm to healthcare are emerging [Gupta et al., 2018; Kim et al., 2017; Lee et al., 2018]. Nevertheless, the translation of federated learning to real-life scenarios still requires to face several challenges. Besides the bureaucratic burden, for federation still requires to establish formal collaboration agreements across partners, the implementation of a federated analysis framework requires to face important technical issues, among which the problem of data harmonization, and the setup of software infrastructures. In particular, from the software standpoint, the practical implementation of federated learning requires the availability of frontend frameworks that can adapt to general modeling approaches and application scenarios, providing researchers with a starting point overcoming problems of deployment, scalability and communication over the internet.

5.3 Goals and Contributions of Fed-BioMed

In this work we propose Fed-BioMed, an open-source production-ready framework for federated learning in healthcare. Fed-BioMed is Python-based and provides modules to deploy general models and optimization methods within federated analysis infrastructures. Besides enabling standard federated learning aggregation schemes, such as federated averaging (FedAvg) [Konečnỳ et al., 2016; McMahan et al., 2017], Fed-BioMed allows the integration of new models and optimization approaches. It is also designed to enable the integration with currently available federated learning frameworks, while guaranteeing secure protocols for broadcasting.

We expect this framework to foster the application of federated learning to real-life analysis scenarios, easing the process of data access, and opening the door to the deployment of new approaches for modeling and federated optimization in healthcare. The code will be freely accessible from our repository page (`https://gitlab.inria.fr/fedbiomed`).

5.4 Related works

NVIDIA Clara is a large initiative focusing on the deployment of federated learning in healthcare [Yuhong et al., 2019], currently providing a service where users can deploy personalized models. The code of the project is not open, and it requires the use of specific hardware components. This may reduce the applicability of federated learning to general use-cases, where client's facilities may face restrictions in the use of proprietary technology.

The Collaborative Informatics and Neuroimaging Suite Toolkit for Anonymous Computation (COINSTAC) [Plis et al., 2016] focuses on single-shot and iterative optimization schemes for decentralized multivariate models, including regression and latent variable analyses (e.g. principal/independent-component analysis, PCA-ICA, or canonical correlation analysis, CCA). This project essentially relies on distributed gradient-based optimization as aggregating paradigm. A frontend distributed analysis framework for multivariate data modeling was also proposed by [Silva et al., 2019]. Similarly to COINSTAC, the framework focuses on the federation of multivariate analysis and dimensionality reduction methods. The `PySyft` initiative [Ryffel et al., 2018] provides an open-source wrapper enabling federated learning as well as encryption and differential privacy. At this moment however this framework focuses essentially on optimization schemes based on stochastic gradient descent, and does not provide natively a deployable production-ready framework. Fed-BioMed is complementary to this initiative and allows interoperability, for example by enabling unit testing based on `PySift` modules.

5.5 Implementation Overview and Architecture of Fed-BioMed

Fed-BioMed is designed as a microservices-based architecture, organized according to the different actors in Federated Learning (FL): (i) the **clients (nodes)**, responsible for managing the data they want to allow modeling on, (ii) a server or **central node (network) service**, in charge of managing communication and model parameter sharing, and (iii) the **researcher**, whose model or optimization method is to be deployed and tested. The architecture is inspired by the 2018 "Guide for Architectural Framework and Application of Federated Machine Learning" (Federated Learning infrastructure and Application standard)[1], basing it on a server-client paradigm with three types of instances: central node, clients, and federators.

5.5.1 Client (Node) Service

Clients, nodes or institutes are responsible for storing the datasets through a dedicated client application. Since not every client is required to store the same type of data, this enables studies with heterogeneous or missing features across participants. Moreover, depending on the target data source and on the analysis paradigm, centers can be either included or ignored from the study. Each client verifies the number of current jobs in queue with the central node, as well as data types and models associated to the running jobs.

[1] `https://standards.ieee.org/project/3652_1.html`

Figure 5.2 shows Fed-BioMed's user interface which allows institutions to configure multiple types of data for modeling consent, such as plain-text based data in the form of CSV or TSV files, default datasets from third-party repositories, or image data. Once configured, a client can start and stop the service whenever considered appropriate. Clients receive commands from researchers that trigger defined events like: ping to check node availability, local training of a specific model, and data description requests (e.g., sample sizes to account for size effects).

5.5.2 Central Node (Network) Service

The central node or network service serves as a trusted broker between researchers and participating clients or institutions. It manages communication channels, model updates, and serves as an ephemeral storage service for model aggregation, ensuring secure and efficient collaboration between all parties involved.

In order to maintain an optimized communication channel, event triggering, messaging, and model parameter sharing are managed by two services. Both commands and data flow can coexist without hindering each other. Messaging is managed through MQTT, a standard industrial protocol for efficient messaging communication. Additionally, model flow is managed using HTTP requests via a REST API developed in Django as the backend framework. This architecture consequently delegates parameters' data flow to the central node instead of being the researcher's responsibility.

These services are also provided within Fed-BioMed in the form of Docker containers, so they can be scaled horizontally. The central node's role in a federated learning setup ensures secure and efficient communication between all parties involved in the study.

Furthermore, the central node is responsible for maintaining a job queue, which keeps track of the current jobs in progress and their associated models and data. This allows clients to verify the number of current jobs in the queue, as well as data types and models related to the running jobs.

5.5.3 Researcher Service

The researcher service in Fed-BioMed provides a comprehensive and efficient platform for researchers to develop, deploy, and manage their models in a federated learning environment. By providing support for various aggregation methods, model development, and orchestration of the training process, Fed-BioMed enables researchers to effectively address data heterogeneity, ensure robust model performance across diverse nodes, and maintain control over the entire federated learning process.

An `Experiment` class is included in the package and it is responsible for orchestrating the training process on available nodes. Such orchestration involves:

- Searching for datasets on active nodes based on specific tags provided by a researcher and used by the nodes to identify the dataset.
- Uploading the training plan file created by the researcher and sending the file URL to the nodes.
- Sending model and training arguments to the nodes.
- Tracking the training process in the nodes during all training rounds.
- Verifying the nodes' responses to ensure that each round is successfully completed in every node.
- Downloading the local model parameters after every round of training.
- Aggregating the local model parameters based on the specified federated approach, and eventually sending the aggregated parameters to the selected nodes for the next round.

As of the publication of this work, Fed-BioMed supports several aggregation methods to tackle the challenges posed by heterogeneous data in federated learning. The first method, FedAvg, is a widely-used standard aggregation scheme in federated learning that performs federated averaging. Additionally, we provide FedProx and SCAFFOLD aggregation methods to further enhance the capabilities of the researcher service in handling diverse data distributions and sample sizes.

5.5.4 Communication scheme

Every instance is packaged and deployed in form of Docker containers [Merkel, 2014] interacting between each other through HTTP requests. Containerized instances help to overcome software/hardware heterogeneity issues by creating an isolated virtualized environment with a predefined operating system (OS), thus improving reproducibility as every center run on the same software environment. This scheme also achieves scalability and modularity when dealing with large amounts of clients. In this case, multiple instances of the API can be created under a load balancer, while federator instances can be separately deployed on a dedicated computation infrastructure.

5.5.5 Security

Fed-BioMed addresses typical security issues regarding communication (e.g. man-in-the-middle and impersonation attacks) and access permissions as follows: 1) all requests are encrypted by using HTTP over SSL, 2) user authentication relies on a password-based scheme, and 3) reading/writing operations are restricted by role definitions. Protection from adversarial or poisoning attacks [Bhagoji et al., 2019] is currently not in the scope of this work, but can be naturally integrated as part of the federator. In the future, malicious attacks will be also prevented by implementing certification protocols attesting the safety of the model source code before deployment [Shen et al., 2016].

5.5.6 Traceability

To allow transparency to the centers and for the sake of technical support, each instance leverages on a logging system that allows to keep track of every request made, shared data and of the current available jobs.

The architecture behind Fed-BioMed is illustrated in Figure 5.1. The common procedure involves the deployment of one or multiple jobs from the researchers. Each job must contain the architecture model to be trained and its initialized parameters for reproducibility, the number of rounds, and the federator instance to be used as optimizer.

5.6 Related work

The interest in federated learning (FL) has led to a significant increase in the number of FL frameworks in recent years. Some well-established frameworks, such as TensorFlow Federated (TFF)[2], FedML [He et al., 2020], IBM-FL [Ludwig et al., 2020], FATE [Liu et al., 2021a], PaddleFL [Ma et al., 2019], and PySyft [Ziller et al., 2021], are not specifically tailored for biomedical applications but could potentially be deployed in this domain. Alternatively, more recent works that were not considered at the time of this work have emerged and have been applied to medial imaging such as SubstraFL [Galtier et al., 2019], OpenFL [Reina et al., 2021], and Flare [Roth et al., 2022].

5.6.1 Non-medical federated learning frameworks

IBM Federated Learning

IBM-FL [Ludwig et al., 2020] is a Python framework designed for federated learning

[2] `https://www.tensorflow.org/federated`

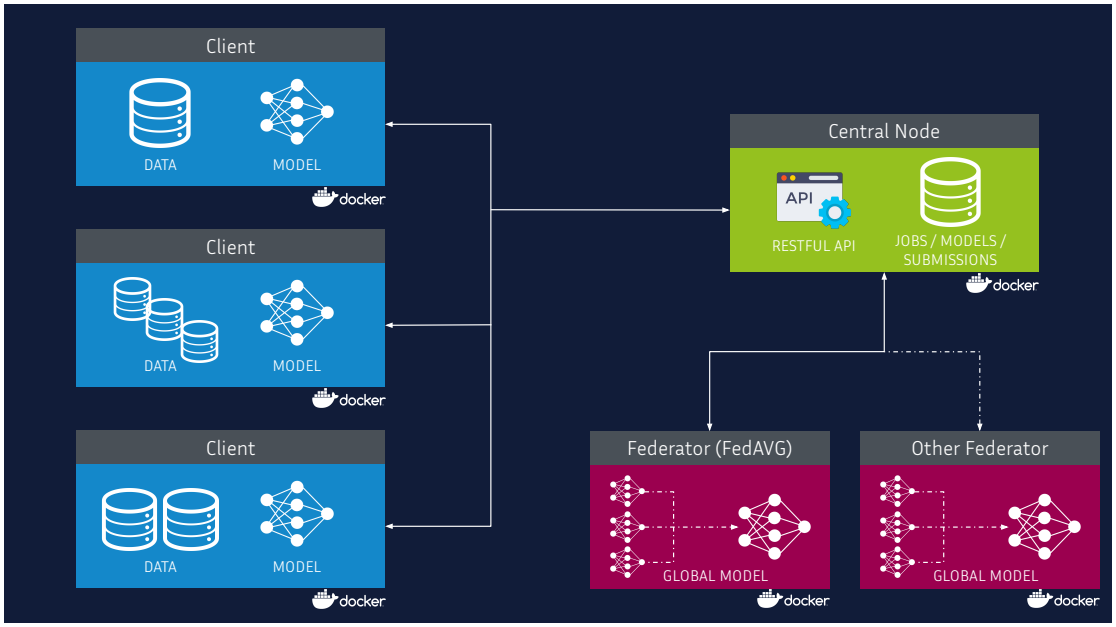


Fig. 5.1.: Clients providing different data sources share their local model parameters with the central node. The central node creates the jobs that will be run by the clients, and transmits the initialization parameters for the models in training. The federator gathers the collected parameters and combines then into a global model that is subsequently shared with the clients for the next training round. As instances are isolated in containers, new instances, such as a new federators (dashed line), can be introduced without altering the behavior of the infrastructure.

in enterprise environments. The framework provides a basic fabric for FL, upon which advanced features can be built. It is not dependent on any specific machine learning framework and supports various learning topologies, such as a shared aggregator, and protocols. IBM Federated Learning aims to offer a solid foundation for federated learning, enabling a wide variety of learning models, topologies, and learning models, particularly in enterprise and hybrid-cloud settings.

PySyft

PySyft[Ziller et al., 2021] is a popular choice, serving as a wrapper for TensorFlow and PyTorch in federated learning scenarios. While PySyft offers a simple wrapper, it lacks security and scalability features for production environments due to its reliance on a single channel for parameter transmission and event triggering. Furthermore, it only supports multiple nodes in development mode through `Virtual workers`. PySyft's production solution, Duet, permits only a single node-to-researcher connection, which is suboptimal for FL in production settings.

TensorFlow Federated

TensorFlow Federated (TFF) is an open-source framework designed for machine learning and other computations on decentralized data. It aims to facilitate research and experimentation with Federated Learning (FL), an approach where a shared global model is

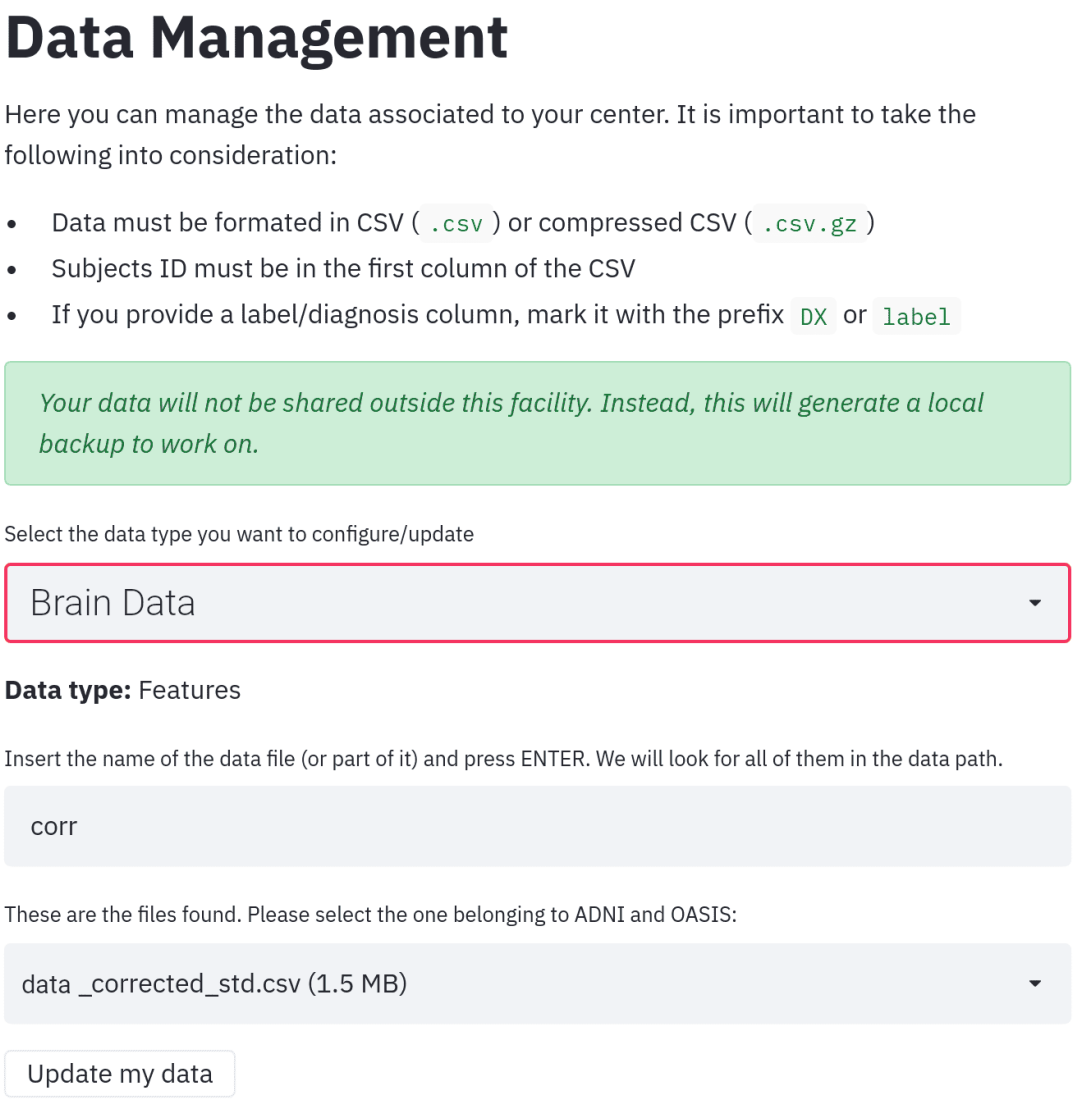


Fig. 5.2.: A screenshot of the data management utility that is integrated within the Fed-BioMed framework. This utility plays a crucial role in facilitating the secure and efficient transfer of data between different healthcare institutions and research centers. With the Fed-BioMed data management utility, users can securely upload, access, and share data with other users within a federated learning setup. The utility also includes features such as data pre-processing, quality control, and privacy-preserving mechanisms to ensure that sensitive information is protected at all times.

trained across multiple participating clients while retaining their training data locally. TFF allows developers to simulate federated learning algorithms on their models and data and experiment with novel algorithms.

The TFF platform comprises two layers: the Federated Learning (FL) layer, which offers high-level interfaces for integrating existing Keras or non-Keras machine learning models into the TFF framework, and the Federated Core (FC) layer, providing lower-

level interfaces for expressing custom federated algorithms. The FC layer combines TensorFlow with distributed communication operators within a strongly-typed functional programming environment.

Currently, TFF is designed for experimentation purposes and does not include tools for production settings, such as deployment on mobile phones. The framework supports a local simulation runtime for experiments. It is expected that the open-source ecosystem surrounding TFF will evolve to incorporate runtimes targeting physical deployment platforms.

Flower

Flower[Beutel et al., 2020], another FL framework, utilizes the same communication channel for both event triggering and model parameter transmission. Although Flower potentially supports a large number of nodes, researchers must distribute the same script across nodes to run an experiment, complicating deployment. Additionally, Flower requires file paths to be specified or hard-coded at each node for data loading, making it unsuitable for continuous delivery-centric applications. The framework also delegates messaging service management to the researcher, reducing system resiliency in the event of researcher connection or availability issues.

FATE (Federated AI Technology Enabler)

FATE[Liu et al., 2021a] offers a production-grade package enabling the deployment of existing multi-party computation (MPC) and FL solutions, including homomorphic encryption capabilities. However, its focus on predefined methods limits flexibility for research purposes.

Paddle FL

Paddle FL[Ma et al., 2019] presents an industrial FL solution specifically designed for the PaddlePaddle[3] machine learning library. While it provides features such as support for differential privacy, Paddle FL remains exclusive to the PaddlePaddle framework, lacking compatibility with PyTorch or TensorFlow.

5.6.2 Medical oriented federated learning frameworks

SubstraFL

SubstraFL [Galtier et al., 2019], developed by Owkin, is a Python library based on the Substra library. Currently SubstraFL is utilized in healthcare applications such as drug discovery within the MELLODY project and oncology. SubstraFL's architecture is built upon collaboration, privacy, and traceability. While it does not provide any tools

[3] `https://github.com/PaddlePaddle/Paddle`

	Fed-BioMed	Flower	Tensorflow FL	FATE	PaddleFL	PySyft
Federated Learning:						
Gradient-descent based models	●	●	●	●	●	●
Support for multiple ML libraries	●	●				●
Node sampling strategies support	●	●	●			
Aggregation methods support	●	●	●			○
Model agnostic	●					○
Privacy:						
f-DP mechanisms support, MPC	○				●	●
Other features:						
Ready for production (deployable)	●			●	●	○
Node data consent management	●					●
Non-gradient-descent based models	●	○				
VPN support provided	●					

Tab. 5.1.: Comparison of Existing Frameworks in Federated Learning (FL) as of July 2021. Green circles represent included features, while yellow circles indicate expected features. Flare was not considered due to its closed-source nature and limited compatibility with NVIDIA hardware.

dedicated to healthcare or biomedical assets, its distributed ledger-based architecture ensures secure and transparent operations. However, its centralized governance and roadmap may hinder the development of an open-source community.

OpenFL

OpenFL [Reina et al., 2021] is a Python-based FL library, was originally designed for healthcare applications and later expanded to be more general from the use case. OpenFL's strong focus on cybersecurity, achieved through measures like TLS-encrypted communication and PKI certificates which is optimal for multi-institutional collaborations. While tutorials are provided for medical applications, no core functionalities are specific to biomedical data sources. Additionally, the adoption of OpenFL may be limited due to its reliance on proprietary hardware.

Flare

Flare, another Python-based FL library developed by NVIDIA, has been used in various healthcare applications, including the development of a triaging model for COVID-19 patients [Dayan et al., 2021] and classification and segmentation tasks on medical images [Roth et al., 2020; Sarma et al., 2021]. Flare's design principles focus on scalability, flexibility, and lightweight, making it suitable for cross-silo FL in production settings and simulated FL for researchers. Although Flare has been showcased in healthcare settings, it is still a generic framework without specific tools for biomedical data sources. Furthermore, Flare's reliance on specialized hardware and configuration files may limit its flexibility and applicability in hospital settings.

5.7 Usage

Deploying a new model.
Fed-BioMed relies on a common convention for deploying a new model. A model must be defined as a PyTorch [Paszke et al., 2019] module class, containing the common methods for any torch module:

- `__init__(**kwargs)` method: defines the model initialization parameters;
- `forward()` method: provides instructions for computing local model updates.

Once the new class is defined, it can be integrated in the model zoo for both clients and federators (Figure 5.3, left).

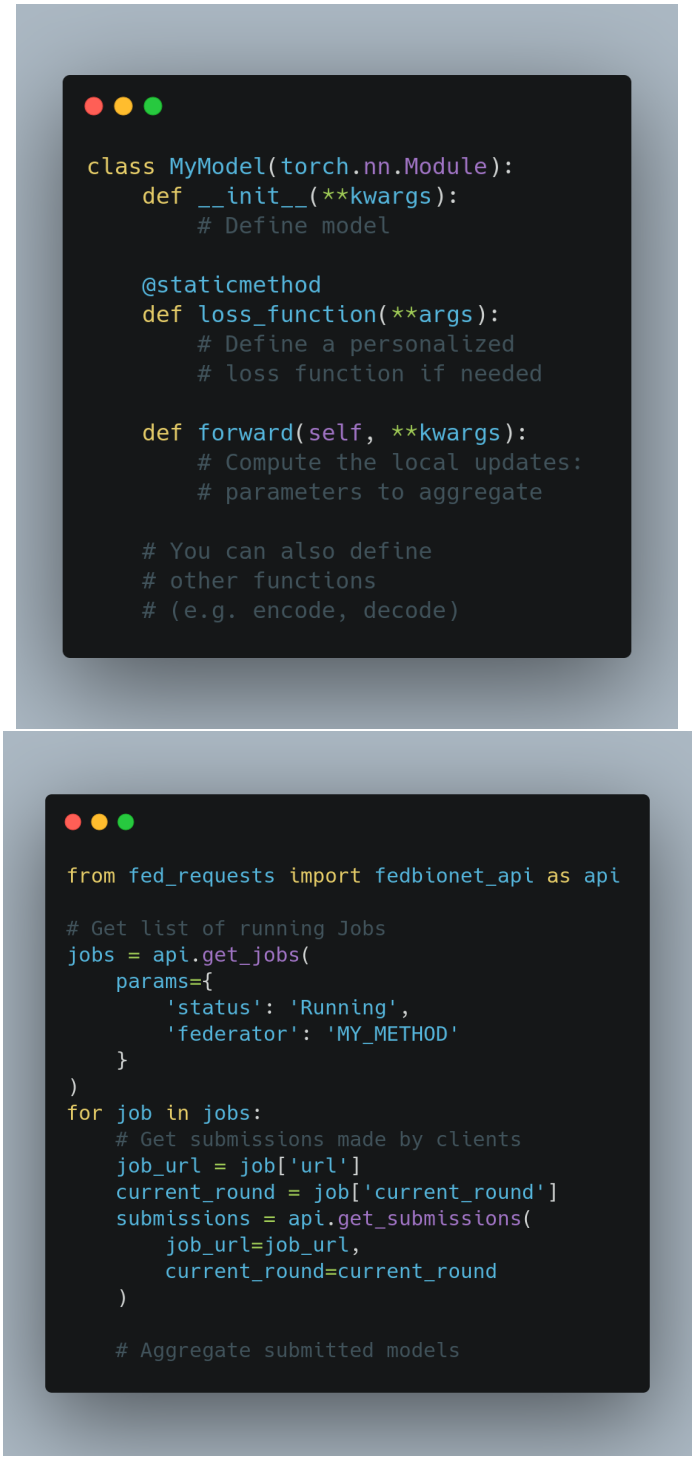
Deploying a new federator or aggregation function in the backend is obtained by creating a containerized service that queries the API for the necessary submissions at each round, and subsequently aggregates the submitted local updates (Figure 5.3, right).

5.8 Applications and Case Studies

This section showcases Fed-BioMed through two different experiments. The first experiment involves analyzing the MNIST dataset [LeCun et al., 2010] with the participation of 25 clients. The second experiment involves a multi-centric analysis of brain imaging data, conducted in collaboration with four research partners located in different geographical regions."

5.8.1 MNIST analysis

The 60000 MNIST images were equally split among 25 centers. Each center was synthetically emulated and setup in order to interact with the centralized API. The model was represented by a variational autoencoder (VAE) implementing non-linear encoding and decoding functions composed by three layers respectively, with a 2-dimensional associated latent space. For training we used a learning rate of $l = 1 \times 10^{-3}$, 10 local epochs for 30 optimization rounds, while the federated aggregating scheme was FedAvg.



```python
class MyModel(torch.nn.Module):
    def __init__(**kwargs):
        # Define model

    @staticmethod
    def loss_function(**args):
        # Define a personalized
        # loss function if needed

    def forward(self, **kwargs):
        # Compute the local updates:
        # parameters to aggregate

    # You can also define
    # other functions
    # (e.g. encode, decode)
```

```python
from fed_requests import fedbionet_api as api

# Get list of running Jobs
jobs = api.get_jobs(
    params={
        'status': 'Running',
        'federator': 'MY_METHOD'
    }
)
for job in jobs:
    # Get submissions made by clients
    job_url = job['url']
    current_round = job['current_round']
    submissions = api.get_submissions(
        job_url=job_url,
        current_round=current_round
    )

    # Aggregate submitted models
```

Fig. 5.3.: Examples of model declaration (left) and creation of a new federator (right).

5.8.2 Brain imaging data analysis

In this experiment we use our framework to perform dimensionality reduction in multi-centric structural brain imaging data (MRI) across datasets from different geographical

locations, providing cohorts of healthy individuals and patients affected by cognitive impairment.

Four centers participated to the study and were based geographically as follows: two centers in France (Center 1 and Center 4), one in the UK (Center 3), and a last one in the US (Center 2). The central node and the federator were also located in France. Each center was running on different OS and the clients were not 100% online during the day allowing to test the robustness of the framework in resuming the optimization in real-life conditions.

For each center, data use permission was obtained through formal data use agreements. Data characteristics across clients are reported in Table 5.2. A total of 4670 participants were part of this study, and we note that the data distribution is heterogeneous with respect to age, range and clinical status. 92 features subcortical volumes and cortical thickness were computed using FreeSurfer [Fischl, 2012] and linearly corrected by sex, age and intra-cranial volume (ICV) at each center.

This analysis involved data standardization with respect to the global mean and standard deviation computed across centers, and dimensionality reduction was performed via a VAE implementing a linear embedding into a 5-dimensional latent space. Federated data standardization was implemented as in [Silva et al., 2019], while VAE's parameters aggregation was performed through FedAvg. Federated learning was performed by specifying a pre-defined budget of 30 rounds of client-server iterations in total with 15 epochs/client-round at a learning rate of $l = 1 \times 10^{-3}$.

	Center 1	**Center 2**	**Center 3**	**Center 4**
No. of participants (M/F)	448/353	454/362	1070/930	573/780
Clinical status				
No. healthy	175	816	2000	695
No. MCI and AD	621	0	0	358
Age $\pm$ sd (range) [years]	73.74 $\pm$ 7.23	28.72 $\pm$ 3.70	63.93 $\pm$ 7.49	67.58 $\pm$ 10.04
Age range [years]	54 - 91	22 - 37	47 - 81	43 - 97

Tab. 5.2.: Demographic information for each of the centers that participated in training models using their MRI-derived brain data is provided below. The cohorts include individuals with Mild Cognitive Impairment (MCI) and Alzheimer's Disease (AD).

5.9 Results

The evolution of the models during training can be assessed by analyzing the weights' norm across iterations (Figure 5.4 and supplementary material for the complete set of weights). MNIST parameters evolution is shown in the top panel, while the related test set projected onto the latent space is shown in Figure 5.5, left panel, describing a meaningful variability across digits and samples.

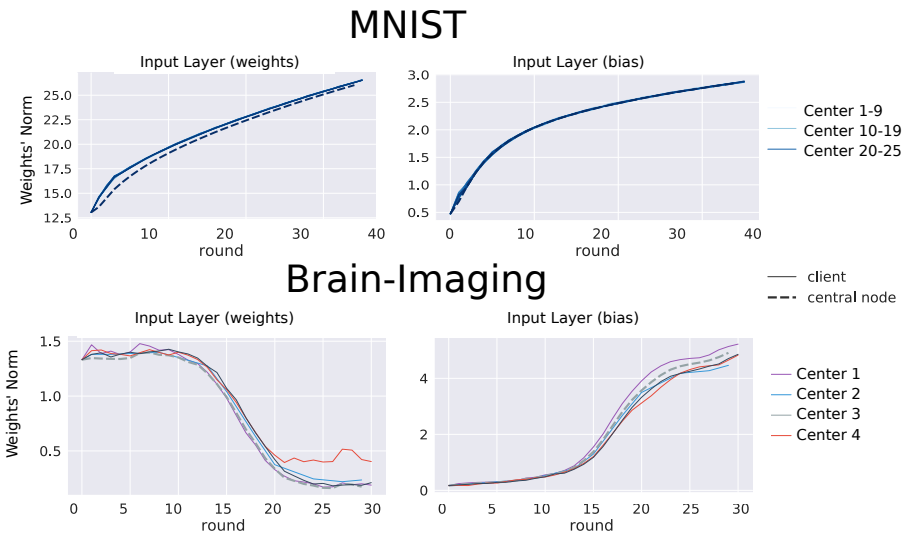


Fig. 5.4.: Illustration of parameter evolution for VAE parameters (input layer). The federated model closely follows the clients' weights distribution. **Top:** MNIST across 25 centers with equally distributed data. **Bottom:** brain-imaging heterogeneous dataset across 4 centers with unevenly distributed data. Continuous lines: clients weights' norm. Dashed lines: federated model weights' norm.

Concerning the brain imaging data analysis experiment, the evolution of the encoding parameters throughout the 30 optimization rounds is shown in Figure 5.4, bottom panel. For this real-world application we also collected each client's elapsed time to report its local updates to the central node, as well as the average time per round in the best scenario (Figure 5.6). As expected, the clients' time varies depending on the geographical proximity with the central node, as well as on the local upload/download speed. The model was further investigated by inspecting the latent space on the subset of the training data available to the coordinating Center 1. The right panel of Figure 5.5 shows that although most of the training data for the VAE comes from healthy and young participants, the model is also able to capture the pathological variability related to cognitive impairment in aging. The latent variables associated to the observations of Center 1 indeed show significantly different distributions across different clinical groups, from healthy controls (CN), to subjects with mild cognitive impairment (MCI), and patients with Alzheimer's disease (AD).

5.10 Conclusions

This work presents an open-source framework for deploying general federated learning studies in healthcare, providing a production ready reference to deploy new studies based on federated models and optimization algorithms. Our experimental results show that the framework is stable in communication, while being robust to handle clients going temporarily out of the grid. Scalability is obtained thanks to the use of

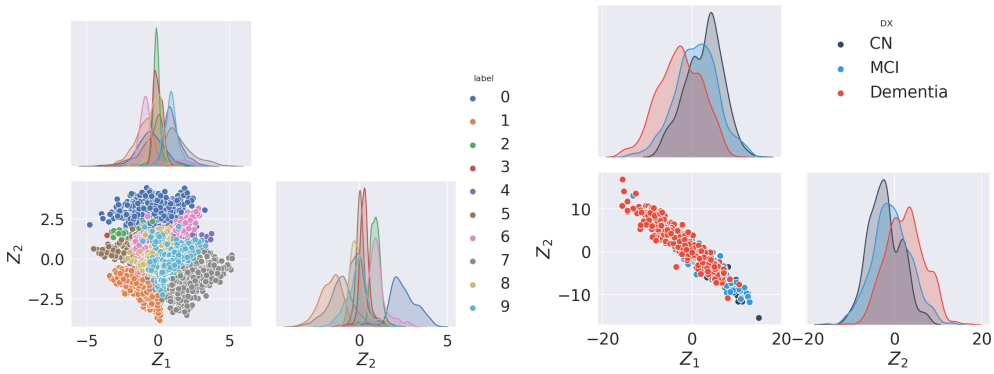


Fig. 5.5.: **Left:** MNIST pixel data projected onto the latent space. **Right:** Brain features of Center 1 projected onto the first 2 components in the latent space. Although the model was trained with unbalanced data, it is still able to capture pathological variability. CN: healthy controls; MCI: mild cognitive impairment; Dementia: dementia due to with Alzheimer's disease.

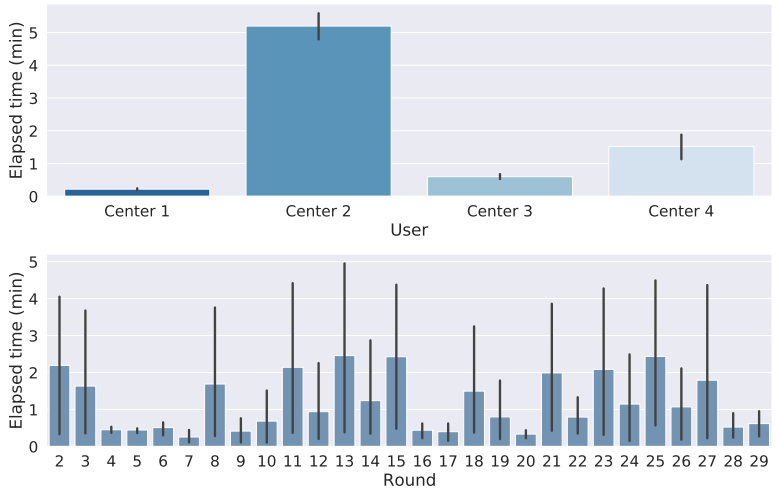


Fig. 5.6.: **Top:** User average elapsed time per round (since a new version of the model is made available and each user to submit its local update). Geographical and data distribution: Center 1 (FR) , Center 2 (US), Center 3 (UK) and Center 4 (FR). **Bottom:** Averaged elapsed time across centers per round of updates.

containerized services. The ability to handle client-authentication and the use of secured broadcasting protocols are also appealing security features of Fed-BioMed. While the experiments mostly focused on VAE and Federated Averaging as aggregating paradigm, our framework is completely extensible to other distributed optimization approaches. Future work will therefore integrate additional models for the analysis of different data modalities and bias, as well as enhanced secure P2P encryption. Concerning the clinical experimental setup, the brain application was chosen to provide a demonstration of our framework to a real-life analysis scenario, and it is not aimed to address a specific clinical question. In the future, the proposed work Fed-BioMed will be a key component for clinical studies tailored to address challenging research questions, such as the analysis of imaging-genetics relationships in current meta-analysis initiatives [Thompson et al., 2014].

5.11 Supplementary Material

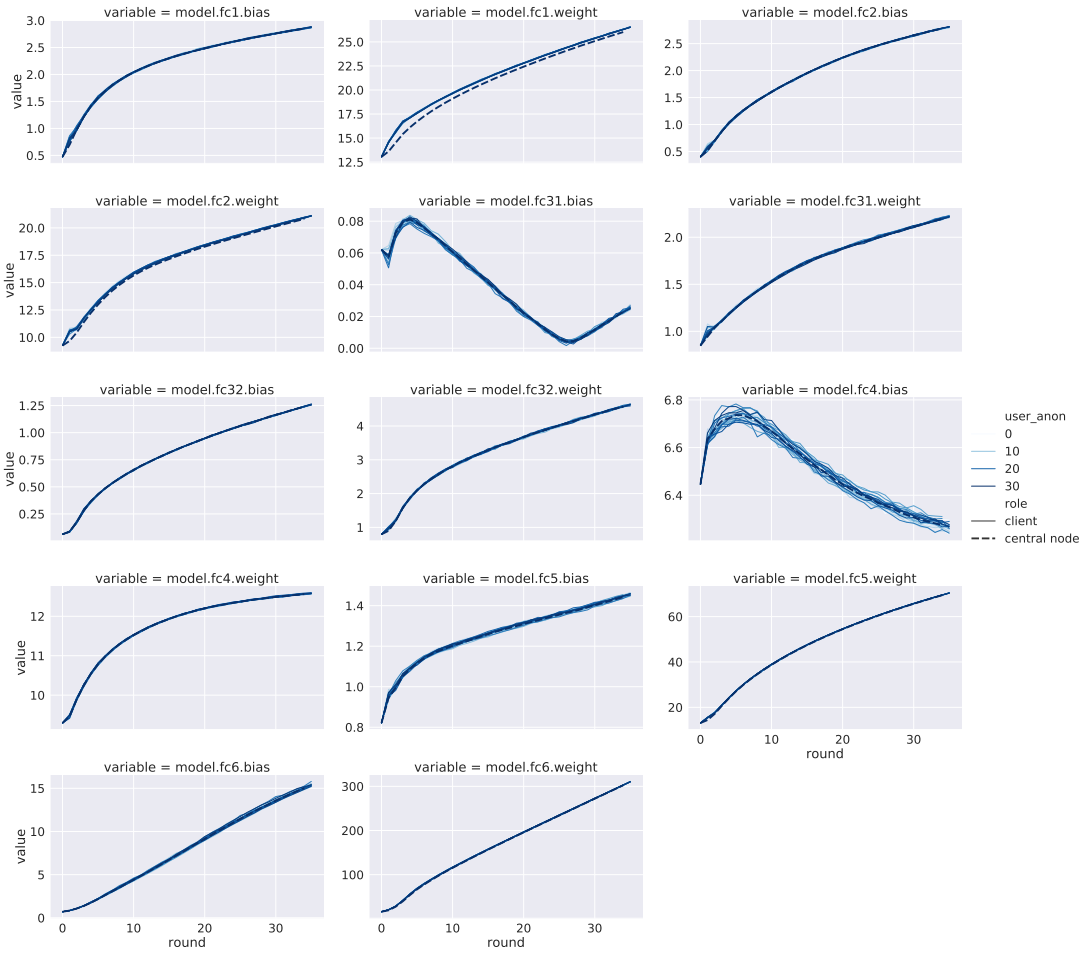


Fig. 5.7.: Parameter convergence for encoding-decoding VAE parameters. The federated model closely follows the clients' weights distribution. MNIST across 25 centers with equally distributed data.Dashed lines: federated model weights' norm.

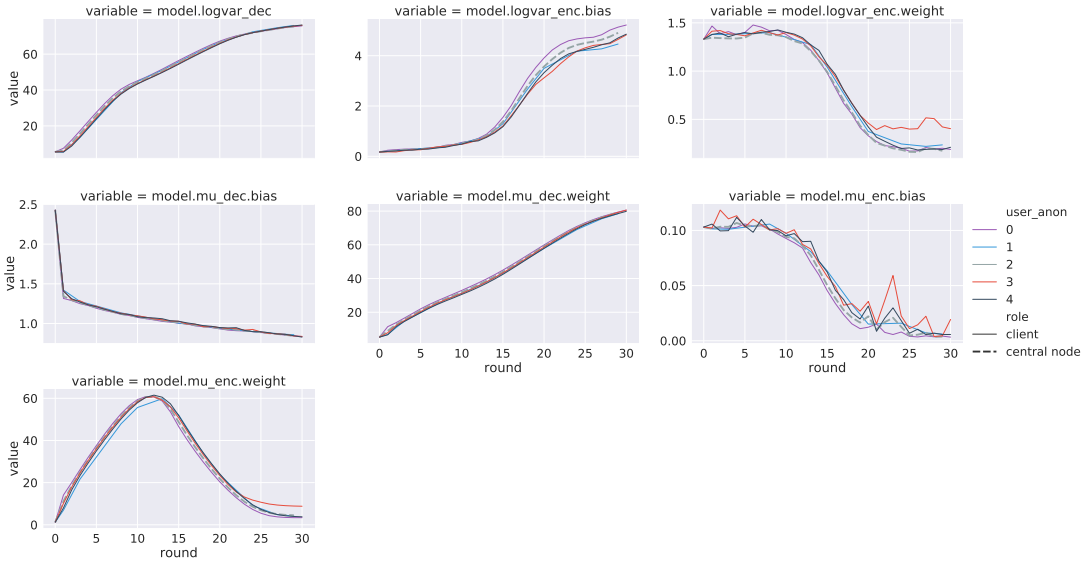


Fig. 5.8.: Parameter convergence for encoding-decoding VAE parameters. Brain-imaging heterogeneous dataset across 4 centers with unevenly distributed data. Continuous lines: clients weights' norm. Dashed lines: federated model weights' norm.

Conclusion

Throughout this thesis, we have presented a comprehensive investigation of federated learning and data harmonization techniques for multi-centric neuroimaging studies. By addressing some of the aforementioned challenges in Section 1.5. In particular, those associated with data heterogeneity due to diverse data sources, the requirements in the optimization process, and the need for standardized infrastructure and a unified framework. As a result, our research contributes to the field of federated learning research in healthcare by proposing more precise methodologies for analyzing heterogeneous large-scale neuroimaging data. The key contributions of this work can be summarized as follows.

Summary of the Main Contributions

Chapter 2: Fed-ComBat: Secure Data Harmonization via Federated Learning

In Chapter 2, We make significant contributions to the field of federated learning for data harmonization in healthcare. We introduce Fed-ComBat, a novel federated framework that enables data harmonization via federated learning. By leveraging federated learning, Fed-ComBat allows for batch effect harmonization on decentralized data without the need for data centralization or sharing. Our contributions can be summarized as follows:

- We propose a generalization of the ComBat formulation by [Johnson et al., 2007b] that allows for the adaptation of any function approximation. This generalization extends existing efforts towards nonlinear covariate effect preservation, as demonstrated by [Pomponio et al., 2020b].

- Thanks to the proposed formulation, we could make use of a Multi-Layer Perceptron (MLP) as a multivariate nonlinear approximator to preserve nonlinearities and leverage FL techniques for optimization. By incorporating the MLP, Fed-ComBat

achieves the preservation of complex nonlinear covariate effects without prior knowledge of which variables exhibit nonlinear behavior or interactions. This capability enables global cohort harmonization without the need to share individual patient data.

- By leveraging federated learning, we enabled global harmonization. Institutions can collaboratively harmonize cohorts without the need to share individual patient data. This privacy-aware approach ensures data control and mitigates the risk of sensitive data leakage, facilitating secure and distributed research collaborations in healthcare.

Chapter 3: Federated Data Harmonization in Biomedical Research using Mixed Effects Models: A Focus on Conditional Variational Autoencoders

This work-in-progress chapter explores the potential of Conditional Variational Autoencoders (CVAEs) for data harmonization in a federated learning context, focusing on neuroimaging data. Preliminary results obtained with synthetic data are promising, indicating the potential of CVAEs for data harmonization. However, validation with real-world neuroimaging datasets is still needed. The ultimate goal is to establish a federated harmonization framework capable of handling various types of data, such as images or time series, moving beyond the current focus on neuroimaging data.

Among the current and expected contributions:

- We present a novel federated harmonization framework based on CVAEs, aiming to overcome limitations of existing data harmonization methods by capturing complex relationships and mixed effects.

- We demonstrate promising outcomes of CVAEs for data harmonization based on synthetic and real data.

- We motivate future work on the validation of the approach and its potential generalizability to different types of data such as imaging and time series, as suggested by [Girin et al., 2021].

Chapter 4: Federated black-box Bayesian optimization

This chapter introduces a novel federated optimization approach based on black-box Bayesian optimization (BO) to address key challenges in federated learning, including

communication usage, hardware heterogeneity, and resource availability across participating institutions. We have also aimed to improve the generalizability of models in heterogeneous setups by relying on BO guarantees. The proposed method was evaluated on synthetic data and multicentric brain data from four different studies for Alzheimer's disease. Despite the expected potential of BO in this context, results highlight scalability issues as number of parameters in the models increase. This informs future research directions, including overcoming scalability limitations and investigating privacy-preserving optimization techniques.

Main contributions:

- We Introduced a new federated optimization method leveraging black-box Bayesian optimization, aiming to centralize the optimization process, reduce communication, and improve generalizability.

- We revealed the scalability challenges when dealing with a larger number of parameters in the model, indicating a need for future research to address this issue.

- We suggested potential research directions, including the investigation of privacy-preserving optimization techniques.

Chapter 5: Fed-BioMed: A General Open-source Frontend Framework for Federated Learning in Healthcare

Our aim in Chapter 5 was to provide Fed-BioMed as a comprehensive common framework for federated learning (FL) in healthcare, highlighting its design principles and the underlying software architecture. Through collaboration with four geographically distributed institutions in the United States, the United Kingdom, and France, we demonstrated the effectiveness and utility of Fed-BioMed in facilitating FL experiments in the healthcare domain. Our main contributions are as follows:

- We propose Fed-BioMed as a common framework for FL in healthcare that is model-agnostic and independent of any specific hardware or machine learning framework. Our open-source framework promotes transparency, collaboration, and easy adoption by the research community and healthcare practitioners, enabling them to leverage FL techniques in their data analysis workflows.

- We have played a pivotal role in developing and nurturing the Fed-BioMed initiative, which extends beyond the scope of production software development. Fed-BioMed now also aims to bring together collaborators from diverse domains, including

data science, software programming, healthcare institutions, and researchers, with the goal of advancing FL research and its practical implementation in real-world medical research applications [Cremonesi et al., 2023].

- Our work stands as a significant contribution by pioneering the utilization of a real-life scenario in FL for healthcare. Through collaboration with institutions across different countries, we have demonstrated the practical applicability of FL in preserving data privacy, optimizing communication efficiency, and addressing heterogeneity in modeling. This pioneering work serves as a foundation for further research and the adoption of FL in real-world healthcare settings.

Perspectives and Future Applications

Direct Application to Imaging

In the preceding sections, our discussion focused primarily on the methodologies proposed in Chapters 2 and 3 as they relate to the derivation of phenotypes. However, it is crucial to recognize the potential of these techniques to address other challenges, particularly in the realm of biomedical imaging.

For instance, imaging data often suffer from an array of technical and methodological biases that can critically distort the information they convey. One such common problem is the "bias field" or intensity non-uniformity, which refers to low-frequency, spatially varying intensity alterations in images, leading to difficulties in data interpretation and analysis. Bias field correction is a routine preprocessing step in biomedical imaging to ensure accurate subsequent analyses.

Leveraging the federated learning methods proposed in this work could serve as a powerful means to harmonize imaging data by correcting bias field artifacts, thereby enhancing the accuracy of biomedical image interpretation. While this application has not been explicitly discussed, the foundation laid in the present work provides a strong starting point for exploring this direction.

Enable Access to researchers through the Fed-BioMed Framework

To maximize the impact of our proposed methodologies, it is vital to consider their accessibility to the broader research community. As such, the integration of the method

described in Chapter 2 into the Fed-BioMed framework as a harmonization plugin is an exciting future direction.

Given the diverse range of biomedical data types and the multitude of possible biases that could affect them, having a tool that can harmonize data through federated learning (FL) would be highly beneficial. Our formulation covers a spectrum of harmonization needs, from linear to non-linear, making it flexible and robust enough to address a wide array of scenarios.

By providing these tools within the Fed-BioMed framework, we could offer a unified solution for data harmonization that is both easy to use and highly effective. In doing so, we could pave the way for enhanced reproducibility and more robust findings in biomedical research, ultimately advancing our understanding of complex biological phenomena.

Expanding the Impact on Large Multicentric Consortia

An additional promising direction for future research is extending the utilization of the methods presented in this thesis to large multicentric consortia, such as the Enhancing NeuroImaging Genetics through Meta-Analysis (ENIGMA) consortium [Thompson et al., 2014]. Federated learning techniques, coupled with the Fed-BioMed framework, can empower researchers within these consortia to collaboratively engage in extensive neuroimaging research projects. This is achievable while ensuring the preservation of data privacy and maintaining compliance with data protection regulations.

The methods proposed in this thesis offer a pathway to tackle the issues of data heterogeneity inherent in multicentric collaborations, optimize models collaboratively, and share computational burdens. By doing so, we can drive more efficient and impactful advancements in healthcare research and practice.

Overall, this thesis has offered methodological and technical contributions that address pivotal challenges - data heterogeneity, optimization, and thee need of an infrastructure standard and a common framework - within federated learning setups for neuroimaging research. Our work aspires to catalyze more efficient and impactful healthcare advancements, whilst ensuring the preservation of patient privacy and adherence to data governance norms.

Through the development of the proposed methods and frameworks, and their integration into real-world applications and large multicentric consortia, we aim to facilitate the wider adoption of federated learning in healthcare research. Ultimately, we hope this

will lead to enhanced understanding, diagnosis, and treatment of neurological disorders, propelling us towards a positive outcome in healthcare.

Bibliography

[, 2021] "IEEE Guide for Architectural Framework and Application of Federated Machine Learning". In: *IEEE Std 3652.1-2020* (2021), pp. 1–69 (cit. on p. 78).

[Ashburner et al., 2005] John Ashburner and Karl J Friston. "Unified segmentation". In: *Neuroimage* 26.3 (2005), pp. 839–851 (cit. on p. 33).

[Auer, 2002] Peter Auer. "Using confidence bounds for exploitation-exploration trade-offs". In: *Journal of Machine Learning Research* 3.Nov (2002), pp. 397–422 (cit. on p. 65).

[Bagdasaryan et al., 2020] Eugene Bagdasaryan et al. "How to backdoor federated learning". In: *International Conference on Artificial Intelligence and Statistics*. 2020, pp. 2938–2948 (cit. on p. 79).

[Bajaj et al., 2021] Ishan Bajaj, Akhil Arora, and MM Faruque Hasan. "Black-box optimization: Methods and applications". In: *Black box optimization, machine learning, and no-free lunch theorems*. Springer, 2021, pp. 35–65 (cit. on p. 61).

[Benkarim et al., 2022] Oualid Benkarim et al. "Population heterogeneity in clinical cohorts affects the predictive accuracy of brain imaging". In: *PLoS biology* 20.4 (2022), e3001627 (cit. on p. 8).

[Bergstra et al., 2012] James Bergstra and Yoshua Bengio. "Random search for hyper-parameter optimization." In: *Journal of machine learning research* 13.2 (2012) (cit. on pp. 61, 62).

[Bethlehem et al., 2022] Richard AI Bethlehem et al. "Brain charts for the human lifespan". In: *Nature* 604.7906 (2022), pp. 525–533 (cit. on pp. 23, 41).

[Beutel et al., 2020] Daniel J Beutel et al. "Flower: A friendly federated learning research framework". In: *arXiv preprint arXiv:2007.14390* (2020) (cit. on p. 87).

[Bhagoji et al., 2019] Arjun Nitin Bhagoji et al. "Analyzing federated learning through an adversarial lens". In: *International Conference on Machine Learning*. 2019, pp. 634–643 (cit. on pp. 79, 84).

[Bonawitz et al., 2017] Keith Bonawitz et al. “Practical secure aggregation for privacy-preserving machine learning”. In: *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 2017, pp. 1175–1191 (cit. on p. 18).

[Brisimi et al., 2018] Theodora S Brisimi et al. “Federated learning of predictive models from federated electronic health records”. In: *International journal of medical informatics* 112 (2018), pp. 59–67 (cit. on pp. 9, 16, 79).

[Brochu et al., 2010] Eric Brochu, Vlad M Cora, and Nando De Freitas. “A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning”. In: *arXiv preprint arXiv:1012.2599* (2010) (cit. on p. 62).

[Bukaty, 2019] Preston Bukaty. *The California Consumer Privacy Act (CCPA): An implementation guide*. IT Governance Publishing, 2019. URL: http://www.jstor.org/stable/j.ctvjghvnn (visited on Oct. 31, 2022) (cit. on p. 22).

[Bull, 2011] Adam D Bull. “Convergence rates of efficient global optimization algorithms.” In: *Journal of Machine Learning Research* 12.10 (2011) (cit. on p. 60).

[Calzada, 2022] Igor Calzada. “Citizens’ data privacy in China: The state of the art of the Personal Information Protection Law (PIPL)”. In: *Smart Cities* 5.3 (2022), pp. 1129–1150 (cit. on p. 22).

[Chang et al., 2018] Ken Chang et al. “Distributed deep learning networks among institutions for medical imaging”. In: *Journal of the American Medical Informatics Association* 25.8 (2018), pp. 945–954 (cit. on p. 16).

[Chen et al., 2022] Andrew A Chen et al. “Privacy-preserving harmonization via distributed ComBat”. In: *Neuroimage* 248 (2022), p. 118822 (cit. on pp. 23, 24, 27, 28).

[Cheung et al., 2023] Alexander TM Cheung et al. “Methods and impact for using federated learning to collaborate on clinical research”. In: *Neurosurgery* 92.2 (2023), pp. 431–438 (cit. on p. 24).

[Costafreda, 2009] Sergi G Costafreda. “Pooling FMRI data: meta-analysis, mega-analysis and multi-center studies”. In: *Frontiers in neuroinformatics* (2009), p. 33 (cit. on p. 9).

[Cremonesi et al., 2023] Francesco Cremonesi et al. “Fed-BioMed: Open, Transparent and Trusted Federated Learning for Real-world Healthcare Applications”. In: *arXiv preprint arXiv:2304.12012* (2023) (cit. on pp. 79, 100).

[Dai et al., 2021] Zhongxiang Dai, Bryan Kian Hsiang Low, and Patrick Jaillet. "Differentially Private Federated Bayesian Optimization with Distributed Exploration". In: *Advances in Neural Information Processing Systems*. Ed. by A. Beygelzimer et al. 2021. URL: https://openreview.net/forum?id=aj8x18_Te9 (cit. on p. 74).

[Dale et al., 1999] Anders M Dale, Bruce Fischl, and Martin I Sereno. "Cortical surface-based analysis: I. Segmentation and surface reconstruction". In: *Neuroimage* 9.2 (1999), pp. 179–194 (cit. on p. 33).

[Dayan et al., 2021] Ittai Dayan et al. "Federated learning for predicting clinical outcomes in patients with COVID-19". In: *Nature medicine* 27.10 (2021), pp. 1735–1743 (cit. on p. 88).

[De Francesco et al., 2021] Silvia De Francesco et al. "Norms for automatic estimation of hippocampal atrophy and a step forward for applicability to the Italian population". In: *Frontiers in Neuroscience* (2021), p. 786 (cit. on p. 38).

[Desikan et al., 2006] Rahul S Desikan et al. "An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest". In: *Neuroimage* 31.3 (2006), pp. 968–980 (cit. on p. 33).

[Devignes et al., 2022] Quentin Devignes et al. "Heterogeneity of PD-MCI in candidates to subthalamic deep brain stimulation: associated cortical and subcortical modifications". In: *Journal of Parkinson's disease* 12.5 (2022), pp. 1507–1526 (cit. on p. 8).

[Di Martino et al., 2014] Adriana Di Martino et al. "The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism". In: *Molecular psychiatry* 19.6 (2014), pp. 659–667 (cit. on pp. 24, 31).

[Di Martino et al., 2017] Adriana Di Martino et al. "Enhancing studies of the connectome in autism using the autism brain imaging data exchange II". In: *Scientific data* 4.1 (2017), pp. 1–15 (cit. on p. 24).

[Dwork, 2006] Cynthia Dwork. "Differential privacy". In: *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33*. Springer. 2006, pp. 1–12 (cit. on p. 11).

[Eisenhauer, 2021] Joseph G Eisenhauer. "Meta-analysis and mega-analysis: A simple introduction". In: *Teaching Statistics* 43.1 (2021), pp. 21–27 (cit. on p. 11).

[El Emam et al., 2014] Khaled El Emam and Bradley Malin. "Concepts and methods for de-identifying clinical trial data". In: *Paper commissioned by the Committee on Strategies for Responsible Sharing of Clinical Trial Data* (2014) (cit. on p. 9).

[Ellis et al., 2009] Kathryn A Ellis et al. “The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer’s disease”. In: *International psychogeriatrics* 21.4 (2009), pp. 672–687 (cit. on pp. 24, 31).

[European Commission, 2016] 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. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj (cit. on p. 22).

[Fischl, 2012] Bruce Fischl. “FreeSurfer”. In: *Neuroimage* 62.2 (2012), pp. 774–781 (cit. on p. 91).

[Fischl et al., 2004] Bruce Fischl et al. “Automatically parcellating the human cerebral cortex”. In: *Cerebral cortex* 14.1 (2004), pp. 11–22 (cit. on p. 33).

[Fortin et al., 2014] Jean-Philippe Fortin et al. “Functional normalization of 450k methylation array data improves replication in large cancer studies”. In: *Genome biology* 15.11 (2014), pp. 1–17 (cit. on p. 45).

[Fortin et al., 2017] Jean-Philippe Fortin et al. “Harmonization of multi-site diffusion tensor imaging data”. In: *Neuroimage* 161 (2017), pp. 149–170 (cit. on pp. 23, 128).

[Fortin et al., 2018] Jean-Philippe Fortin et al. “Harmonization of cortical thickness measurements across scanners and sites”. In: *Neuroimage* 167 (2018), pp. 104–120 (cit. on pp. 23, 37, 117).

[Galtier et al., 2019] Mathieu N Galtier and Camille Marini. “Substra: a framework for privacy-preserving, traceable and collaborative machine learning”. In: *arXiv preprint arXiv:1910.11567* (2019) (cit. on pp. 84, 87).

[Garcia-Dias et al., 2020] Rafael Garcia-Dias et al. “Neuroharmony: A new tool for harmonizing volumetric MRI data from unseen scanners”. In: *Neuroimage* 220 (2020) (cit. on p. 44).

[Gebre et al., 2023] Robel K Gebre et al. “Cross–scanner harmonization methods for structural MRI may need further work: A comparison study”. In: *NeuroImage* 269 (2023), p. 119912 (cit. on pp. 23, 39).

[Geyer et al., 2017] Robin C Geyer, Tassilo Klein, and Moin Nabi. “Differentially private federated learning: A client level perspective”. In: *arXiv preprint arXiv:1712.07557* (2017) (cit. on p. 18).

[Girin et al., 2021] Laurent Girin et al. “Dynamical Variational Autoencoders: A Comprehensive Review”. In: *Foundations and Trends® in Machine Learning* 15.1-2 (2021), pp. 1–175. URL: http://dx.doi.org/10.1561/2200000089 (cit. on p. 98).

[Glass, 1976] Gene V Glass. “Primary, secondary, and meta-analysis of research”. In: *Educational researcher* 5.10 (1976), pp. 3–8 (cit. on p. 9).

[Gupta et al., 2018] Otkrist Gupta and Ramesh Raskar. “Distributed learning of deep neural network over multiple agents”. In: *Journal of Network and Computer Applications* 116 (2018), pp. 1–8 (cit. on p. 80).

[He et al., 2020] Chaoyang He et al. “Fedml: A research library and benchmark for federated machine learning”. In: *arXiv preprint arXiv:2007.13518* (2020) (cit. on p. 84).

[Hedges, 1992] Larry V Hedges. “Meta-analysis”. In: *Journal of Educational Statistics* 17.4 (1992), pp. 279–296 (cit. on p. 10).

[Higgins et al., 2019] Julian PT Higgins et al. *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons, 2019 (cit. on pp. 9, 10).

[Hoffman et al., 2013] Matthew D Hoffman et al. “Stochastic variational inference”. In: *Journal of Machine Learning Research* (2013) (cit. on p. 49).

[Holly et al., 2022] Stephanie Holly et al. “Evaluation of Hyperparameter-Optimization Approaches in an Industrial Federated Learning System”. In: *Data Science – Analytics and Applications*. Ed. by Peter Haber et al. Wiesbaden: Springer Fachmedien Wiesbaden, 2022, pp. 6–13 (cit. on p. 75).

[Hsu et al., 2019] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. “Measuring the effects of non-identical data distribution for federated visual classification”. In: *arXiv preprint arXiv:1909.06335* (2019) (cit. on p. 30).

[Ibanez et al., 2021] Agustin Ibanez et al. “The multi-partner consortium to expand dementia research in Latin America (ReDLat): driving multicentric research and implementation science”. In: *Frontiers in neurology* 12 (2021), p. 631722 (cit. on p. 8).

[Jensen et al., 2012] Peter B Jensen, Lars J Jensen, and Søren Brunak. “Mining electronic health records: towards better research applications and clinical care”. In: *Nature Reviews Genetics* 13.6 (2012), pp. 395–405 (cit. on p. 8).

[Johnson et al., 2007a] W Evan Johnson, Cheng Li, and Ariel Rabinovic. “Adjusting batch effects in microarray expression data using empirical Bayes methods”. In: *Biostatistics* 8.1 (2007), pp. 118–127 (cit. on pp. 17, 23–26, 40, 44, 128).

[Johnson et al., 2007b] W Evan Johnson, Cheng Li, and Ariel Rabinovic. “Adjusting batch effects in microarray expression data using empirical Bayes methods”. In: *Biostatistics* 8.1 (2007), pp. 118–127 (cit. on p. 97).

[Jones et al., 1998a] Donald R Jones, Matthias Schonlau, and William J Welch. "Efficient Global Optimization of Expensive Black-Box Functions". In: *Journal of Global Optimization* 13.4 (1998), pp. 455–492. URL: https://doi.org/10.1023/A:1008306431147 (cit. on p. 63).

[Jones et al., 1998b] Donald R Jones, Matthias Schonlau, and William J Welch. "Efficient global optimization of expensive black-box functions". In: *Journal of Global optimization* 13.4 (1998), p. 455 (cit. on p. 64).

[Kesteren et al., 2019] Erik-Jan van Kesteren et al. "Privacy-preserving generalized linear models using distributed block coordinate descent". In: *arXiv preprint arXiv:1911.03183* (2019) (cit. on pp. 23, 25).

[Kim et al., 2017] Yejin Kim et al. "Federated tensor factorization for computational phenotyping". In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2017, pp. 887–895 (cit. on p. 80).

[Kim et al., 2022] SeungWook Kim et al. "Harmonization of Multicenter Cortical Thickness Data by Linear Mixed Effect Model". In: *Frontiers in Aging Neuroscience* 14 (2022) (cit. on p. 44).

[Kingma et al., 2014] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014) (cit. on p. 16).

[Konečnỳ et al., 2016] Jakub Konečnỳ et al. "Federated learning: Strategies for improving communication efficiency". In: *arXiv preprint arXiv:1610.05492* (2016) (cit. on pp. 13, 17, 23, 79, 80).

[LaMontagne et al., 2019] Pamela J LaMontagne et al. "OASIS-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease". In: *MedRxiv* (2019), pp. 2019–12 (cit. on pp. 24, 32).

[LeCun et al., 2010] Yann LeCun and Corinna Cortes. "MNIST handwritten digit database". In: (2010). URL: http://yann.lecun.com/exdb/mnist/ (cit. on p. 89).

[Lee et al., 2018] Junghye Lee et al. "Privacy-preserving patient similarity learning in a federated environment: development and analysis". In: *JMIR medical informatics* 6.2 (2018), e20 (cit. on p. 80).

[Leek et al., 2012] Jeffrey T Leek et al. "The sva package for removing batch effects and other unwanted variation in high-throughput experiments". In: *Bioinformatics* 28.6 (2012), pp. 882–883 (cit. on p. 44).

[Li et al., 2018] Tian Li et al. "Federated optimization in heterogeneous networks". In: *arXiv preprint arXiv:1812.06127* (2018) (cit. on pp. 18, 60, 79).

[Li et al., 2019a] Tian Li et al. "Feddane: A federated newton-type method". In: *2019 53rd Asilomar Conference on Signals, Systems, and Computers*. IEEE. 2019, pp. 1227–1231 (cit. on pp. 60, 79).

[Li et al., 2019b] Wenqi Li et al. “Privacy-preserving federated brain tumour segmentation”. In: *International workshop on machine learning in medical imaging*. Springer. 2019, pp. 133–141 (cit. on p. 24).

[Li et al., 2019c] Xiang Li et al. “On the convergence of fedavg on non-iid data”. In: *arXiv preprint arXiv:1907.02189* (2019) (cit. on pp. 17, 27).

[Li et al., 2020a] Tian Li et al. “Federated learning: Challenges, methods, and future directions”. In: *IEEE Signal Processing Magazine* 37.3 (2020), pp. 50–60 (cit. on pp. 16, 17, 40, 44, 79).

[Li et al., 2020b] Xiaoxiao Li et al. “Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results”. In: *Medical Image Analysis* 65 (2020), p. 101765 (cit. on p. 24).

[Lim et al., 2020] Wei Yang Bryan Lim et al. “Federated learning in mobile edge networks: A comprehensive survey”. In: *IEEE Communications Surveys & Tutorials* 22.3 (2020), pp. 2031–2063 (cit. on p. 17).

[Liu et al., 2021a] Yang Liu et al. “Fate: An industrial grade platform for collaborative learning with data protection”. In: *The Journal of Machine Learning Research* 22.1 (2021), pp. 10320–10325 (cit. on pp. 84, 87).

[Liu et al., 2021b] Yu Liu et al. “Morphometry of the hippocampus across the adult life-span in patients with depressive disorders: Association with neuroticism”. In: *Brain Topography* 34.5 (2021), pp. 587–597 (cit. on p. 38).

[Lorenzi et al., 2018] Marco Lorenzi et al. “Susceptibility of brain atrophy to TRIB3 in Alzheimer’s disease, evidence from functional prioritization in imaging genetics”. In: *Proceedings of the National Academy of Sciences* 115.12 (2018), pp. 3162–3167 (cit. on p. 29).

[Ludwig et al., 2020] Heiko Ludwig et al. “Ibm federated learning: an enterprise framework white paper v0. 1”. In: *arXiv preprint arXiv:2007.10987* (2020) (cit. on p. 84).

[Lydia et al., 2019] Agnes Lydia and Sagayaraj Francis. “Adagrad—an optimizer for stochastic gradient descent”. In: *Int. J. Inf. Comput. Sci* 6.5 (2019), pp. 566–568 (cit. on p. 16).

[Ma et al., 2019] Yanjun Ma et al. “PaddlePaddle: An open-source deep learning platform from industrial practice”. In: *Frontiers of Data and Domputing* 1.1 (2019), pp. 105–115 (cit. on pp. 84, 87).

[MacKay et al., 2003] David JC MacKay and David JC Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003 (cit. on p. 62).

[Mahlool et al., 2022] Dhurgham Hassan Mahlool and Mohamed Hamzah Abed. "Distributed brain tumor diagnosis using a federated learning environment". In: *Bulletin of Electrical Engineering and Informatics* 11.6 (2022). Cited by: 0; All Open Access, Gold Open Access, Green Open Access, pp. 3313–3321. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85139133360&doi=10.11591%2feei.v11i6.4131&partnerID=40&md5=1ad3123e9db42a98e13e51bd26fdcbc1 (cit. on p. 24).

[Maito et al., 2023] Marcelo Adrián Maito et al. "Classification of Alzheimer's disease and frontotemporal dementia using routine clinical and cognitive measures across multicentric underrepresented samples: A cross sectional observational study". In: *The Lancet Regional Health-Americas* 17 (2023), p. 100387 (cit. on p. 8).

[Malone et al., 2013] Ian B Malone et al. "MIRIAD—Public release of a multiple time point Alzheimer's MR imaging dataset". In: *NeuroImage* 70 (2013), pp. 33–36 (cit. on p. 24).

[Marek et al., 2018] Kenneth Marek et al. "The Parkinson's progression markers initiative (PPMI)–establishing a PD biomarker cohort". In: *Annals of clinical and translational neurology* 5.12 (2018), pp. 1460–1477 (cit. on pp. 24, 32).

[Mathew et al., 1999] Thomas Mathew and Kenneth Nordstrom. "On the equivalence of meta-analysis using literature and using individual patient data". In: *Biometrics* 55.4 (1999), pp. 1221–1223 (cit. on p. 11).

[McMahan et al., 2017] Brendan McMahan et al. "Communication-efficient learning of deep networks from decentralized data". In: *Artificial intelligence and statistics*. PMLR. 2017, pp. 1273–1282 (cit. on pp. 13, 15, 24, 27, 46, 49, 60, 66, 80).

[Merkel, 2014] Dirk Merkel. "Docker: lightweight linux containers for consistent development and deployment". In: *Linux journal* 2014.239 (2014), p. 2 (cit. on p. 83).

[Miotto et al., 2018] Riccardo Miotto et al. "Deep learning for healthcare: review, opportunities and challenges". In: *Briefings in bioinformatics* 19.6 (2018), pp. 1236–1246 (cit. on p. 7).

[Mockus, 1998] Jonas Mockus. "The application of Bayesian methods for seeking the extremum". In: *Towards global optimization* 2 (1998), p. 117 (cit. on p. 64).

[Murdoch et al., 2013] Travis B Murdoch and Allan S Detsky. "The inevitable application of big data to health care". In: *Jama* 309.13 (2013), pp. 1351–1352 (cit. on p. 7).

[Nobis et al., 2019] Lisa Nobis et al. "Hippocampal volume across age: Nomograms derived from over 19,700 people in UK Biobank". In: *NeuroImage: Clinical* 23 (2019), p. 101904. URL: https://www.sciencedirect.com/science/article/pii/S2213158219302542 (cit. on p. 38).

[Orlhac et al., 2022] Fanny Orlhac et al. “A guide to ComBat harmonization of imaging biomarkers in multicenter studies”. In: *Journal of Nuclear Medicine* 63.2 (2022), pp. 172–179 (cit. on pp. 17, 44).

[Pardalos et al., 2021] Panos M Pardalos, Varvara Rasskazova, Michael N Vrahatis, et al. *Black Box Optimization, Machine Learning, and No-Free Lunch Theorems*. Springer, 2021 (cit. on p. 68).

[Paszke et al., 2019] Adam Paszke et al. “Pytorch: An imperative style, high-performance deep learning library”. In: *Advances in neural information processing systems*. 2019, pp. 8026–8037 (cit. on p. 89).

[Perrone et al., 2019] Valerio Perrone et al. *Learning search spaces for Bayesian optimization: Another view of hyperparameter transfer learning*. Tech. rep. 2019 (cit. on pp. 67, 69).

[Plis et al., 2016] Sergey M Plis et al. “COINSTAC: a privacy enabled model and prototype for leveraging and processing decentralized brain imaging data”. In: *Frontiers in neuroscience* 10 (2016), p. 365 (cit. on p. 81).

[Pomponio et al., 2020a] Raymond Pomponio et al. “Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan”. In: *NeuroImage* 208 (2020), p. 116450 (cit. on pp. xv, 17, 23–26, 29, 37, 38, 123, 124, 129).

[Pomponio et al., 2020b] Raymond Pomponio et al. “Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan”. In: *NeuroImage* 208 (2020), p. 116450 (cit. on p. 97).

[Real et al., 2019] Esteban Real et al. “Regularized evolution for image classifier architecture search”. In: *Proceedings of the aaai conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 4780–4789 (cit. on p. 61).

[Reddi et al., 2020] Sashank Reddi et al. “Adaptive federated optimization”. In: *arXiv preprint arXiv:2003.00295* (2020) (cit. on p. 16).

[Reina et al., 2021] G Anthony Reina et al. “OpenFL: An open-source framework for Federated Learning”. In: *arXiv preprint arXiv:2105.06413* (2021) (cit. on pp. 84, 88).

[Reynolds et al., 2022] Maxwell Reynolds et al. “Combat harmonization: Empirical bayes versus fully bayes approaches”. In: *bioRxiv* (2022) (cit. on pp. 30, 33).

[Rieke et al., 2020] Nicola Rieke et al. “The future of digital health with federated learning”. In: *NPJ digital medicine* 3.1 (2020), pp. 1–7 (cit. on p. 60).

[Rosenthal et al., 2016] Liana S Rosenthal et al. “The NINDS Parkinson’s disease biomarkers program”. In: *Movement Disorders* 31.6 (2016), pp. 915–923 (cit. on pp. 24, 32).

[Roth et al., 2020] Holger R Roth et al. “Federated learning for breast density classification: A real-world implementation”. In: *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning: Second MICCAI Workshop, DART 2020, and First MICCAI Workshop, DCL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 2*. Springer. 2020, pp. 181–191 (cit. on p. 88).

[Roth et al., 2022] Holger R Roth et al. “NVIDIA FLARE: Federated Learning from Simulation to Real-World”. In: *arXiv preprint arXiv:2210.13291* (2022) (cit. on p. 84).

[Rumsfeld et al., 2016] John S Rumsfeld, Karen E Joynt, and Thomas M Maddox. “Big data analytics to improve cardiovascular care: promise and challenges”. In: *Nature Reviews Cardiology* 13.6 (2016), pp. 350–359 (cit. on p. 8).

[Russkikh et al., 2020] Nikolai Russkikh et al. “Style transfer with variational autoencoders is a promising approach to RNA-Seq data harmonization and analysis”. In: *Bioinformatics* 36.20 (2020), pp. 5076–5085 (cit. on p. 47).

[Ryffel et al., 2018] Theo Ryffel et al. “A generic framework for privacy preserving deep learning”. In: *arXiv preprint arXiv:1811.04017* (2018) (cit. on p. 81).

[Sarma et al., 2021] Karthik V Sarma et al. “Federated learning improves site performance in multicenter deep learning without data sharing”. In: *Journal of the American Medical Informatics Association* 28.6 (2021), pp. 1259–1264 (cit. on p. 88).

[Satizabal et al., 2019] Claudia L Satizabal et al. “Genetic architecture of subcortical brain structures in 38,851 individuals”. In: *Nature genetics* 51.11 (2019), pp. 1624–1636 (cit. on p. 11).

[Sattler et al., 2019] Felix Sattler et al. “Robust and communication-efficient federated learning from non-iid data”. In: *IEEE transactions on neural networks and learning systems* 31.9 (2019), pp. 3400–3413 (cit. on p. 18).

[Schmaal et al., 2016] Lianne Schmaal et al. “Subcortical brain alterations in major depressive disorder: findings from the ENIGMA Major Depressive Disorder working group”. In: *Molecular psychiatry* 21.6 (2016), pp. 806–812 (cit. on p. 11).

[Schneeweiss, 2014] Sebastian Schneeweiss. “Learning from big health care data”. In: *N Engl J Med* 370.23 (2014), pp. 2161–2163 (cit. on p. 7).

[Schwarz et al., 2021] Christopher G Schwarz et al. “Changing the face of neuroimaging research: Comparing a new MRI de-facing technique with popular alternatives”. In: *NeuroImage* 231 (2021), p. 117845 (cit. on p. 22).

[Sheller et al., 2019] Micah J Sheller et al. “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: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I 4*. Springer. 2019, pp. 92–104 (cit. on p. 16).

[Shen et al., 2016] Shiqi Shen, Shruti Tople, and Prateek Saxena. “Auror: Defending against poisoning attacks in collaborative deep learning systems”. In: *Proceedings of the 32nd Annual Conference on Computer Security Applications*. 2016, pp. 508–519 (cit. on p. 84).

[Silva et al., 2019] Santiago Silva et al. “Federated learning in distributed medical databases: Meta-analysis of large-scale subcortical brain data”. In: *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*. IEEE. 2019, pp. 270–274 (cit. on pp. 27, 29, 78, 81, 91).

[Silva et al., 2020] Santiago Silva et al. “Fed-biomed: A general open-source frontend framework for federated learning in healthcare”. In: *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning: Second MICCAI Workshop, DART 2020, and First MICCAI Workshop, DCL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 2*. Springer. 2020, pp. 201–210 (cit. on p. 78).

[Sled et al., 1998] John G Sled, Alex P Zijdenbos, and Alan C Evans. “A nonparametric method for automatic correction of intensity nonuniformity in MRI data”. In: *IEEE transactions on medical imaging* 17.1 (1998), pp. 87–97 (cit. on p. 33).

[Smith et al., 2017] Virginia Smith et al. “Federated multi-task learning”. In: *Advances in neural information processing systems* 30 (2017) (cit. on p. 17).

[Sobester et al., 2008] András Sobester, Alexander Forrester, and Andy Keane. *Engineering design via surrogate modelling: a practical guide*. John Wiley & Sons, 2008 (cit. on p. 61).

[Sohn et al., 2015] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. “Learning structured output representation using deep conditional generative models”. In: *Advances in neural information processing systems* 28 (2015) (cit. on pp. 46, 47).

[Solove et al., 2020] Daniel J Solove and Paul M Schwartz. *Information privacy law*. Aspen Publishing, 2020 (cit. on pp. 8, 9, 13).

[Sperling et al., 2014] Reisa A Sperling et al. “The A4 study: stopping AD before symptoms begin?” In: *Science translational medicine* 6.228 (2014), 228fs13–228fs13 (cit. on pp. 24, 31).

[Sun et al., 2013] Jimeng Sun and Chandan K Reddy. “Big data analytics for healthcare”. In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2013, pp. 1525–1525 (cit. on p. 8).

[Thompson et al., 2014] Paul M Thompson et al. “The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data”. In: *Brain imaging and behavior* 8.2 (2014), pp. 153–182 (cit. on pp. 93, 101).

[Torous et al., 2018] John Torous et al. “Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements”. In: *BMJ Ment Health* 21.3 (2018), pp. 116–119 (cit. on p. 7).

[Undavia et al., 2020] Jaimin Navinchandra Undavia and Atul Manubhai Patel. “Big Data Analytics in Healthcare: Applications and Challenges”. In: *International Journal of Big Data and Analytics in Healthcare (IJBDAH)* 5.1 (2020), pp. 19–27 (cit. on p. 7).

[Voigt et al., 2017] Paul Voigt and Axel Von dem Bussche. “The eu general data protection regulation (gdpr)”. In: *A Practical Guide, 1st Ed., Cham: Springer International Publishing* (2017) (cit. on pp. 8, 9, 13, 79).

[Wachinger et al., 2021] Christian Wachinger et al. “Detect and correct bias in multi-site neuroimaging datasets”. In: *Medical Image Analysis* 67 (2021), p. 101879 (cit. on pp. 17, 23).

[Webb-Vargas et al., 2017] Yenny Webb-Vargas et al. “Big data and neuroimaging”. In: *Statistics in biosciences* 9 (2017), pp. 543–558 (cit. on p. 8).

[Weiner et al., 2013] Michael W Weiner et al. “The Alzheimer’s Disease Neuroimaging Initiative: a review of papers published since its inception”. In: *Alzheimer’s & Dementia* 9.5 (2013), e111–e194 (cit. on pp. 24, 31).

[Wosik et al., 2020] Jedrek Wosik et al. “Telehealth transformation: COVID-19 and the rise of virtual care”. In: *Journal of the American Medical Informatics Association* 27.6 (2020), pp. 957–962 (cit. on p. 7).

[Yang et al., 2019] Qiang Yang et al. “Federated machine learning: Concept and applications”. In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 10.2 (2019), pp. 1–19 (cit. on p. 79).

[Yuhong et al., 2019] Wen Yuhong et al. *Federated Learning powered by NVIDIA Clara*. 2019. URL: `https://devblogs.nvidia.com/federated-learning-clara/` (visited on June 22, 2020) (cit. on p. 80).

[Zaheer et al., 2018] Manzil Zaheer et al. “Adaptive methods for nonconvex optimization”. In: *Advances in neural information processing systems* 31 (2018) (cit. on p. 16).

[Zhao et al., 2018] Yue Zhao et al. “Federated learning with non-iid data”. In: *arXiv preprint arXiv:1806.00582* (2018) (cit. on p. 17).

[Zhu et al., 2021] Hangyu Zhu et al. “Federated learning on non-IID data: A survey”. In: *Neurocomputing* 465 (2021), pp. 371–390 (cit. on p. 30).

[Ziller et al., 2021] Alexander Ziller et al. “Pysyft: A library for easy federated learning”. In: *Federated Learning Systems: Towards Next-Generation AI* (2021), pp. 111–139 (cit. on pp. 84, 85).

[Zugman et al., 2022] André Zugman et al. “Mega-analysis methods in ENIGMA: The experience of the generalized anxiety disorder working group”. In: *Human Brain Mapping* 43.1 (2022), pp. 255–277 (cit. on p. 11).

[Zuo et al., 2021] Lianrui Zuo et al. “Unsupervised MR harmonization by learning disentangled representations using information bottleneck theory”. In: *NeuroImage* 243 (2021), p. 118569 (cit. on pp. 44, 47).

List of Figures

List of Tables

A

Supplementary material for Chapter 2: Fed-ComBat: Secure Data Harmonization via Federated Learning

A.1 Comparing centralized and Fed-ComBat's formulation

As the random effects (batch effects) in this work are assumed to be site-specific, the model's design matrix $\mathbf{X}$ is constructed under a linear assumption, jointly optimizing the parameters $\hat{\alpha}_g$ and $\hat{\gamma}_{ig}$. However, for nonlinear scenarios, the model needs to be reformulated, as previously done by [Pomponio et al., 2020a]. To provide a more comprehensive understanding of the optimization problem, we propose the following generalized approach:

$$y_{ijg} = \phi(\mathbf{x}_{ij}; \boldsymbol{\theta}_g) + \mathbf{z}_{ij}\boldsymbol{\beta}_g + \delta_{ig}\varepsilon_{ijg} \tag{A.1}$$

Where with respect to the model in Equation (2.1), $\mathbf{Z} := \{\mathbf{z}_{ij}\}_{\forall_{i,j}}$ is an indicator matrix as indicated in Equation (A.2) on the batch effects that allows the elements of $\boldsymbol{\beta}_g$ be composites of α_g and γ_{ig} as shown in Appendix A.1.

$$\mathbf{z}_{ij_{(1,S)}} := \begin{cases} 1 & \text{if } \mathbf{x}_{ij} \in \text{batch i,} \\ 0 & \text{if } \mathbf{x}_{ij} \notin \text{batch i.} \end{cases} \tag{A.2}$$

$$\boldsymbol{\beta}_g = \alpha_g + \begin{pmatrix} \gamma_{1g} \\ \vdots \\ \gamma_{Sg} \end{pmatrix} \tag{A.3}$$

$$\text{s.t.} \quad \mathbb{E}[\gamma_i] = \textstyle\sum_{i=1}^{S} \frac{n_i}{N} \gamma_{ig} = 0 \quad \forall_{g \in \{1,\ldots,G\}} \tag{A.4}$$

Which with the constraint in Eqn A.4 the estimation of α_g and γ_{ig} from β_g becomes straightforward:

$$\alpha_g = \mathbb{E}[\boldsymbol{\beta}_g] = \sum_{i=0}^{S} p_i \beta_{ig} = \sum_{i=0}^{S} \frac{n_i}{N} \alpha_g + \mathbb{E}[\gamma_{ig}] \tag{A.5}$$

$$\gamma_{ig} = \beta_{ig} - \alpha_g \tag{A.6}$$

A.2 Harmonization on synthetic data

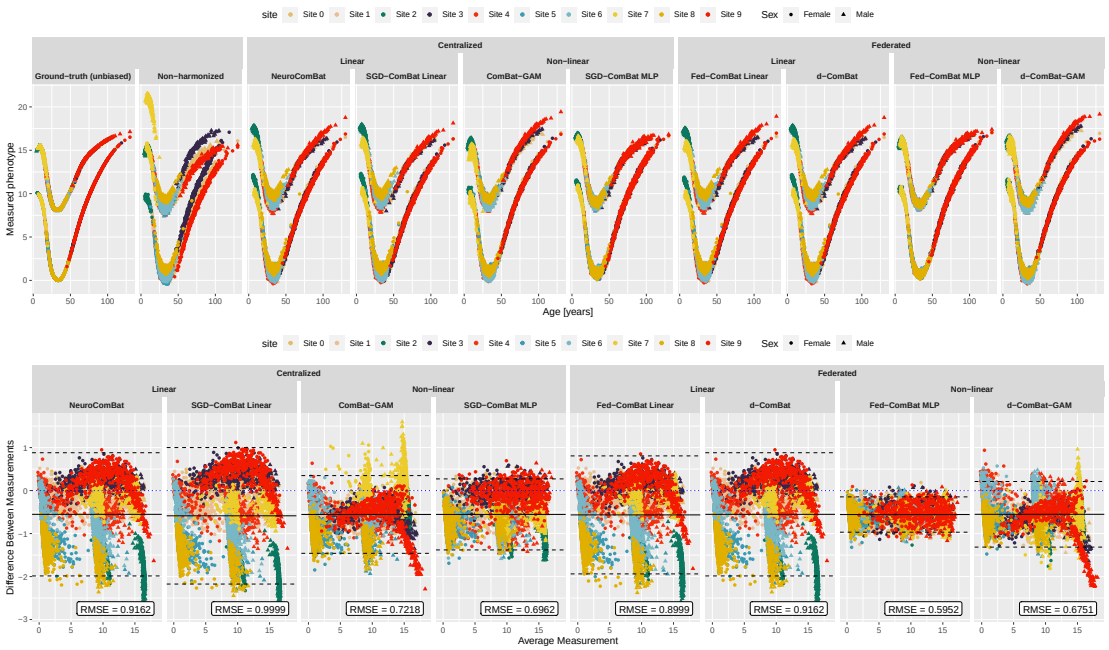


Fig. A.1.: Comparison between centralized and federated methods on synthetic data. The objective is to reconstruct the unbiased (ground-truth) function. As well as their respective bland Altman plots.

A.3 Harmonization on real data

A.3.1 Evidence of non linear effects on brain phenotypes

Table A.1 shows the difference in Akaike information criterion (AIC) between a GAM and a GLM approximating the phenotypes (brain measures) as proposed by [Pomponio et al., 2020a] (B-Splines for approximating age with DOF=10 and degree 3) as a function of Age, Sex, ICV and group (diagnosis). Negative values indicate a better fit using GAMs and hence the presence of nonlinear age effect.

A.3.2 Residual ComBat effects on real data.

Figure A.2 presents the residual plots obtained after fitting a comprehensive model to the harmonized MRI-derived phenotype data, incorporating variables such as age, sex, and group. These residual plots provide insightful visual representations of the model's performance. Notably, the observed absence of pronounced shifts or variance scaling in all the models suggests that a linear correction approach proves to be more than adequate for addressing the particular use case (see Section 2.6).

Region	ΔAIC	Region	ΔAIC	Region	ΔAIC
rh_rostralmiddlefrontal_thickness	-737.827	rh_lateralorbitofrontal_thickness	-294.661	rh_fusiform_thickness	-128.352
Right-Hippocampus	-666.327	rh_superiorfrontal_thickness	-284.615	lh_caudalmiddlefrontal_thickness	-126.349
Left-VentralDC	-631.412	rh_lateraloccipital_thickness	-284.338	CC_Anterior	-124.647
Left-Hippocampus	-588.296	rh_posteriorcingulate_thickness	-277.702	lh_fusiform_thickness	-124.049
lh_precuneus_thickness	-586.472	rh_pericalcarine_thickness	-277.491	rh_isthmuscingulate_thickness	-111.719
Right-VentralDC	-577.552	CC_Mid_Posterior	-276.176	rh_parsopercularis_thickness	-106.362
lh_rostralmiddlefrontal_thickness	-570.793	lh_lateralorbitofrontal_thickness	-258.546	rh_parahippocampal_thickness	-105.893
lh_cuneus_thickness	-544.505	rh_parsorbitalis_thickness	-246.652	Right-Cerebellum-Cortex	-100.931
Left-Lateral-Ventricle	-530.619	rh_rostralanteriorcingulate_thickness	-243.986	Right-Pallidum	-94.757
rh_cuneus_thickness	-515.156	lh_parsorbitalis_thickness	-230.732	rh_middletemporal_thickness	-87.857
Right-Lateral-Ventricle	-514.104	Left-Caudate	-229.629	lh_parahippocampal_thickness	-83.225
rh_precuneus_thickness	-505.154	Right-Caudate	-226.743	lh_parsopercularis_thickness	-67.404
rh_lingual_thickness	-489.542	CC_Central	-217.088	lh_transversetemporal_thickness	-67.285
Brain-Stem	-451.604	lh_parstriangularis_thickness	-215.885	lh_superiortemporal_thickness	-63.792
lh_paracentral_thickness	-443.599	lh_lateraloccipital_thickness	-213.35	lh_middletemporal_thickness	-59.581
lh_lingual_thickness	-443.392	CC_Mid_Anterior	-209.57	rh_superiortemporal_thickness	-57.407
lh_inferiorparietal_thickness	-439.908	lh_superiorfrontal_thickness	-206.472	rh_transversetemporal_thickness	-56.076
lh_superiorparietal_thickness	-424.113	rh_caudalmiddlefrontal_thickness	-206.01	lh_insula_thickness	-54.416
rh_superiorparietal_thickness	-420.171	rh_entorhinal_thickness	-194.472	rh_insula_thickness	-53.683
rh_medialorbitofrontal_thickness	-417.312	lh_caudalanteriorcingulate_thickness	-191.647	lh_precentral_thickness	-51.688
Left-Amygdala	-409.543	lh_medialorbitofrontal_thickness	-189.768	lh_inferiortemporal_thickness	-48.254
Right-Cerebellum-White-Matter	-392.708	lh_entorhinal_thickness	-176.719	Left-Pallidum	-40.941
rh_inferiorparietal_thickness	-381.635	lh_isthmuscingulate_thickness	-168.909	rh_precentral_thickness	-37.279
rh_paracentral_thickness	-358.113	rh_supramarginal_thickness	-161.94	rh_inferiortemporal_thickness	-37.05
Right-Amygdala	-324.232	lh_postcentral_thickness	-157.198	Left-Putamen	-34.233
CC_Posterior	-323.508	rh_postcentral_thickness	-154.755	Right-Putamen	-24.207
rh_parstriangularis_thickness	-310.166	lh_supramarginal_thickness	-148.167		
lh_posteriorcingulate_thickness	-306.379	rh_caudalanteriorcingulate_thickness	-144.195		
lh_pericalcarine_thickness	-305.481	lh_rostralanteriorcingulate_thickness	-128.495		

Tab. A.1.: Difference of Akaike information criterion (AIC) between a nonlinear and a linear model using age as the exogenous variable across cortical thickness measures. $\Delta\text{AIC} < 0$ suggest that the nonlinear model is a best fit hence suggesting non linear effects of age in such negative differences.

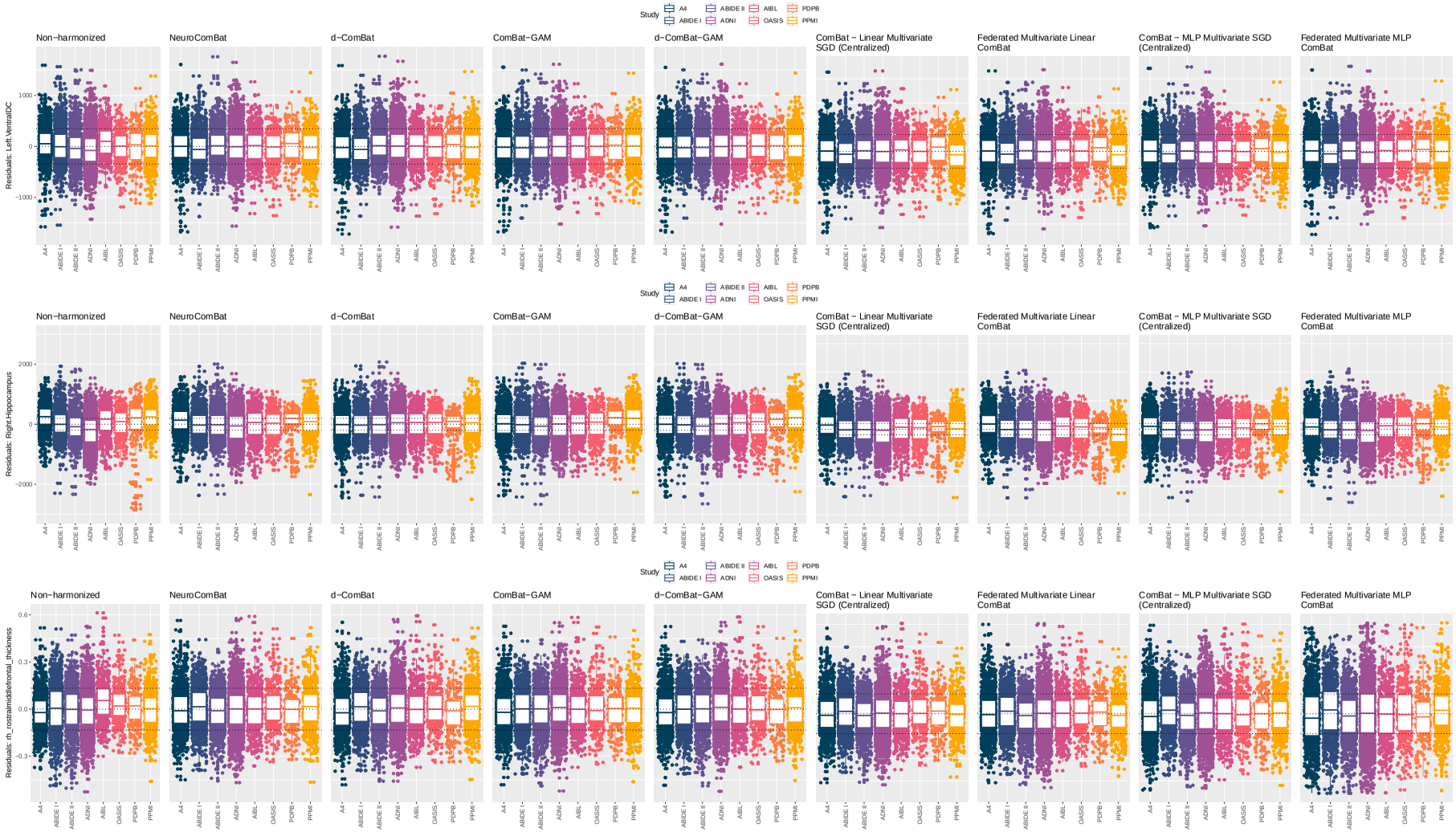


Fig. A.2.: Extension of Figure 2.6 for all the methods evaluated in this work and top regions in Figure 2.5a.

A.4 Identifiability of ComBat parameters

A statistical model is said to be identifiable if there exists a unique set of parameter values that can explain the observed data. Mathematically, a statistical model is identifiable if the following conditions hold:

Let $\mathcal{X}$ be the sample space and $f_\theta(x)$ be the probability density function or probability mass function of the statistical model, indexed by the parameter vector $\theta \in \Theta$.

1. θ is finite-dimensional and $\Theta \subseteq \mathbb{R}^d$ for some integer d.

2. For any $\theta_1, \theta_2 \in \Theta$, if $f_{\theta_1}(x) = f_{\theta_2}(x)$ for all $x \in \mathcal{X}$, then $\theta_1 = \theta_2$. In other words, the likelihood function, which is defined as $L(\theta; x) = f_\theta(x)$, is injective with respect to θ.

Condition 1 ensures that the parameter vector is finite-dimensional and belongs to some subset of Euclidean space. Condition 2 ensures that there exists a one-to-one correspondence between the parameter values and the probability distribution, which is necessary for unique estimation of the parameters. These one-to-one correspondence in the context of Fed-ComBat in Equation (2.3) ensures that there the true expectation effect is captured by $\phi(\mathbf{x}_{ij}; \boldsymbol{\theta}) + \alpha_g$ and not by $\hat{\gamma_{ig}}$. Equivalently for Equation (2.4). In practice this is translated to:

- No allowing intercepts in the neural network architectures so only one is estimated by α_g.

- Using a design matrix $\mathbf{z}$ shown in Equation (A.1) that allow the estimation of a single set of intercepts that can then be decomposed $\boldsymbol{\beta}_g = \{\alpha_g + \gamma_{ig}\}_{i=1}^{S}$.

A.5 ComBat formulations

A.5.1 Linear Combat – NeuroCombat [Johnson et al., 2007a; Fortin et al., 2017]

$$y_{ijg} = \alpha_g + \mathbf{x}_{ij}^{\intercal}\boldsymbol{\theta}_g + \gamma_{ig} + \delta_{ig}\varepsilon_{ig} \tag{A.7}$$

A.5.2 ComBat-GAM [Pomponio et al., 2020a]

$$y_{ijg} = \alpha_g + \sum_{c=1}^{C} \sum_{k=1}^{K_c} \theta_{gk} b_{gk} \left(x_{ijc} \right) + \gamma_{ig} + \delta_{ig} \varepsilon_{ig} \tag{A.8}$$

A.6 CRediT author statement

Santiago Silva: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Writing - Original Draft, Visualization. **Neil Oxtoby:** Resources, Data Curation, Writing - Review & Editing. **Andre Altmann:** Supervision, Validation, Writing - Review & Editing. **Marco Lorenzi:** Supervision, Methodology, Validation, Writing - Review & Editing.