When to harmonize? Evaluating Stage-Specific Harmonization in Federated Brain Age Estimation

Tanurima Halder*
Trust Lab
Indian Institute of Technology
Bombay
haldertanurima98@gmail.com

Kunal Deo*
Trust Lab
Indian Institute of Technology
Bombay
deo.kunal.a@gmail.com

Nicolás Nieto
Research Center Jülich
Brain and Behaviour (INM-7)
Jülich, Germany
n.nieto@fz-juelich.de

Kaustubh R. Patil#
Research Center Jülich
Brain and Behaviour (INM-7)
Jülich, Germany,
k.patil@fz-juelich.de

Kshitij Jadhav#
Koita Centre for Digital Health
Indian Institute of Technology
Bombay, dr.kshitij31@gmail.com

***Abstract*—Federated learning (FL) is a promising solution for healthcare Artificial Intelligence (AI), striking a balance between patient privacy and the need for diverse datasets. FL enables collaborative model training across institutions, preserving confidentiality and advancing clinical tasks such as diagnosis and treatment planning. However, a key challenge in this setting is the inherent heterogeneity of medical datasets acquired in different institutions, which can undermine the generalizability and performance of the model. This issue is particularly pronounced in neuroimaging applications, such as Magnetic resonance imaging (MRI), where site-specific biases arise from variations in scanner hardware, acquisition protocols, and preprocessing pipelines. These differences introduce non-biological variability that can jeopardize downstream analyses and model training. To remove the effect of sites, harmonization techniques are essential tools to improve robustness and reliability. Harmonization techniques are usually applied at the feature level; however, given the limited access to the data possessed by FL schemes, feature-level harmonization may not be enough to remove site effects. In this work, we propose two complementary harmonization strategies within the FL framework: (1) the traditional feature harmonization, by applying ComBat to directly correct the MRI-derived features; and (2) gradient harmonization, which aligns local model updates, particularly the gradients of fully connected layers, across sites to mitigate inter-site distributional shifts before global aggregation. Together, these approaches aim to improve cross-site consistency and improve the model's overall performance in federated medical imaging tasks.**

***Index Terms*—Federated Learning, Brain Age Prediction, Brain MRI, Gradient Harmonization, Feature Harmonization.**

*,#These authors contributed equally to this work.

I. Introduction

Traditional centralized Artificial Intelligence (AI) systems in healthcare require collecting and storing sensitive patient data, including electronic health records and Magnetic resonance imaging (MRI) scans, in a single location or server. This approach raises significant concerns regarding patient privacy, data ethics, and regulatory compliance [1] [2]. Given the violation of patient privacy, and the logistical impracticality of transferring all patient data to a central location, federated learning (FL) has emerged as a compelling alternative. FL allows different institutions to collaboratively train machine learning (ML) models while keeping sensitive patient data unshared, and only sharing model parameters [3].

In FL, a central server coordinates the training of multiple participating sites. The server and the sites are also known as silos in the context of medical applications, where typically a few institutions with a large amount of data are included. The central server initializes and shares a model with each silo, which is then trained with the local data. After the models are trained on the local silo data, the silo-model parameters, or gradients, are shared with the central server, avoiding the communication of sensitive patient information [3] [4]. This privacy-preserving approach has been successfully applied in healthcare applications such as electronic health records [5] and segmentation of medical images [6], among others. In several cases, FL achieves performance levels on par with centralized

training [7].

However, a major challenge in FL is the presence of non-independent and non-identically distributed (non-IID) data across sites [8], which conflicts with a fundamental assumption of ML. In the context of MRI, these distributional shifts are presented in each site and can arise from differences in patient populations, imaging protocols [9], acquisition settings [10] [9], preprocessing pipelines [11], and other site-specific characteristics [12]. These systematic differences between sites, called effects of site, are particularly problematic in neuroimaging because they significantly distort image properties [10] [13]. These site-specific shifts, always presented in federated settings, represent a key problem with FL models' convergence and generalization [14] [8].

To mitigate these challenges, different harmonization techniques can be employed. Harmonization refers to the process of removing site-effects while preserving meaningful biological signals [13] [15] [14] [9]. In the work we are presenting, we have utilized neuroHarmonize (a ComBat-based method) [9] in two different federated learning setups. This is done to reduce site-effects and enable better generalization of the ML model for the task of brain-age prediction.

The goal of this paper is to compare two different harmonization methodologies (Feature Harmonization and Gradient Harmonization) to determine which one yields better overall performance while preserving data privacy.

II. Related Work

Federated Learning has shown great promise in the field of healthcare [2]. It has been used for various applications in the medical field, such as multiomics [16], electronic patient records [5], and medical imaging [17].

There have been multiple efforts in the removal of site effects from MRI data, with neuroHarmonize being a popular method that models the data as a combination of biological information, site effects, and residuals [9]. FedHarmony [18] combines the usage of harmonization to reduce site-specific effects in a Federated setup. While these methods succeeded in removing the influence of site effects from the images, they only harmonized features extracted from images and did not explore the harmonization of model gradients [10].

The term "gradient harmonization" was previously introduced in the context of unsupervised domain adaptation by [19]. However, to the best of our knowledge, this approach has not yet been explored in the context of FL. In this work, we propose a methodology for gradient harmonization focused on its application within a federated learning framework. We explore the impact of feature-level harmonization, gradient-level harmonization, and the combination of both to mitigate the impact of site effects in a brain age prediction application developed in an FL setup.

III. Our Contributions

We propose and evaluate two complementary harmonization strategies within an FL framework for MRI-based applications. First, a Feature Harmonization, where a copy of the central data (assumed public) is shared with each site, and the local data is harmonized using ComBat-based harmonization. Age was not used as a covariate as it also represents our target, and ComBat needs it at test time, if provided in training time, unavoidably incurring in data leakage [20].

Second, Gradient Harmonization was introduced through a unified federated learning framework. The model is trained across multiple sites without sharing raw data. The global model is shared and trained locally at each site, after which site-correction is applied, using ComBat-based harmonization, to the gradients from the locally trained models before aggregation into the global model, helping the final central model updates to be free from domain shifts caused by site-effects.

By combining these different harmonization levels (feature and gradient), our approach aims to address site heterogeneity in a flexible and modular way, improving both the reliability and performance of federated models in neuroimaging.

IV. Methodology

A. *Dataset and Preprocessing*

We conducted experiments on structural MRI data collected from three publicly available datasets, each representing a separate data silo: The Enhanced Nathan Kline Institute (eNKI) [21], Cambridge Centre for Ageing Neuroscience (CamCAN) [22], and the Southwest University Adult Lifespan Dataset (SALD) [23]. These datasets encompass a broad age range, with subjects aged between 18 and 88 years. For all MRI images, gray matter volume (GMV) was extracted using the Computational Anatomy Toolbox (CAT) version 12.8 [24]. The resulting cortical voxel GMV was aggregated using 1000 cortical regions defined in the Schaefer atlas [25]. Additionally, 42 subcortical features and 31 cerebellum features obtained from CAT were included, giving a total of 1073 features.

B. *Loss Functions*

*1) **Primary Loss (L_1 Loss):*** We use the L_1 loss to minimize the mean absolute error (MAE) between predictions and ground truth values, promoting pointwise accuracy in regression. While effective as a primary loss,

it does not account for distributional alignment, which may lead to suboptimal performance in cases of non-IID target distributions. To prevent the silo models from over-training on their own data and forgetting previous information, further additions were made to the loss functions.

*2) **Proximal Weight Regularization (FedProx)**:* This loss function penalizes the silo weights w for straying too far from the central weights w_g during local training in a federated setup.

$$\mathcal{L}_{\text{prox}} = \frac{\mu}{2}\|w - w_g\|_2^2$$

FedProx adds a quadratic "proximal" term to each silo's local objective, penalizing deviations from the global model. This anchoring (i) prevents local updates from drifting too far during Stochastic Gradient Descent (SGD), (ii) reduces variance among silos updates caused by non-IID data or uneven compute budgets, and (iii) yields more stable convergence with better worst-case guarantees under heterogeneity. As shown in [26], FedProx converges faster and achieves higher accuracy than FedAvg when silo data distributions or resources are highly non-uniform.

*3) **Gaussian KL Divergence Loss**:* To regularize the latent representation and encourage the model's predicted distribution to match a target Gaussian in both mean and variance, we employ the Kullback–Leibler (KL) divergence between two univariate normal distributions. Formally, given two univariate Gaussian distributions:

Let the predicted distribution p be $\mathcal{N}(\mu_p, \sigma_p^2)$, and the target distribution t be $\mathcal{N}(\mu_t, \sigma_t^2)$. The KL divergence from the predicted to the target distribution is given by [7]:

$$\mathrm{KL}(p \,\|\, t) = \frac{1}{2}\left[\log\left(\frac{\sigma_t^2}{\sigma_p^2}\right) + \frac{\sigma_p^2 + (\mu_p - \mu_t)^2}{\sigma_t^2} - 1\right] \tag{1}$$

This loss term penalizes predictions that are either overconfident (σ_p^2 too small) or underconfident (σ_p^2 too large), promoting calibrated uncertainty estimates in the model's outputs.

*4) **Maximum Mean Discrepancy Loss**:* To improve performance beyond the baseline MAE loss, we add the Maximum Mean Discrepancy (MMD) loss—a nonparametric, kernel-based measure of distributional alignment. Unlike parametric metrics (e.g., assuming Gaussianity), MMD compares two distributions P and Q without explicit density estimation. MMD operates by embedding the distributions into a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ via a feature mapping $\phi(\cdot)$ associated with a positive-definite kernel function $k(x, y) = \langle \phi(x), \phi(y) \rangle$. The squared MMD between P and Q is then defined as the squared distance between their mean embeddings in $\mathcal{H}$ [27]:

$$\mathrm{MMD}^2(P, Q) = \|\mathbb{E}_{x\sim P}[\phi(x)] - \mathbb{E}_{y\sim Q}[\phi(y)]\|_{\mathcal{H}}^2$$

By minimizing MMD^2, the model's predicted-age (or feature) distribution is forced to match the target age distribution nonparametrically, capturing higher-order moments beyond mean and variance. Adding the MMD term yields more accurate and generalizable brain-age estimates across cohorts [27].

V. Model Architecture

We used a 4-layer fully connected neural network (FCNN) for regression-based brain age prediction. The network maps an input feature vector extracted from regional neuroimaging measures to a single continuous age estimate. The FCNN was designed to promote stable training and good generalization.

Input Layer: The input to each model is a d-dimensional feature vector, 1073 in our experiments (1000 GMV ROI values + 73 subcortical regions).

Hidden Layers 3 hidden layers were used following the structure:

- Linear (fully connected) transformation
- Batch Normalization (to reduce internal covariate shift)
- LeakyReLU activation (negative $slope = 0.01$)
- Dropout (probability $p = 0.2$)

The combination of batch normalization and LeakyReLU stabilizes gradient flow even when network depth increases, while dropout acts as a regularizer to discourage co-adaptation among neurons.

Output Layer: Each model's final layer is a single neuron with a linear activation function—appropriate for predicting a continuous target (chronological age in years).

VI. Evaluated Harmonization Strategies

To evaluate the impact of harmonization strategies on brain age prediction in a multi-silo setting, we designed a series of experiments encompassing individual, centralized, and federated training paradigms. The experiments are categorized into baseline and harmonization-based settings, as described below:

A. *Decentralized Training (Single-site)*

Each silo trains its own ML model using only local data. This fully decentralized setup ensures privacy but lacks cross-silo knowledge integration.

B. *Centralized Training (Pooled)*

All silo data are combined into a central silo, and a single model is trained. This configuration assumes full data accessibility without privacy constraints, serving as a comparison between the Federated and Centralized approaches. The trained model is then evaluated on each of the silo test sets separately.

C. *Federated Learning without harmonization*

We consider a federated learning setup comprising a central server and S data silos, each holding private data (X_s, y_s). Our goal is to establish an FL comparative baseline model that does not utilize any harmonization but maintains a federated setup (Figure 1).

On client s, with local model parameters θ_s initialized from the global model θ, we define the following composite loss to be minimized over E local epochs: The per-minibatch loss is

$$\mathcal{L}_{\text{batch}} = \mathcal{L}_{\text{MAE}} + \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{MMD}} + \mathcal{L}_{\text{prox}}, \quad (2)$$

and is minimized with gradient clipping ($\|\nabla\|_\infty \leq 1.0$) and a cosine-annealed learning-rate schedule.

After E local epochs on silo s, we compute the client's empirical gradient

$$g_s \;=\; \frac{n_s}{N}\,\nabla_\theta\,\mathcal{L}_s\big(\theta_s^E\big), \quad N = \sum_{s=1}^{S} n_s. \quad (3)$$

Clients send $\{g_s\}$ to the central server, which aggregates and clips:

$$g_{\text{agg}} = \sum_{s=1}^{S} g_s, \qquad g_{\text{agg}} \leftarrow \text{clip}(g_{\text{agg}}, -G, +G). \quad (4)$$

Using the global parameters at communication round t by θ^t, and maintain a momentum buffer v^t. The server update is:

$$v^{t+1} = \beta\, v^t + g_{\text{agg}}, \quad \theta^{t+1} \;= \theta^t - \eta\, v^{t+1}, \quad (5)$$

where β is the momentum coefficient and η the server learning rate. The new θ^{t+1} is broadcast to all clients.

1) Validation and Model Selection: Each client evaluates θ_s^e on its held-out validation set after each epoch $e \leq E$ and retains

$$\theta_s^* \;=\; \arg\min_{e \leq E} \mathcal{L}_{\text{val}}\big(\theta_s^e\big). \quad (6)$$

Similarly, the central server monitors a central validation set and selects the best global model across epochs.

D. *Feature Harmonization*

To reduce scanner-related variability between the local (s) and global (g) datasets, we applied the ComBat harmonization method as implemented in the neuroHarmonize (Python package [28]). Let $\mathbf{X}^{(s)} \in \mathbb{R}^{F \times N_s}$ and $\mathbf{X}^{(g)} \in \mathbb{R}^{F \times N_g}$ denote the feature matrices for the silo-specific and global training sets, respectively, where $F = 1073$ is the number of features and N_s, N_g are the numbers of subjects. The corresponding age vectors are $\mathbf{y}^{(s)} \in \mathbb{R}^{N_s}$ and $\mathbf{y}^{(g)} \in \mathbb{R}^{N_g}$. First, the global and local data are concatenated in each site, both in X and y, and a site vector is created, setting 1 for local and 2 for global data.

A neuroHarmonize model was trained using the concatenated feature matrix X, specifying the batch as covariates C, where the column "batch" encodes site labels. No additional categorical covariates were retained. Each silo training dataset was harmonized, and each neuroHarmonize model, which was trained at the silo, was saved for later application to the test data.

Each silo received a copy of the global model parameters $\theta^{(\text{global})}$ and trained locally for E epochs on its harmonized data $(\widehat{\mathbf{X}}^{(s)}, \widehat{\mathbf{y}}^{(s)})$. At each minibatch, the local parameters θ were updated to minimize the composite loss:

$$\mathcal{L}(\theta) = \mathcal{L}_{\ell_1}(\theta) + \frac{\mu}{2}\left\|\theta - \theta^{(\text{global})}\right\|_2^2 + \lambda_{\text{KL}}\,\mathcal{L}_{\text{KL}}(\theta) + \lambda_{\text{MMD}}\,\mathcal{L}_{\text{MMD}}(\theta)$$

Gradients were clipped for stability, and a cosine-annealing scheduler adjusted the learning rate. The model with the lowest validation loss was retained.

After local training, each silo computes its gradient:

$$g^{(s)} = \nabla_\theta \mathcal{L}^{(s)}(\theta)\big|_{\theta=\theta^{(\text{local})}}$$

These gradients were aggregated across S silos by summation:

$$g^{\text{agg}} = \sum_{s=1}^{S} g^{(s)}$$

After gradient clipping, the global model parameters were updated using momentum on the fully connected layers:

$$v_t = m\, v_{t-1} + g^{\text{agg}}, \quad \theta_t = \theta_{t-1} - \eta\, v_t,$$

where m is the momentum coefficient and η the learning rate. A held-out central validation set was evaluated each round, and the best global model was saved upon improvement. This process was repeated for a fixed number of global epochs.

After the limit of federated epochs was reached, each silo's test dataset was harmonized using the trained neuroHarmonize model and then used to make a prediction using the best-trained model.

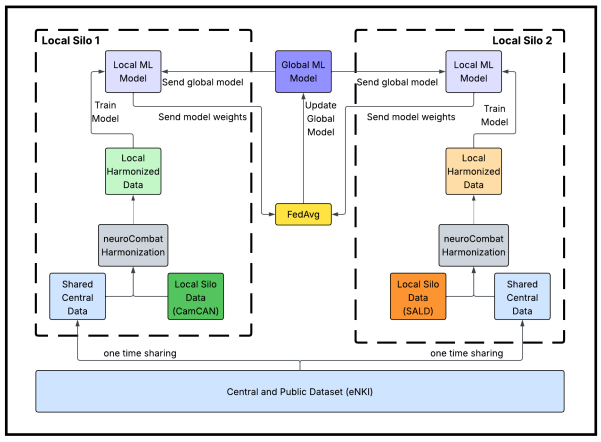

Fig. 1: The Feature Harmonization Method

E. *Local Gradient Harmonization*

We proposed a federated learning framework that integrates distributional harmonization of gradients with regularized local training, mitigating site-related biases while preserving privacy (Figure 2). This approach consists of three main components:

1) Local Silo Optimization: Each site (silo) s holds its own dataset $\{(x_i^s, y_i^s)\}_{i=1}^{n_s}$. We denote by θ the parameters of the global model, which are broadcast to each silo at the start of each federated round. Within silo s, a local copy was created $\theta \to \theta^s$ and perform T epochs of gradient-based training, minimizing the composite loss:

$$\mathcal{L}^s_{\text{local}}(\theta^s) = \frac{1}{|B|} \sum_{(x,y)\in B} \underbrace{|f_{\theta^s}(x) - y|}_{\ell_1} + \underbrace{\frac{\mu}{2}\|\theta^s - \theta\|_2^2}_{\text{proximal term}} + \lambda_{\text{KL}}\,\mathcal{L}_{\text{KL}}(\theta) \quad + \lambda_{\text{MMD}}\,\mathcal{L}_{\text{MMD}}(\theta),$$

At each epoch, the total loss was back-propagated, applied gradient clipping to norm G, and updated θ^s via SGD with momentum. After each epoch t, we recorded the full gradient vector $\nabla\mathcal{L}^s_{\text{local}}(\theta^s)$ for each layer.

2) Harmonization of Silo Gradients: We treat each recorded gradient vector for layer l as a sample labeled by silo identity.

We train and apply a neuroHarmonize model to remove site-related effect. The harmonized gradients were then averaged per-silo and then aggregated.

3) Global Model Update: Harmonized gradients $\widehat{g}_l$ for each fully connected layer l are used in a momentum-SGD update:

$$v_l \leftarrow m\,v_l + \widehat{g}_l, \quad \theta_l \leftarrow \theta_l - \eta\,v_l \tag{7}$$

where v_l is the velocity buffer (initialized to zero), m the momentum coefficient, and η the learning rate. This framework combines proximal federated learning, distributional regularizers (KL, MMD), and statistical harmonization of gradients—tailored for multi-site neuroimaging-based age prediction.

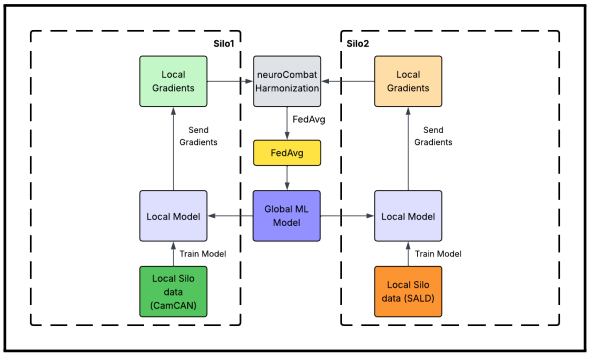

Fig. 2: Local Gradient Harmonization

F. *Dual Gradient Harmonization*

1) Gradient Harmonization: To mitigate domain-specific biases arising between silo (local) updates and centralized (global) updates, the gradient distributions per layer were harmonized using a neuroHarmonize model (Figure 3). Let $\mathbf{G}_{\text{L}}^e(\ell)$ and $\mathbf{G}_{\text{G}}^e(\ell)$ denote the vectorized gradients for layer ℓ at epoch e from the local and global models, respectively. We stack these across E epochs to form two matrices of size $\mathbb{R}^{P_\ell \times E}$, where P_ℓ is the number of parameters in layer ℓ. We then concatenate the local and global matrices to form:

$$\mathbf{X}_\ell = [\mathbf{G}_{\text{L}}^e(\ell)\,|\,\mathbf{G}_{\text{G}}^e(\ell)] \in \mathbb{R}^{P_\ell \times 2E},$$

and construct a batch-covariate vector indicating **local** and **global**. Applying neuroHarmonize yields a harmonized matrix $\widetilde{\mathbf{X}}_\ell$ with site effects removed.

Finally, we recover two harmonized gradient vectors by averaging each half of $\widetilde{\mathbf{X}}_\ell$ along the epoch dimension and reshaping them back to their original parameter tensor shapes:

$$\widetilde{\mathbf{G}}_{\text{L}}(\ell) = \frac{1}{E}\sum_{e=1}^{E} \widetilde{\mathbf{X}}_\ell[:, e], \quad \widetilde{\mathbf{G}}_{\text{G}}(\ell) = \frac{1}{E}\sum_{e=E+1}^{2E} \widetilde{\mathbf{X}}_\ell[:, e].$$

The final aggregated gradient for layer ℓ is computed as:

$$\widehat{\mathbf{G}}(\ell) = \widetilde{\mathbf{G}}_{\text{L}}(\ell) + \widetilde{\mathbf{G}}_{\text{G}}(\ell).$$

2) Local-Global Training Framework: As in the standard FL framework, the training alternates between *local* and *global* phases. Starting from a shared global model M, we perform the following steps:

3) 1. Local Model Update: Each silo s trains a local model M_s on its private data set $(\mathbf{X}_s, \mathbf{y}_s)$ for E epochs while incorporating proximal, KL divergence, and MMD penalties to align with the global model per-batch losses.

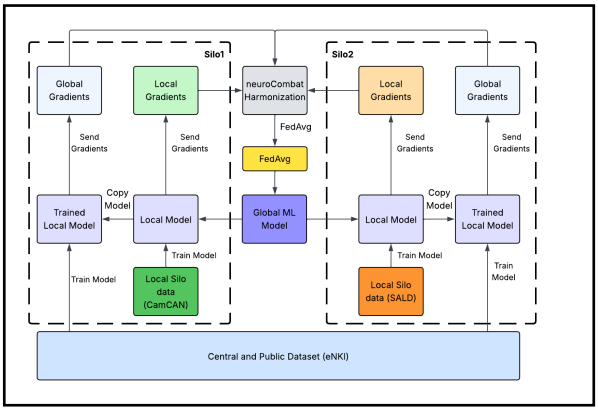

Fig. 3: Dual Gradient Harmonization

Gradients are backpropagated, with gradient-norm clipping applied for stability.

2. Gradient extraction and harmonization: After each local epoch, we extract $\mathbf{G}_{\mathrm{L}}^{e}$ for each layer. Once local training is completed, we harmonize these with concurrently computed global gradients $\{\mathbf{G}_{\mathrm{G}}^{e}\}$ as described in Section VI-F1.

3. Global Model Update: The aggregated harmonized gradients $\{\widehat{\mathbf{G}}(\ell)\}$ from all silos are summed to form the global update. We apply a momentum-SGD update:

$$\mathbf{v}_t(\ell) = m\,\mathbf{v}_{t-1}(\ell) + \widehat{\mathbf{G}}(\ell), \quad \boldsymbol{\theta}_t(\ell) \leftarrow \boldsymbol{\theta}_{t-1}(\ell) - \eta\,\mathbf{v}_t(\ell)$$

This framework harmonizes gradient distributions between 'local' and 'global' batches; we explicitly remove systematic site effects, which we expect to produce better cross-site generalization while preserving the private information of each silo.

VII. Experimental Setup

We utilized three publicly available datasets, CamCAN, SALD, and eNKI, each of which was split into training and test subsets to produce feature matrices X and corresponding age labels y.

To ensure consistency across all experiments and account for age-related variability, we adopted an age-stratified 70/30 train-test split within each silo. Specifically, 70% of the data from each site was allocated to the training set and the remaining 30% was held out for testing, while maintaining the age distribution across both subsets. This stratified split was performed once and kept fixed throughout all subsequent experiments described in this study.

In our federated learning (FL) framework, CamCAN and SALD acted as local silos, while eNKI served as the central server dataset. Before initiating the federated optimization, the global model was "warmed up" by training for 20 epochs on the centralized eNKI training split. Following this, federated training proceeded for 200 global communication rounds, with each client performing 15 local epochs per round.

Across all training phases, we used a fixed learning rate of 1×10^{-3}, a dropout rate of 0.2, momentum of 0.8, and gradient clipping with a maximum norm of 1.0. The loss-term weights for proximal, KL-divergence, and MMD regularization were set to 0.05,0.3, and 0.5, respectively.

All experiments were conducted on a Linux server equipped with four NVIDIA RTX A6000 GPUs (48 GB GDDR6 each, running CUDA 12.7 with driver 565.57.01). During training and evaluation, workloads were pinned to a GPU with 49.140 MiB total memory, of which approximately 10.225 MiB were used per run. Model implementation was carried out in PyTorch 2.4.0 on a single RTX A6000 to ensure reproducibility and to limit cross-practice interference.

VIII. Results

A. Harmonization strategies comparison

We report mean absolute errors on the held-out test sets of CamCAN, SALD, and eNKI under four conditions: (i) the best-performing single-silo model, (ii) the best centralized global model, (iii) the federated global model without harmonization (baseline), and (iv) three harmonization-enhanced federated variants [Table I]. In addition to absolute MAE values, we quantify both absolute and relative improvements of each harmonization method over the unharmonized federated baseline.

We compare the test MAE on each of the three datasets (CamCAN, SALD, eNKI) for all the training and harmonization strategies, with lower MAE indicating better performance. Table II and Table III summarize the MAE of the silo-specific and global best models, respectively, for each harmonization and training scheme.

*a) **Local (Silo-Level) Performance Analysis**:* When examining silo-specific training (TableII), harmonization yields consistent but silo-dependent improvements over the decentralized (Single-Site) baseline:

CamCAN: Feature-level harmonization achieves the lowest MAE (5.392), improving compared with Single-Site (5.632). No-H, Local-H, and Dual-H all underperform relative to Single-Site, indicating that CamCAN benefits most from harmonizing feature distributions before training.

SALD: Dual-H produces the best MAE (4.610), substantially better than Single-Site (5.080). No-H also reduces error (4.755), whereas Feature-H and Gradient-H remain close to or worse than Single-Site, demonstrating

TABLE I: *Test MAE results before and after different harmonization strategies across CamCAN, SALD, and eNKI test sets.*

(a) No-Harmonization

Silo	CamCAN	SALD	eNKI	Avg
Best silo model	6.87	4.75	4.66	5.43
Best global model	7.08	4.61	4.66	5.52
Global federated model	6.86	4.77	4.74	5.46

(b) Feature-Level Harmonization (neuroHarmonize)

Silo	CamCAN	SALD	eNKI	Avg
Best silo model	5.39	5.10	4.56	5.02
Best global model	5.76	4.24	4.56	4.85
Global federated model	5.39	5.10	4.75	5.08

(c) Local Gradient Harmonization

Silo	CamCAN	SALD	eNKI	Avg
Best silo model	6.79	5.03	4.55	5.40
Best global model	7.25	4.48	4.55	5.43
Global federated model	6.78	5.04	4.96	5.59

(d) Dual Gradient Harmonization

Silo	CamCAN	SALD	eNKI	Avg
Best silo model	6.64	4.61	4.68	5.31
Best global model	7.48	4.74	4.68	5.63
Global federated model	6.64	4.62	4.72	5.33

TABLE II: *Silo-specific model MAE by method (lower is better). Single-Site one ML model is trained on single site; No-Harmonization FL with no harmonization; Feature-H, Gradient-H, Dual-H.*

Method	CamCAN	SALD	eNKI	Avg
Single-Site	5.632	5.080	4.796	5.169
No-Harmonization	6.874	4.755	4.657	5.429
Feature-H	5.392	5.104	4.563	5.020
Gradient-H	6.788	5.031	4.547	5.455
Dual-H	6.638	4.610	4.685	5.311

TABLE III: *Global (pooled) model MAE by method. (Pooled) is the centralized baseline; other rows are models trained with harmonization.*

Method	CamCAN	SALD	eNKI	Avg
Pooled	6.779	4.864	4.579	5.4073
No-Harmonization	6.862	4.774	4.738	5.458
Feature-H	5.386	5.103	4.752	5.080
Gradient-H	6.784	5.037	4.959	5.593
Dual-H	6.639	4.616	4.716	5.3236

that gradient-level harmonization (especially Dual-H) is most effective for SALD.

eNKI: Gradient-H yields the lowest MAE (4.547) compared to Single-Site (4.796). Feature-H (4.563) and No-H (4.657) also improve on Single-Site, but to a lesser extent. Thus, eNKI favours a local gradient harmonization during silo training.

*b) **Global (Pooled) Performance Analysis**:* For the globally trained (Pooled) model (TableIII), harmonization again has mixed effects:

CamCAN: Feature-H reduces MAE from 6.779 (Pooled) to 5.386, a substantial gain that even outperforms its silo-specific models. Other global harmonization methods (Gradient-H, Dual-H) achieve results similar to or slightly worse than the Pooled baseline.

SALD: Dual-H yields the best pooled MAE (4.616), slightly better than Pooled (4.864) and No-H (4.774). In contrast, Feature-H and Gradient-H degrade SALD performance relative to Polled, indicating that only dual-gradient alignment benefits SALD in a pooled setting.

eNKI: The Pooled baseline remains the best (4.579). None of the harmonised global models outperforms Pooled, and some (Gradient-H, Dual-H) even degrade MAE. Therefore, eNKI does not benefit from harmonisation when training on fully pooled data.

B. Computational timing

The total training time for each harmonisation strategy was as follows: No-H: $10.75min$; Feature-H: $21.24min$; Gradient-H: $79.43min$, and Dual-H: $164.45min$.

IX. Conclusion and Future Work

In this work, we addressed the challenge of accurately predicting brain age in a federated learning setting where MRI scans from different institutions exhibit distinct scanner-induced biases. To this end, we compared harmonization strategies: feature- gradient-level harmonization and a combination of both (Dual-H), which aligns model updates across silos. Our experiments on three publicly available datasets, CamCAN, SALD, and eNKI, showed that Dual-H achieved the lowest silo-specific MAE (e.g., 4.55 years on eNKI and 4.75 years on SALD), outperforming both the silo-only baseline (MAE $\approx$ 5.60 years) and Feature harmonization (Feature-H) (MAE $\approx$ 5.00 years). On average, feature-level harmonization showed the best across sites performance while gradient-level harmonization showed the worst performance, highlighting the importance of the complementary approach. We acknowledge several limitations: our current setup relies on GMV features rather than end-to-end MRI volumes, and we have yet to integrate formal privacy guarantees or test on clinical populations. In future work, we will extend our approach to Convolutional Neural Network (CNN)-based predictors, incorporate differential-privacy techniques to quantify privacy leakage. Overall, our findings demonstrate that gradient-level harmonization is a promising direction for federated medical AI, enhancing feature harmonization and enabling more accurate and privacy-preserving brain-age prediction. The code for reproducing the results obtained in this work is available at: https://github.com/kadcoder/IIT_Julich_Federated_Learning

X. Acknowledgment

This work was partly supported by H2020 Research Infrastructures Grant EBRAIN-Health 101058516 and the Helmholtz Imaging Platform (project BrainShapes, ZT-I-PF-4-062). Also, by the MODS project funded from the program "Profilbildung 2020" (grant no. PROFILNRW-2020-107-A), an initiative of the Ministry of Culture and Science of the State of Northrhine Westphalia. Along with that, the authors also thank Trust Lab, IIT Bombay for supporting this research.

References

[1] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," *Nature Machine Intelligence*, vol. 2, no. 6, pp. 305–311, 2020.

[2] N. Rieke, J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, S. Bakas, M. Galtier, B. Landman, K. H. Maier-Hein *et al.*, "The future of digital health with federated learning," *NPJ Digital Medicine*, vol. 3, no. 1, p. 119, 2020.

[3] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Singh and J. Zhu, Eds., vol. 54. PMLR, 20–22 Apr 2017, pp. 1273–1282. [Online]. Available: https://proceedings.mlr.press/v54/mcmahan17a.html

[4] M. J. Sheller, G. A. Reina, B. Edwards, J. Martin, and S. Bakas, "Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data," *Scientific Reports*, vol. 10, no. 1, p. 12598, 2020.

[5] C. Ju, R. Zhao, J. Sun, X. Wei, B. Zhao, Y. Liu, H. Li, T. Chen, X. Zhang, D. Gao, B. Tan, H. Yu, C. He, and Y. Jin, "Privacy-preserving technology to help millions of people: Federated prediction model for stroke prevention," 2020. [Online]. Available: https://arxiv.org/abs/2006.10517

[6] S. Galada, T. Halder, K. Deo, R. P. Krish, and K. Jadhav, "Prism: Privacy-preserving inter-site mri harmonization via disentangled representation learning," 2024. [Online]. Available: https://arxiv.org/abs/2411.06513

[7] S. Garst, J. Dekker, and M. Reinders, "A comprehensive experimental comparison between federated and centralized learning," *Database*, vol. 2025, p. baaf016, 03 2025. [Online]. Available: https://doi.org/10.1093/database/baaf016

[8] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[9] J.-P. Fortin, N. Cullen, Y. I. Sheline, W. D. Taylor, I. Aselcioglu, P. A. Cook, P. Adams, C. Cooper, M. Fava, P. J. McGrath *et al.*, "Harmonization of cortical thickness measurements across scanners and sites," *NeuroImage*, vol. 167, pp. 104–120, 2018.

[10] N. K. Dinsdale, M. Jenkinson, and A. I. T. Namburete, "Unlearning scanner bias for mri harmonisation," *Medical Image Analysis*, vol. 67, p. 101841, 2021.

[11] G. Antonopoulos, S. More, F. Raimondo, S. B. Eickhoff, F. Hoffstaedter, and K. R. Patil, "A systematic comparison of vbm pipelines and their application to age prediction," *NeuroImage*, vol. 279, p. 120292, 2023.

[12] V. Komeyer, N. Nieto, S. B. Eickhoff, F. Raimondo, and K. R. Patil, "Overview of challenges in brain-based predictive modeling: Towards meaningful predictive insights," *Biological Psychiatry*, 2025.

[13] R. Pomponio, G. Erus, M. Habes, J. Doshi, D. Srinivasan, E. Mamourian, V. Bashyam, I. M. Nasrallah, T. D. Satterthwaite, Y. Fan *et al.*, "Harmonization of large mri datasets for the analysis of brain imaging patterns throughout the lifespan," *NeuroImage*, vol. 208, p. 116450, 2020.

[14] Y. Yan, S. Gupta, X. Song, S. Shah, and C. Zhang, "Fedharmonizer: Semi-supervised federated learning with feature and gradient harmonization for medical image analysis," *IEEE Transactions on Medical Imaging*, 2023.

[15] A. A. Chen, J. C. Beer, N. J. Tustison, P. A. Cook, R. T. Shinohara, and H. Shou, "Harmonizing functional connectivity reduces scanner effects in multi-site fmri data," *NeuroImage*, vol. 241, p. 118403, 2021.

[16] B. P. Danek, M. B. Makarious, A. Dadu, D. Vitale, P. S. Lee, A. B. Singleton, M. A. Nalls, J. Sun, and F. Faghri, "Federated learning for multi-omics: A performance evaluation in parkinson's disease," *Patterns*, vol. 5, no. 3, p. 100945, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666389924000448

[17] M. Adnan, S. Kalra, J. C. Cresswell, G. W. Taylor, and H. R. Tizhoosh, "Federated learning and differential privacy for medical image analysis," *Sci. Rep.*, vol. 12, no. 1, p. 1953, Feb. 2022.

[18] N. K. Dinsdale, M. Jenkinson, and A. I. Namburete, "Fedharmony: Unlearning scanner bias with distributed data," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 695–704.

[19] F. Huang, S. Song, and L. Zhang, "Gradient harmonization in unsupervised domain adaptation," 2024. [Online]. Available: https://arxiv.org/abs/2408.00288

[20] N. Nieto, S. B. Eickhoff, C. Jung, M. Reuter, K. Diers, M. Kelm, A. Lichtenberg, F. Raimondo, and K. R. Patil, "Impact of leakage on data harmonization in machine learning pipelines in class imbalance across sites," 2024. [Online]. Available: https://arxiv.org/abs/2410.19643

[21] K. B. Nooner, S. J. Colcombe, R. H. Tobe, M. Mennes, M. M. Benedict, A. L. Moreno, L. J. Panek, S. Brown, S. T. Zavitz, Q. Li *et al.*, "The nki-rockland sample: a model for accelerating the pace of discovery science in psychiatry," *Frontiers in neuroscience*, vol. 6, p. 152, 2012.

[22] M. A. Shafto, L. K. Tyler, M. Dixon, J. R. Taylor, J. B. Rowe, R. Cusack, A. J. Calder, W. D. Marslen-Wilson, J. Duncan, T. Dalgleish *et al.*, "The cambridge centre for ageing and neuroscience (cam-can) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing," *BMC neurology*, vol. 14, no. 1, p. 204, 2014.

[23] D. Wei, K. Zhuang, L. Ai, Q. Chen, W. Yang, W. Liu, K. Wang, J. Sun, and J. Qiu, "Structural and functional brain scans from the cross-sectional southwest university adult lifespan dataset," *Scientific data*, vol. 5, no. 1, pp. 1–10, 2018.

[24] P. Zheng, Y. Zhu, Z. Zhang, Y. Hu, and A. Schmeink, "Federated learning in heterogeneous networks with unreliable communication," in *2021 IEEE Globecom Workshops (GC Wkshps)*, 2021, pp. 1–6.

[25] A. Schaefer, R. Kong, E. M. Gordon, T. O. Laumann, X.-N. Zuo, A. J. Holmes, S. B. Eickhoff, and B. T. Yeo, "Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri," *Cerebral cortex*, vol. 28, no. 9, pp. 3095–3114, 2018.

[26] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proceedings of the USENIX Conference on Annual Technical Conference (USENIX ATC)*, 2020, pp. 43–57.

[27] A. Zapaishchykova, D. Tak, Z. Ye, K. X. Liu, J. Likitlersuang, S. Vajapeyam, R. B. Chopra, J. Seidlitz, R. A. Bethlehem, L. B. C. Consortium, R. H. Mak, S. Mueller, D. A. Haas-Kogan, T. Y. Poussaint, H. J. Aerts, and B. H. Kann, "Diffusion deep learning for brain age prediction and longitudinal tracking in children through adulthood," *medRxiv*, 2023. [Online]. Available: https://www.medrxiv.org/content/early/2023/10/23/2023.10.17.23297166

[28] J.-P. Fortin, N. Parker, Z. I. Tunçay, E. M. Lake, M. Reuter, T. Santini, C. M. Crainiceanu, B. S. McEwen, D. D. Dougherty, M. M. Trivedi *et al.*, "Harmonization of multi-site neuroimaging data," *Neuroimage*, vol. 161, pp. 142–152, 2017.