Exploring Explainability in Federated Learning: A Comparative Study on Brain Age Prediction

Giuseppe Fasano(✉), Angela Lombardi, Antonio Ferrara, Eugenio Di Sciascio, and Tommaso Di Noia

Department of Electrical and Information Engineering, Politecnico di Bari, 70125 Bari, BA, Italy
{giuseppe.fasano,angela.lombardi}@poliba.it

Abstract. Predicting brain age from neuroimaging data is increasingly used to study aging trajectories and detect deviations linked to neurological conditions. Machine learning models trained on large datasets have shown promising results, but data privacy regulations and the challenge of sharing medical data across institutions limit the feasibility of centralized training. Federated Learning (FL) offers a solution by allowing multiple sites to collaboratively train a model without sharing raw data. However, it remains unclear how FL affects the explainability of these models, raising concerns about the consistency and reliability of their predictions.

In this study, we analyze the consistency of model explanations between centralized and federated training paradigms. Using DeepSHAP we compare feature attributions in brain age prediction models trained on the multi-site, publicly available OpenBHB dataset. We examine the impact of how data is distributed across sites (IID vs. non-IID), the number of sites participating per training round (sampling rate), and different FL aggregation methods (FedAVG, FedProx).

Our findings show that federated models provide different explanations compared to centralized models, even when trained on the same data and task. Non-IID data distributions reduce the consistency of explanations, while including a larger number of sites per training round improves stability. Interestingly, some federated models trained on non-IID data capture biologically meaningful patterns of brain aging even more effectively than centralized models. These results suggest that careful choices in how data is distributed and how training is conducted in FL can impact model accuracy and interpretability.

Keywords: Brain Age Prediciton · Federated Learning · eXplainable Artificial Intelligence · DeepSHAP · Non-IID Data

1 Introduction

Brain age prediction has emerged as a crucial task in computational neuroscience, leveraging machine learning techniques to estimate an individual's brain

R. Guidotti et al. (Eds.): xAI 2025, CCIS 2576, pp. 295–315, 2026.
https://doi.org/10.1007/978-3-032-08317-3_14

age from neuroimaging data [23,30]. This approach provides valuable insights into neurological health, serving as a biomarker for cognitive decline and various neurodegenerative disorders [8,22].

Traditional deep learning models for brain age prediction rely on centralized training paradigms, where large datasets are collected and processed in a single location [6]. However, this methodology is constrained by data privacy concerns, regulatory restrictions, and the challenges of aggregating multi-institutional datasets [36]. These limitations hinder the scalability and generalizability of foundational models for brain age prediction [15].

Federated Learning (FL) has emerged as a promising solution to address these challenges. FL enables decentralized model training across multiple nodes (e.g., hospitals, research centers) without requiring data sharing [33]. This approach preserves data privacy while leveraging different real-world datasets, ultimately improving model robustness and generalizability [31]. However, despite the advantages of FL, the implications of this decentralized training paradigm on explainability remain largely unexplored. Given the critical role of interpretability in AI-based medical applications, it is essential to assess whether FL models provide consistent and reliable explanations compared to their centralized counterparts.

Explainable Artificial Intelligence (XAI) methods, such as SHAP (Shapley Additive Explanations) and Grad-CAM, are increasingly employed to interpret deep learning models in neuroimaging [9,18,35]. These techniques enhance model transparency by identifying important features and decision-making patterns. However, applying XAI in federated settings presents unique challenges, including the potential variability in feature importance across distributed nodes and the effect of aggregation strategies on interpretability [26]. Understanding how FL affects explainability is crucial to ensuring trustworthiness and clinical applicability [11].

This study aims to systematically investigate the consistency of XAI outputs in centralized versus federated training paradigms, as well as to explore the consistency of the explanation among different federated configurations. Specifically, we address the following research questions (RQs):

1. RQ1: given the same task, dataset, and architecture, are XAI outputs consistent among centralized and federated models?
2. RQ2: given the same task, dataset, and architecture, are XAI outputs consistent among different configurations of the Federated Learning approach?
3. RQ3: given the same task, dataset, and architecture, are XAI outputs consistent among federated models trained on different node data distributions?
4. RQ4: do the identified features exhibit biological significance across different training paradigms?

To address these questions, we conduct a comparative evaluation of centralized and federated models trained on the task of the brain age prediction, analyzing their interpretability using state-of-the-art XAI techniques. We employ a dense convolutional architecture and the publicly available multi-site dataset by

the OpenBHB project [12]. Then, we compare centralized training with federated learning approaches under different configurations.

The organization of this paper unfolds as follows: Sect. 2 provides an overview of FL and the efforts found in literature to bridge explainability within the federated paradigm. Section 3 delves into the dataset and the pre-processing step, while Sect. 4 details the training pipelines and the analysis framework we implemented to conduct the study. In Sect. 5 we present and discuss the results of the analysis. Section 6 concludes the paper by addressing the limitations of the work and potential future directions.

2 Related Work

Federated learning is a decentralized machine learning paradigm initially proposed by Google [19,20,29] to enable collaborative model training across multiple data sources without requiring direct data sharing. By preserving data locality, federated learning mitigates privacy risks and regulatory concerns associated with centralized data aggregation [4,17]. As data privacy concerns continue to grow, this paradigm has turned into one of the most popular frameworks for privacy-oriented training of machine learning models and is nowadays applied in various privacy-sensitive domains, such as healthcare [31,32], resource-constrained environments [39], recommendation systems [2].

Federated learning operates by iteratively updating the parameters of a machine learning model through distributed optimization, where participating clients periodically perform local training on their private data to minimize a global loss function and share model updates with a central server. The latter then aggregates these updates to refine the global model before redistributing it to clients for the next training round.

A key distinction in federated learning lies between cross-device and cross-silo settings. The former typically involves a vast number of heterogeneous, resource-constrained edge devices, such as smartphones, participating in model training with intermittent availability. The latter, in contrast, focuses on a limited number of reliable and well-resourced entities, such as hospitals or research institutions, each holding substantial local datasets. This scenario is particularly relevant for medical applications, where data is collected by separate institutions on different patients, and data privacy regulations prevent its centralized collection.

While cross-silo federated learning may not suffer from device and communication heterogeneity across clients, it still presents significant statistical challenges due to data heterogeneity, wherein data across clients is non-independent and identically distributed (non-IID). This heteogeneity can manifest in several ways [38], namely i) label skewness, e.g., when one institution primarily collect data from younger patients, while another may focus on older individuals, or the same MRI characteristics are associated to different ages based on the geographical area; ii) feature skewness, e.g., when the same labels across the clients are associated to different MRI characteristics (also due to different resolutions, protocols, hardware); iii) quality skewness; iv) quantity skewness. Such

statistical heterogeneity poses fundamental optimization challenges in federated learning [1]. Due to non-uniform data distributions, local optimization objectives on individual clients may diverge from the global objective [24], leading to inconsistent convergence behavior. Clients may reach disparate local optima rather than contributing cohesively to a globally optimal model. In brain age prediction tasks, which can benefit in principle from multi-institutional collaboration while respecting privacy constraints, these issues are further exacerbated by domain-specific challenges such as differences in patient demographics, imaging modalities, and clinical annotation standards. Consequently, federated models can suffer from degraded performance compared to traditional centralized learning and, more importantly, cast doubts on the consistency of explainable AI (XAI) outputs.

Indeed, the growing importance of trustworthy AI has propelled research into XAI, particularly in high-stakes domains like healthcare where model interpretability is critical for clinical trust and decision-making. Regulatory frameworks such as the GDPR further mandate that AI-driven decisions be accompanied by human-understandable explanations. These imperatives have spurred the development of Federated Explainable AI (Fed-XAI), which aims to merge the privacy benefits of FL with the transparency of XAI. For instance, Bárcena et al. [3] introduced Fed-XAI to ensure that federated systems remain both private and interpretable, even though their approach has largely been limited to simpler, transparent models (e.g., rule-based systems and decision trees) that do not directly extend to complex modalities like images or text.

Complementary efforts have further broadened the scope of explainability within FL. Gill et al. [13] proposed TraceFL, which dynamically tracks the lineage of the global model, identifying the most influential clients for a given prediction without modifying the underlying algorithm. Meanwhile, adaptations of SHAP have been explored for federated settings: Wang [37] and Corbucci et al. [7] introduced variants that enable the attribution of feature contributions even when raw data remains decentralized. In the digital manufacturing domain, Kusiak et al. [21] developed fXAI, executing explainability algorithms sequentially on local datasets to construct globally interpretable models. In medical applications, Briola et al. [5] and studies on Parkinson's Disease stage prediction [10] have demonstrated the potential of XAI in FL, albeit under the simplifying assumption of IID data distributions.

While existing literature has made valuable strides in federated explainability, our work sets itself apart by tackling the practical challenges of realistic, non-IID scenarios—conditions that are more representative of medical applications. Rather than confining evaluations to simple or IID settings, we conduct a comprehensive comparison between centralized and federated training paradigms under various FL configurations. In doing so, we scrutinize how data heterogeneity influences the consistency of XAI outputs and investigate whether the features highlighted by these methods carry biological significance.

3 Materials

In this work, we used the publicly accessible data from the OpenBHB project[1], which collects data from 10 publicly available datasets, including IXI, ABIDE 1, ABIDE 2, CoRR, GSP, LOCALIZER, MPI-Leipzig, NAR, NPC, and RBP. The project provides the dataset already split into training and validation sets, which we used to train and test our models. The original validation partition was created using stratified sampling based on age, sex, and site [12].

The OpenBHB dataset collects N = 3984 observations of Healthy Controls (HC) from 62 sites around the world. By utilizing a multi-site international dataset, we simulated a distributed learning scenario in which each site functioned as an independent node within a federated network, enabling the implementation and evaluation of various federated learning approaches.

It is important to note that we excluded 5 of the 62 sites from the study since they were only present in the test partition. The remaining 57 sites were treated as nodes in the federated setting.

As illustrated in Fig. 1, the age distribution across sites in the training set is heterogeneous, with some sites covering a narrow age range and only a few including data from elderly participants. Additionally, the number of observations per site varies considerably, ranging from fewer than 30 to more than 700 samples.

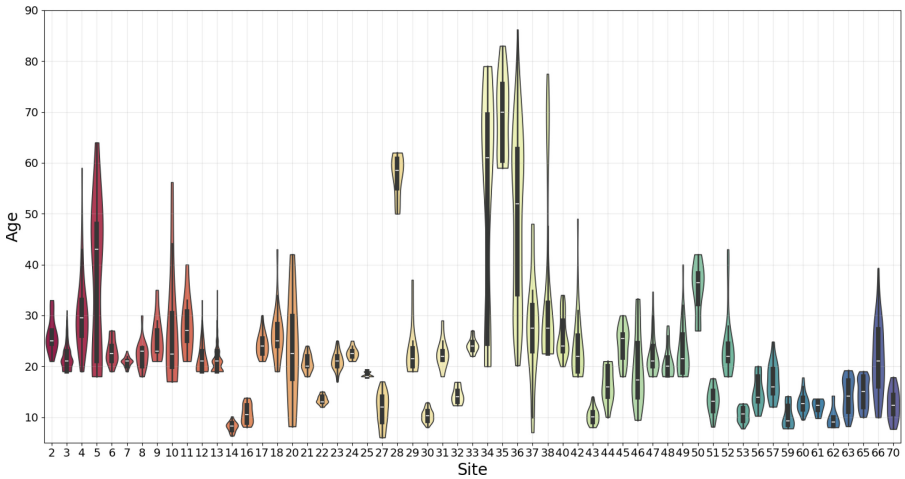


Fig. 1. Age distribution per site in the OpenBHB dataset. The labels on the x-axis refer to the site ID given by the project.

Among the different data and modalities provided by the OpenBHB project, we considered quasi-raw (minimally preprocessed) T1-weighted MRI images for

[1] https://ieee-dataport.org/open-access/openbhb-multi-site-brain-mri-dataset-age-prediction-and-debiasing.

our work, which underwent further preprocessing. The first step involved undersampling of MRI images to decrease computational demands and storage needs. We chose to reduce the dimensions of the images by half along each axis, leading to final dimensions of 91 × 109 × 91 voxels. Subsequently, z-score normalization was applied to standardize the intensity values across the dataset. For each sample, this process involved computing the mean and standard deviation of the voxel intensities and transforming each voxel's intensity value into its corresponding z-score.

Finally, two federated datasets were derived from the OpenBHB data. The first was constructed by partitioning the training and test data based on their site of origin, resulting in 57 sub-datasets. This partitioning reflects a realistic scenario in which data is non-independent and non-identically distributed (non-IID) across the nodes of the network. To investigate the impact of non-IID data on the explanations provided by federated models, we generated a second dataset using a synthetic partitioning strategy to simulate an IID distribution. This dataset was created through stratified sampling based on the age distribution of the original training set, resulting in 57 sub-datasets with comparable sizes (5657 samples per site) and similar age distributions.

4 Methods

4.1 Federated Scenario Anslysis

In this study, we examined how different implementations of the Federated Learning paradigm affect the explanations of a model trained for brain age prediction. First, we characterized the scenario to determine the most appropriate federated approach for training, considering the following three key aspects: the data distribution among nodes, the number of involved sites, and the architecture of the federated network.

The training dataset is partitioned into 57 sites, all sharing the same feature space (T1-weighted MRI images) but differing in sample space. Each site contains a distinct population with variations in age, the target variable for prediction, and no overlapping samples across sites. This condition is commonly referred to as sample-based Federated Learning or Horizontal Federated Learning (HFL) [14].

Compared to other settings, such as Vertical Federated Learning and Federated Transfer Learning, HFL offers a more straightforward training approach. In HFL, each node trains a local instance of the same model architecture using its own data. Subsequently, the trained model parameters from all nodes are collected and aggregated into a global model, which is then redistributed to all nodes. This iterative process, known as a learning round, is repeated until the global model converges.

Depending on the architecture of the federated network, the aggregation of local models can be performed either by individual nodes (decentralized FL) or a central server (centralized FL). In a real-world application of the federated paradigm for a medical network, a decentralized approach should be preferred.

Although requiring expensive communication costs and a complex implementation, a decentralized approach shows a higher level of security and data privacy than a centralized counterpart. Actually, the centralized approach relies on a server, which is vulnerable to attack and might not be trusted [28]. However, since our focus is not on developing a real-world FL system but rather on studying the impact of FL on model explanations, we opted for a centralized FL implementation for simplicity.

In the simulated network, 57 clients—one per site—communicate with a single central server responsible for model aggregation. Finally, we categorized the scenario based on the size of the node cohort. Federated Learning is typically classified as cross-silo or cross-device, depending on the number of participating clients. Cross-silo FL is used for networks with a limited number of nodes, whereas cross-device FL involves millions of clients. Given the number of nodes in our study, our scenario falls under the cross-silo FL category.

The final essential element of a federated framework is the aggregation protocol. Let $D_k = \{(x_i^k, y_i^k)\}_{i=1,\ldots,|D_k|}$ be a private collection of $|D_k|$ observations of the k-th node of a network of K clients. Each node trains a local model $f_k(x^k; w^k)$ and shares its learned parameters w^k. The aggregation algorithm is responsible for collecting the w^k and combining them into a single global model with parameters w^g. Federated Averaging (FedAVG) [29] is one of the earliest and most widely used aggregation algorithms in FL. In FedAVG, a group of $L \leq K$ nodes is randomly selected at each learning round, and their w^l parameters are weighted and averaged as follows:

$$w^g = \sum_{l=1}^{L} \frac{|D_l|}{V} w^l \tag{1}$$

Defined V as the total number of observations among clients, the contribution to the aggregated model by the l-th node is weighted by the ratio of its data volume $|D_l|$ to the global data volume V.

It is important to note that the FedAVG protocol does not specify the number of local training epochs. Since numerous local steps may cause clients to prioritize their objectives rather than minimizing the global cost function, FedProx [25] was introduced. In the FedProx protocol, a proximal term is added to the global cost function to balance the influence of local models:

$$\frac{\mu}{2} ||w^k - w^g||^2 \tag{2}$$

The proximal term acts as a regularization, which forces the local model to the global one. The parameter μ is the penalty constant, which controls the strength of the regularization, and FedProx is equal to FedAVG when $\mu = 0$.

In this study, we adopted the FedAVG strategy as it is a widely used approach for HFL applications. Additionally, we selected FedProx as a comparative method to examine the impact of regularization on the explainability of federated models.

4.2 Selection of the CNN Model

We addressed the problem of brain age prediction using a convolutional architecture, trained under different configurations to predict the chronological age of individuals from their MRI T1-weighted brain images. A DenseNet-121 [16] was chosen for the study, which we slightly modified to adapt to our task. To better accommodate the resolution of our input images, we implemented a 3D adaptation of the original 2D design, as proposed in the reference paper, resulting in a model with approximately 11 million parameters. First, a DenseNet was trained in a conventional centralized manner on the entire OpenBHB training set, and we used this model as a baseline for our analysis. Subsequently, the same architecture was employed across various FL implementations to evaluate their impact.

The choice of DenseNet-121 was moved by the results of our previous work [9], where we investigated the performance of different widely recognized CNN solutions for the brain age prediction task on the OpenBHB dataset. Our results identified the DenseNet-based model as the most effective. Following the same training setup, we trained the baseline model[2], consisting of 100 training epochs with the Stochastic Gradient Descent (SGD) optimizer and the cosine annealing scheduler with warm restarts. The scheduler was configured with an initial period of 17 epochs, which doubles after each cycle. The initial learning rate was 0.01 and the model was trained with mini-batches of size 16.

4.3 Federated Training

In this study, we explored multiple FL implementations to assess the impact of different design choices on the explainability outputs of the federated models.

Federated Learning with FedAVG on Non-IID Data. We first investigated FL frameworks based on the FedAVG protocol with random client sampling, configured to train collaborative models on non-IID data. The number of local training epochs was fixed at one, and we implemented four different settings based on varying node sampling rates: 10%, 25%, 50%, and 100%.

Next, we fixed the client sampling rate at 100% and varied the number of local epochs, considering 1, 3, and 5 local epochs.

Federated Learning with FedAVG on IID Data. We then evaluated FedAVG in a setting with IID data, fixing the number of local epochs to one and exploring different node sampling rates: 2%, 10%, 25%, 50%, and 100%. In this context, a rate of 2% corresponds to selecting only one node per learning round, a scenario expected to degenerate into conventional mini-batch training since no parameter averaging occurs.

[2] Our pre-trained DenseNet is made available at the link: https://huggingface.co/SisInfLab-AIBio/BrainAge_DenseNet.

Federated Learning with FedProx on Non-IID Data. Finally, we implemented federated training integrating the FedProx protocol on non-IID data, with μ parameter equal to 1. We fixed the sampling rate to 100% and considered 1, 3, and 5 local epochs.

Training Configuration. Beyond the centralized baseline model, we trained 14 federated models under different conditions, varying:

- data distribution (IID vs. non-IID);
- number of local epochs;
- client sampling rate;
- aggregation protocol (FedAVG vs. FedProx).

All federated models were trained for 500 learning rounds, with the SGD optimizer, an initial learning rate of 0.01, and mini-batches of size 16.

4.4 Performance Evaluation

We assessed the performance of the models using the Mean Absolute Error (MAE), calculated as:

$$MAE = \frac{1}{t}\sum_{i=1}^{t}|\hat{y}_i - y_i| \tag{3}$$

where t represents the number of samples in the test set, y_i is the actual chronological age, and $\hat{y}_i$ is the predicted brain age. MAE expresses the average error between the predicted and actual ages, so lower MAE values denote better model accuracy. MAE was also used as the training cost function for the baseline model and the federated models.

4.5 XAI Analysis

The explainability analysis in this study was conducted using DeepSHAP, a widely recognized explanation algorithm designed for convolutional neural networks (CNNs). The primary objective was to assess whether federated training, along with different federated learning configurations, influences model explanations compared to its centralized counterpart.

DeepSHAP is an adaptation of SHAP [27] (SHapley Additive exPlanations) for convolutional architectures. SHAP is a post-hoc, model-agnostic technique based on the cooperative game theory of Shapley, and it is an established XAI method of feature attribution for traditional machine learning models. This means that SHAP computes the relevance of each feature to the determination of a specific prediction. However, computer vision problems are usually solved with deep learning solutions that work on images, shifting the attribution from features to pixels or voxels, as in our case. DeepSHAP is a method of pixel (or

voxel) attribution that estimates SHAP values with a DeepLIFT-inspired approach [34]. Among XAI techniques, DeepSHAP is one of the few methods that adhere to the three principles of additivity, monotonicity, and missingness, which further justify our choice of considering DeepSHAP for the XAI analysis of this work.

To systematically compare explanations across different federated models and the baseline, we implemented a dedicated XAI pipeline, consisting of two main modules:

- Generation module that computes and processes model explanations.
- Analysis module that examines and compares XAI outputs across different models.

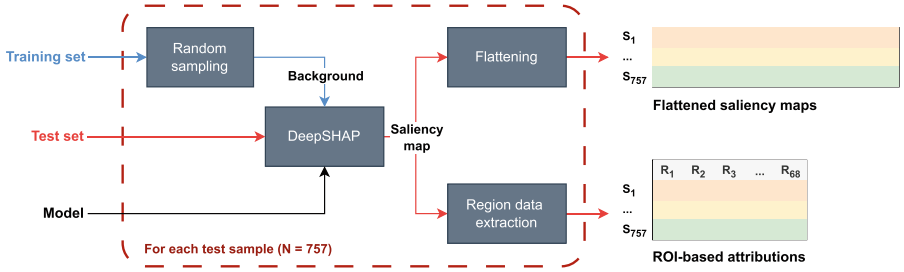


Fig. 2. Generation module of the XAI pipeline.

Generation Module. The workflow of the generation module is illustrated in Fig. 2. The first step involves computing DeepSHAP values for each of the $M = 757$ test samples. Since DeepSHAP requires both a trained model and a reference background B to estimate SHAP values, we randomly sampled $|B| = 200$ training images as the background for each computation. The adoption of a simple random sampling to collect the background is motivated by the findings of our previous study [9], in which we explored the impact of different background sampling techniques on the final DeepSHAP scores. The random sampling method led to similar explanations generated with a more sophisticated collection method, i.e., stratified sampling.

The adoption of backgrounds sampled from the overall distribution of data among nodes violates the privacy of the sites. However, the adaptation of the DeepSHAP method to the federated paradigm lies outside the scope of this work, which focuses on the investigation of the consistency of XAI outputs across models trained with a federated approach rather than a conventional centralized pipeline. Since the final federated model is computed by aggregating the local trained models, it is fair to assume that its training set is the union of all local training sets, from which we can sample a background for DeepSHAP.

For a given test sample, the DeepSHAP block outputs a saliency map, a voxel-wise attribution map with the same dimensions as the input image. The module then processes these attributions to prepare them for further analysis in the comparison module. This is achieved through two transformation steps:

- Flattening phase: converts the 3D DeepSHAP arrays into 1D vectors, resulting in 757 vectors of scores per model.
- Region data extraction phase: groups voxel contributions into 68 anatomical regions based on the Desikan-Killiany atlas. Given a 3D saliency map, the module sums the contributions within each of the 68 atlas-defined regions. The final output of this branch is an ROI-based table with 757 rows (test samples) and 68 columns (brain regions) for each model.

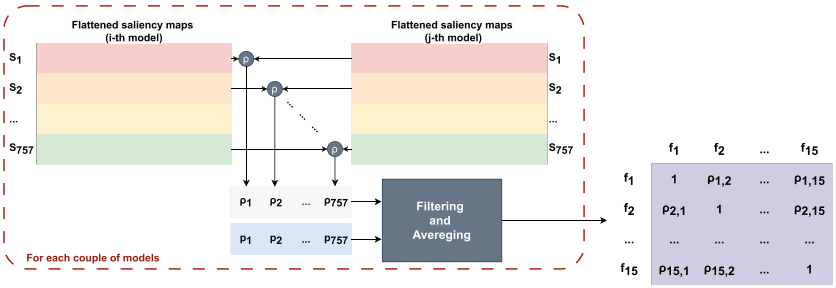
Analysis Module. The processed outputs from the generation module are analyzed through two distinct approaches:

- Inter-model analysis (Fig. 3a): investigates the similarity of explanations across models trained on the same task and dataset but under different training paradigms. To quantify the relationship between explanations from different models, we computed the Pearson correlation coefficient ρ between pairs of flattened DeepSHAP saliency maps corresponding to the same patient. Given two models f_i and f_j, we calculated ρ for the saliency maps of each patient and averaged all significant correlations (significance threshold $\alpha = 0.05$) to obtain the average correlation coefficient $\bar{\rho}_{ij}$.
- Intra-model analysis (Fig. 3b): the relationship between DeepSHAP scores and chronological age is examined to assess the biological significance of federated model explanations. Considering the ROI-based table relative to model f_i, we correlated the scores of each feature with the chronological age of patients, resulting in the vector of coefficients P_i. To determine whether all models exhibited similar relationships between region attributions and age, we conducted a Kruskal-Wallis test across the P_i distributions of all models. A Bonferroni correction was applied to account for multiple comparisons, with a significance threshold of $\alpha = 0.05$.

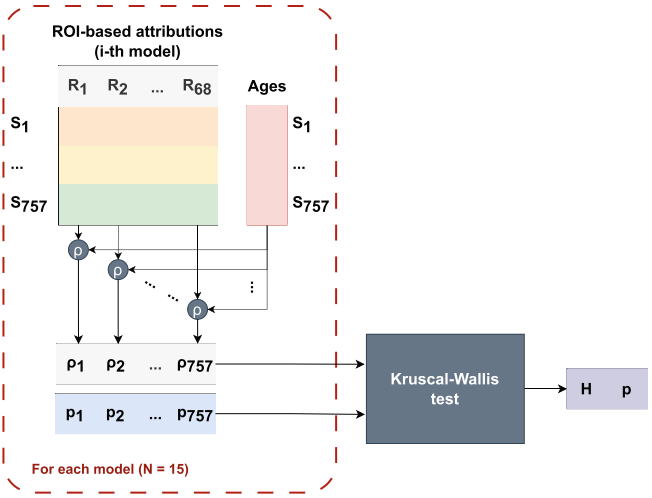
5 Results and Discussion

5.1 Performance of the Models

The evaluation results of the trained federated model on the test set are summarized in Table 1. None of the configurations achieved the performance of the baseline model ($MAE = 3.34$). However, the models trained on the IID dataset resulted in the lowest test errors among the federated models. This outcome aligns with well-documented findings in Federated Learning, where the non-IID nature of data distribution negatively impacts the performance of the final aggregated model.



(a) Inter-model analysis



(b) Intra-model analysis.

Fig. 3. Approaches of the analysis module (a) **Inter-model analysis** and (b) **Intra-model analysis**.

In the IID setting with the FedAVG protocol, the model that achieved the lowest error was the one with a 50% node sampling rate ($MAE = 4.32$). In the non-IID scenario, the best-performing model trained with FedAVG used a 25% sampling rate and 1 local epoch per learning round ($MAE = 5.81$). For FedProx-based models, the lowest error was observed in the configuration with a 100% sampling rate and 5 local epochs ($MAE = 5.86$). Interestingly, these results suggest that effective brain age prediction can be achieved in a federated setting without requiring all sites to participate in each learning round. Instead, a fraction of 50% or even fewer clients per round appears sufficient to train a performant federated model.

In the non-IID setting with a 100% sampling rate, the lowest test errors were observed for the models trained with FedAVG using 3 local epochs ($MAE = 6.20$) and FedProx using 5 local epochs ($MAE = 5.86$). These findings suggest that federated models optimized for brain age prediction may benefit from mul-

Table 1. Performance comparison of different federated models on the test set.

Dataset	Aggregator	Sampling **rate** [%]	Local **epochs**	MAE
IID	FedAVG	2	1	5.18
		10	1	6.61
		25	1	4.55
		50	1	**4.32**
		100	1	4.37
non-IID	FedAVG	10	1	8.71
		25	1	**5.81**
		50	1	7.35
		100	1	6.81
		100	3	6.20
		100	5	6.43
	FedProx	100	1	6.65
		100	3	6.72
		100	5	**5.86**

tiple local epochs, potentially improving their generalization ability despite data heterogeneity.

Finally, we analyze the results obtained on the IID dataset for the configuration with a 2% sampling rate ($MAE = 5.18$). Although this setup resembles traditional mini-batch training, it is important to note that the model is exposed to only 57 training samples per learning epoch. Given that the training set contains over 3000 samples and the model is trained for 500 learning rounds, the effective training duration is equivalent to fewer than 10 epochs in a centralized setting. This explains the relatively high error compared to other configurations in the IID scenario.

5.2 XAI Analysis

Inter-Model Analysis. In this study, we examined the consistency of model explanations between centralized and federated training across different FL configurations. To achieve this, we computed DeepSHAP scores for the same test set across all trained federated models and the centralized model.

To quantify the similarity of explanations, we calculated the Pearson correlation coefficients between the attributions of each model pair. The average and standard deviation of these correlation coefficients are summarized in the heatmaps in Fig. 4. Notably, only statistically significant correlations were considered when computing the averages. Across all model pairs, the proportion of significant correlations was consistently high, never falling below 93% of the total number of coefficients.

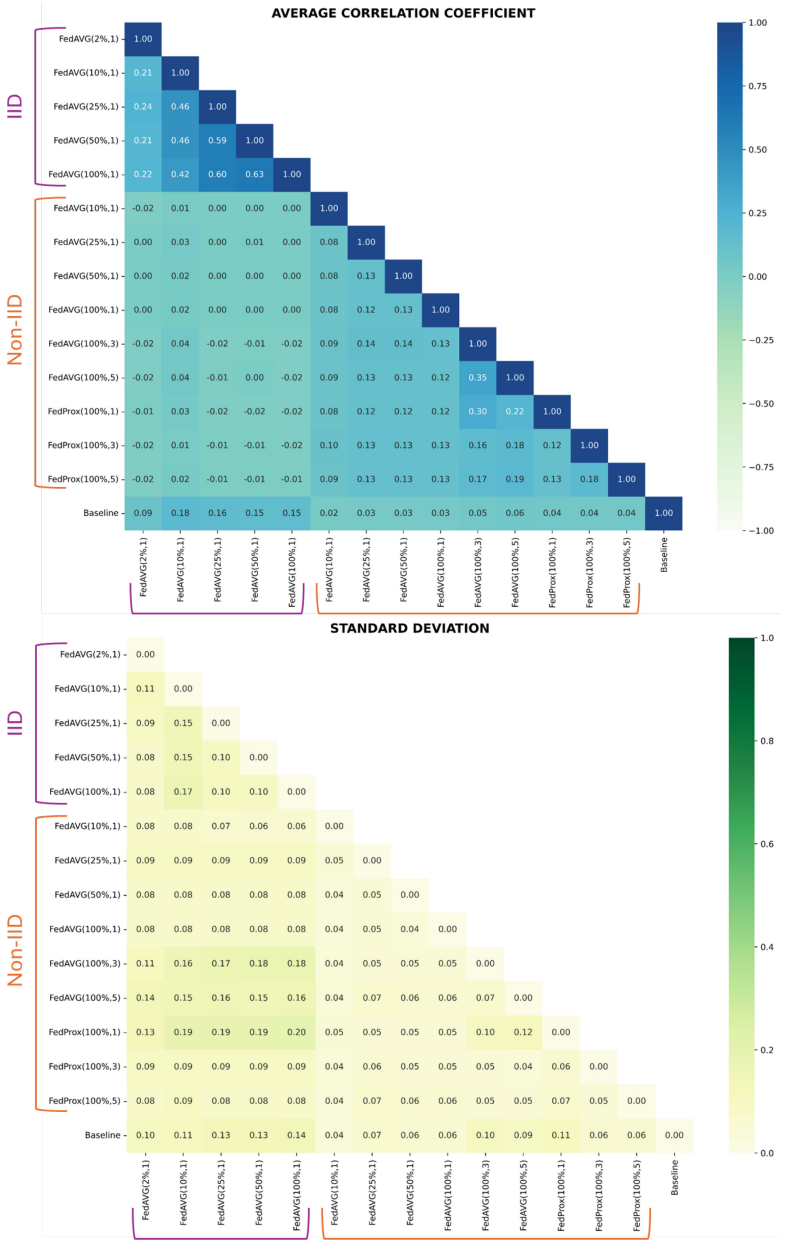


Fig. 4. Heatmaps of the average and the standard deviation of significant correlations between explanation distributions per couple of models. For each couple of models, the proportion of significant correlations never falls below 93% of the total number of coefficients.

RQ1: given the same task, dataset, and architecture, are XAI outputs consistent among centralized and federated models? The last row of the first heatmap in Fig. 4 represents the average correlation values between the explanations of the baseline model and those of the federated models. The results clearly indicate a lack of strong correlation between the attributions generated by the baseline model and any of the federated models. Despite sharing the same task, dataset, and architecture, the federated models appear to have learned different representational patterns for solving the brain age prediction task compared to the centralized baseline.

A notable difference emerges when considering the nature of the training dataset. While models trained on the non-IID dataset exhibit near-zero correlations with the baseline, those trained on the IID dataset demonstrate small but positive correlations, with values ranging between 0.10 and 0.20. This finding suggests that data distribution plays a significant role in shaping model explanations, as no similar trend is observed for other federated learning parameters, such as the number of local epochs, sampling rate, or aggregation protocol.

Among the IID-trained models, the configuration with a 2% sampling rate exhibits the lowest average correlation with the baseline. Although this setting was expected to resemble traditional mini-batch training, the extremely low sampling rate likely hindered the learning process by limiting the amount of data seen during each federated round. As previously discussed, this configuration can be interpreted as a centralized model trained on a very limited number of epochs, which partially explains the weak correlation with baseline attributions.

These findings highlight a key consideration: explanations generated by federated models are inconsistent with those of the centralized model, despite using the same task and dataset. While most federated learning parameters appear to have minimal influence on this discrepancy, the distribution of training data does exert a slight effect on model explanations.

RQ2: Given the same task, dataset, and architecture, are XAI outputs consistent among different configurations of the Federated Learning approach? Among the considered aspects of FL, the sampling rate appears to have a notable influence on explanation consistency, particularly in the IID dataset setting. Specifically, higher sampling rates correspond to higher correlation values among models trained under the IID condition. A similar, albeit weaker, trend is observed in the non-IID setting, where the configuration with a 10% sampling rate exhibits the lowest correlations within this group.

These findings suggest that neither the number of local epochs nor the aggregation protocol significantly impacts the consistency of model explanations across different FL configurations. Instead, a high node sampling rate should be prioritized to enhance explanation consistency in both IID and non-IID scenarios. However, it is important to note that a higher sampling rate does not always correspond to the best-performing model (see Table 1).

Interestingly, the effect of the node sampling rate per learning round is more pronounced in the IID setting, further emphasizing the importance of selecting a high sampling rate to maintain explanation consistency in federated models.

RQ3: Given the same task, dataset, and architecture, are XAI outputs consistent among federated models trained on different node data distributions? Regarding the consistency of model explanations across different data distributions among nodes, the heatmap reveals key patterns. First, models exhibit higher correlations with other models trained on the same dataset type rather than with configurations involving a different data distribution. Indeed, the average correlations between models trained on different dataset distributions are effectively null.

Furthermore, models trained with the IID dataset show stronger and higher correlations compared to those trained on the non-IID dataset, with some values exceeding 0.50. These results suggest that IID data distribution leads to more stable and consistent DeepSHAP attributions across different federated learning configurations.

Overall, these findings highlight that ensuring IID data distribution should be a priority when implementing a federated learning framework to enhance explanation stability across configurations. In a real-world scenario, collaborating clinics should establish standardized data acquisition protocols to facilitate IID sample distribution across nodes. This preprocessing step should be integrated into the data collection phase, as it precedes the implementation of any federated framework.

These observations align with the findings from RQ1, further reinforcing that maintaining IIDness in federated datasets is a fundamental prerequisite for ensuring consistency in model explanations.

Table 2. Number and percentage of features with a statistically significant correlation with chronological age per federated model.

Dataset	Aggregator	Sampling rate [%]	Local epochs	Features	
				Number	Percentage
IID	FedAVG	2	1	60	88
		10	1	49	72
		25	1	57	83
		50	1	54	79
		100	1	54	79
non-IID	FedAVG	10	1	62	91
		25	1	61	89
		50	1	54	79
		100	1	61	89
		100	3	63	92
		100	5	63	92
	FedProx	100	1	61	89
		100	3	61	89
		100	5	62	91

Intra-model Analysis. The last aim of this study was to explore the biological significance of ROI-based features across different training paradigms. We extracted region-based attributions by masking each saliency map with the Desikan-Killiany atlas and summing voxel attributions included in the same region. Next, the correlation between ROI-based features and chronological age was considered. The results of the analysis are displayed in Table 2 and Fig. 5.

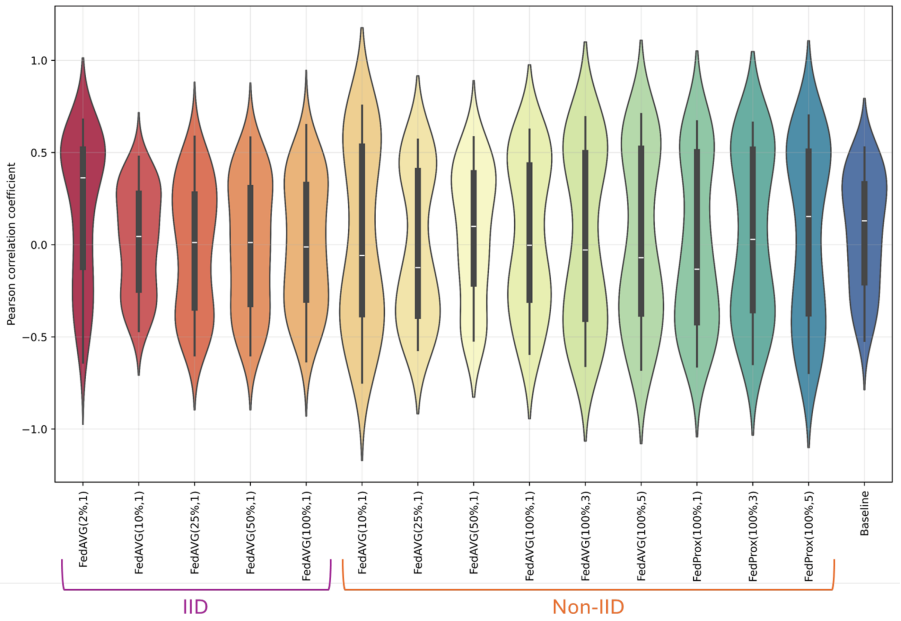


Fig. 5. Violin plots of Pearson correlation coefficient distribution between each ROI-based feature and the chronological age of the subjects per model.

RQ4: Do the identified features exhibit biological significance across different training paradigms? The number and percentage of ROI-based feature attributions that showed statistically significant correlations with chronological age for each federated model are summarized in Table 2. Most federated configurations identified a greater number of significant features compared to the centralized baseline model, which identified 55 significant features (80%). These results suggest that DeepSHAP effectively captures the influence of age on brain regions across both federated and centralized models.

Interestingly, among federated models trained on non-IID data, only the FedAVG model with a 50% sampling rate and 1 local epoch exhibited a lower percentage of significant features (79%), while all other non-IID configurations exceeded 89%. In contrast, federated models trained on the IID dataset exhibited lower percentages of relevant features, with three out of five configurations identifying fewer than 80% of significant features. These findings suggest that

federated models trained on IID datasets may struggle to capture meaningful age-related variations in certain brain regions, whereas non-IID federated models demonstrate an improved ability to detect age-related changes—sometimes even outperforming the centralized model.

To further assess the consistency of age-related attributions, we compared the distributions of Pearson correlation coefficients per model, as depicted in the violin plots in Fig. 5. Except for the 2% sampling rate configuration, all IID-trained models exhibited near-zero median correlations. In contrast, non-IID-trained models displayed medians fluctuating around zero. A Kruskal-Wallis test revealed no significant difference in Pearson correlation coefficients across models ($H = 17.96$, $p = 0.2085$), using a significance level of $\alpha = 0.05$.

6 Conclusions

In this study, we systematically examined the consistency of XAI attributions across centralized and federated training paradigms. To address our research questions, we conducted a comparative evaluation of centralized and federated models trained on the brain age prediction task, leveraging a publicly available multi-site dataset. Explanations were computed using DeepSHAP, a state-of-the-art XAI technique.

Our findings reveal that federated models produce substantially different explanations compared to their centralized counterparts, regardless of the specific federated learning configuration. This suggests that federated training alters feature attribution patterns, which could have implications for the interpretability and trustworthiness of federated models in medical applications. Within federated models, the non-IID nature of the dataset strongly reduces the consistency of DeepSHAP attributions, emphasizing the need for careful data distribution strategies in real-world FL applications.

Among the evaluated FL parameters, node sampling rate was the only factor that significantly influenced the consistency of XAI outputs. Our results suggest that high node sampling rates and IID data distributions should be preferred to enhance the reliability of federated explanations. However, it is important to note that a higher sampling rate does not always lead to the best predictive performance, highlighting a potential trade-off between model accuracy and explanation consistency.

A key limitation of this study is that we did not explore aggregation protocols beyond FedAVG and FedProx, nor did we investigate the potential influence of training hyperparameters (e.g., optimizer choice, learning rate) on explanation consistency. Future research will address these limitations by extending the analysis to alternative FL configurations, including different aggregation strategies and common extensions of the FL paradigm, such as differential privacy or secure aggregation. Future studies should also explore whether the observed inconsistencies in explanations affect the usability and trustworthiness of federated models in clinical decision-making. Additionally, the found biological significance of the models should be assessed through human evaluation by domain experts.

Acknowledgement. This work was partially supported by the following projects: "LIFE: the itaLian system wIde Frailty nEtwork"; DEMETRA: "Development of an ensemble learning-based, multidimensional sensory impairment score to predict cognitive impairment in an elderly cohort of Southern Italy" (CUP D99J22001970006) Missione 6/componente 2/Investimento: 2.1 "Rafforzamento e potenziamento della ricerca biomedica del SSN", funded by European Commission NextGenerationEU; "IDENTITA - rete Integrata meDiterranea per l'osservazione ed Elaborazione di percorsi di Nutrizione"; GenoCollab: Bando a Cascata - Prima Edizione - progetto PNRR, MISURA 4 - COMPONENTE 2—INVESTIMENTO 1.4 —"Programma di Ricerca del Centro Nazionale Di Ricerca - Sviluppo di Terapia Genica e Farmaci con Tecnologia a Rna" CN00000041, SPOKE 9 "From target to therapy: pharmacology, safety and regulatory competence center", finanziato dall'Unione Europea —NextGenerationEU.

Disclosure of Interests. None.

References

1. Anelli, V.W., Deldjoo, Y., Noia, T.D., Ferrara, A.: Prioritized multi-criteria federated learning. Intelligenza Artificiale **14**(2), 183–200 (2020)
2. Anelli, V.W., Deldjoo, Y., Noia, T.D., Ferrara, A., Narducci, F.: User-controlled federated matrix factorization for recommender systems. J. Intell. Inf. Syst. **58**(2), 287–309 (2022)
3. Bárcena, J.L.C., et al.: Fed-XAI: federated learning of explainable artificial intelligence models. In: Musto, C., Guidotti, R., Monreale, A., Semeraro, G. (eds.) Proceedings of the 3rd Italian Workshop on Explainable Artificial Intelligence co-located with 21th International Conference of the Italian Association for Artificial Intelligence (AIxIA 2022), Udine, Italy, November 28 - December 3, 2022. CEUR Workshop Proceedings, vol. 3277, pp. 104–117. CEUR-WS.org (2022). https://ceur-ws.org/Vol-3277/paper8.pdf
4. Bonawitz, K., et al.: Towards federated learning at scale: System design. CoRR abs/1902.01046 (2019)
5. Briola, E., Nikolaidis, C.C., Perifanis, V., Pavlidis, N., Efraimidis, P.S.: A federated explainable AI model for breast cancer classification. In: Li, S., Coopamootoo, K.P.L., Sirivianos, M. (eds.) European Interdisciplinary Cybersecurity Conference, EICC 2024, Xanthi, Greece, June 5-6, 2024, pp. 194–201. ACM (2024). https://doi.org/10.1145/3655693.3660255
6. Cole, J.H., et al.: Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. Neuroimage **163**, 115–124 (2017)
7. Corbucci, L., Guidotti, R., Monreale, A.: Explaining black-boxes in federated learning. In: Longo, L. (ed.) Explainable Artificial Intelligence - First World Conference, xAI 2023, Lisbon, Portugal, July 26-28, 2023, Proceedings, Part II. Communications in Computer and Information Science, vol. 1902, pp. 151–163. Springer (2023). https://doi.org/10.1007/978-3-031-44067-0_8
8. Corps, J., Rekik, I.: Morphological brain age prediction using multi-view brain networks derived from cortical morphology in healthy and disordered participants. Sci. Rep. **9**(1), 9676 (2019)
9. Bonis, M.L.N., et al.: Explainable brain age prediction: a comparative evaluation of morphometric and deep learning pipelines. Brain Inf. **11**(1), 33 (2024)

10. Ducange, P., Marcelloni, F., Renda, A., Ruffini, F.: Federated learning of XAI models in healthcare: a case study on Parkinson's disease. Cogn. Comput. **16**(6), 3051–3076 (2024). https://doi.org/10.1007/S12559-024-10332-X
11. Ducange, P., Marcelloni, F., Renda, A., Ruffini, F.: Federated learning of XAI models in healthcare: a case study on Parkinson's disease. Cogn. Comput. **16**(6), 3051–3076 (2024)
12. Dufumier, B., Grigis, A., Victor, J., Ambroise, C., Frouin, V., Duchesnay, E.: OpenBHB: a large-scale multi-site brain MRI data-set for age prediction and debiasing. Neuroimage **263**, 119637 (2022)
13. Gill, W., Anwar, A., Gulzar, M.A.: TraceFL: Interpretability-driven debugging in federated learning via neuron provenance. arXiv preprint arXiv:2312.13632 (2023)
14. Guendouzi, S.B., Ouchani, S., Assaad, H.E., El-Zaher, M.: A systematic review of federated learning: challenges, aggregation methods, and development tools. J. Netw. Comput. Appl. **220**, 103714 (2023). https://doi.org/10.1016/J.JNCA.2023.103714
15. He, Y., et al.: Foundation model for advancing healthcare: challenges, opportunities and future directions. IEEE Reviews in Biomedical Engineering (2024)
16. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
17. Kairouz, P., et al.: Advances and open problems in federated learning. arXiv preprint arXiv:1912.04977 (2019)
18. Kiani, M., Andreu-Perez, J., Hagras, H., Rigato, S., Filippetti, M.L.: Towards understanding human functional brain development with explainable artificial intelligence: Challenges and perspectives. IEEE Comput. Intell. Mag. **17**(1), 16–33 (2022)
19. Konecný, J., McMahan, B., Ramage, D.: Federated optimization: Distributed optimization beyond the datacenter. CoRR abs/1511.03575 (2015)
20. Konecný, J., McMahan, H.B., Ramage, D., Richtárik, P.: Federated optimization: Distributed machine learning for on-device intelligence. CoRR abs/1610.02527 (2016)
21. Kusiak, A.: Federated explainable artificial intelligence (fXAI): a digital manufacturing perspective. Int. J. Prod. Res. **62**(1-2), 171–182 (2024). https://doi.org/10.1080/00207543.2023.2238083
22. Lee, J., et al.: Deep learning-based brain age prediction in normal aging and dementia. Nat. Aging **2**(5), 412–424 (2022)
23. Leonardsen, E.H., et al.: Deep neural networks learn general and clinically relevant representations of the ageing brain. Neuroimage **256**, 119210 (2022)
24. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. In: MLSys. mlsys.org (2020)
25. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. Proc. Mach. Learn. Syst. **2**, 429–450 (2020)
26. López-Blanco, R., Alonso, R.S., González-Arrieta, A., Chamoso, P., Prieto, J.: Federated learning of explainable artificial intelligence (fed-XAI): a review. In: International Symposium on Distributed Computing and Artificial Intelligence, pp. 318–326. Springer (2023)
27. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 4768–4777. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)

28. Luzón, M.V., et al.: A tutorial on federated learning from theory to practice: foundations, software frameworks, exemplary use cases, and selected trends. IEEE/CAA J. Automatica Sinica **11**(4), 824–850 (2024)
29. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Proceedings of 20th International Conference on Artificial Intelligence and Statistics, pp. 1273–1282 (2017). http://proceedings.mlr.press/v54/mcmahan17a.html
30. More, S., et al.: Brain-age prediction: a systematic comparison of machine learning workflows. Neuroimage **270**, 119947 (2023)
31. Nguyen, D.C., et al.: Federated learning for smart healthcare: a survey. ACM Comput. Surv. **55**(3), 60:1–60:37 (2023)
32. Pfitzner, B., Steckhan, N., Arnrich, B.: Federated learning in a medical context: a systematic literature review. ACM Trans. Internet Techn. **21**(2), 50:1–50:31 (2021)
33. Rauniyar, A., et al.: Federated learning for medical applications: a taxonomy, current trends, challenges, and future research directions. IEEE Internet Things J. **11**(5), 7374–7398 (2023)
34. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. ICML **70**, 3145–3153 (2017). https://proceedings.mlr.press/v70/shrikumar17a.html
35. Sihag, S., Mateos, G., McMillan, C., Ribeiro, A.: Explainable brain age prediction using covariance neural networks. Adv. Neural. Inf. Process. Syst. **36**, 46958–46988 (2023)
36. Tafuri, B., et al.: The impact of harmonization on radiomic features in Parkinson's disease and healthy controls: a multicenter study. Front. Neurosci. **16**, 1012287 (2022)
37. Wang, G.: Interpret federated learning with shapley values. CoRR abs/1905.04519 (2019). http://arxiv.org/abs/1905.04519
38. Ye, M., Fang, X., Du, B., Yuen, P.C., Tao, D.: Heterogeneous federated learning: state-of-the-art and research challenges. ACM Comput. Surv. **56**(3), 79:1–79:44 (2024)
39. Yin, H., et al.: On-device recommender systems: A comprehensive survey. CoRR abs/2401.11441 (2024)