Supplementary Material: Feasibility of Federated Learning from Client Databases with Different Brain Diseases and MRI Modalities

A. Algorithm

Alg. 1 shows the pseudo-code of the FedUniBrain framework. We omit the mini-batch dimensionality for clarity.

Algorithm 1 FedUniBrain for Multi-modal MRI Seg.

Input: Set of clients C, each with dataset $\mathcal{D}_c$
Output: Trained global model parameters θ
SERVER EXECUTES:
  ▷ *First, get unique modalities and intialize*
  for each client $c \in C$ **in parallel do**
    $\mathcal{M}_c \leftarrow \text{CLIENTGETMODALITIES}(c)(\mathcal{D}_c)$
  $\mathcal{M} \leftarrow \bigcup_{c \in C} \mathcal{M}_c$
  Initialize global model θ with $|\mathcal{M}|$ input channels
  for $e = 1$ to E **do** ▷ *Begin training*
    for each client $c \in C$ **in parallel do**
      $\theta_c \leftarrow \text{CLIENTUPDATE}(c, \theta)$
    for each client model c and each layer l **do**
      if keep client-specific BN params = True:
        if $l \neq$ BN layer:
          $\theta^l = \frac{1}{C} \sum_{c=1}^{C} \theta_c^l$
      else:
        $\theta^l = \frac{1}{C} \sum_{c=1}^{C} \theta_c^l$
▷ *Below is executed on the clients*
CLIENTGETMODALITIES(c): ▷ *Run on client c*
  $\mathcal{M}_c \leftarrow$ Set of input MRI modalities from $\mathcal{D}_c$
  return $\mathcal{M}_c$ to server
CLIENTUPDATE($c, \theta, \mathcal{M}$): ▷ *Run on client c*
  $\theta_c \leftarrow \theta$
  if client-specific BN params = True :
    Overwrite all BN params with $\gamma_c, \beta_c, \hat{\mu}_c, \hat{\sigma}_c$
  for $t = 1$ to τ **do**
    $d_c \sim \mathcal{D}_c$ ▷ *Sample mini-batch*
    Initialize blank input tensor $b \in \mathbb{R}^{|\mathcal{M}| \times w \times h \times d}$
    ▷ *Copy client data to input tensor*
    $b[i, :, :, :] \leftarrow d_c[i, :, :, :] \quad \forall i \in \mathcal{M}_c$
    $k \leftarrow \text{RANDINT}(1, |\mathcal{M}_c|)$
    $\mathcal{M}_b \leftarrow$ Rand. sample k modalities from $\mathcal{M}_c$
    ▷ *Modality Drop by setting modalities blank*
    $b[j, :, :, :] \leftarrow 0. \quad \forall j \in \mathcal{M}_b$
    Perform local update on θ_c using: $\mathcal{L}_c(\theta_c; b; \mathcal{T}_c)$
  return θ_c

B. Additional Training Details

Additional Hyperparameter Group Norm (GN): In the results presented in Tab. 2, we use FedUniBrain with GN as an alternative non-client-specific feature normalization technique. The number of groups is an additional hyperparameter for GN, which we set to 16 for our experiments.

Centralised MultiUNet experiments: For the MultiUNet results in Tab. 4, we use the same training setup as described in [2]. However, to enable a direct comparison, we use the same number of epochs, learning rates, and loss function as in our Federated Learning experiments.

C. Datasets and their Different MRI Modalities

In Fig. C1, we show visual representations demonstrating that each MRI modality provides complementary information about tissues and anatomical structures through different contrasts. For each of the four databases - BRATS, ISLES, MSSEG, TBI - we show the same anatomical slices across different modalities, illustrating that each modality offers different diagnostic insights into the brain.

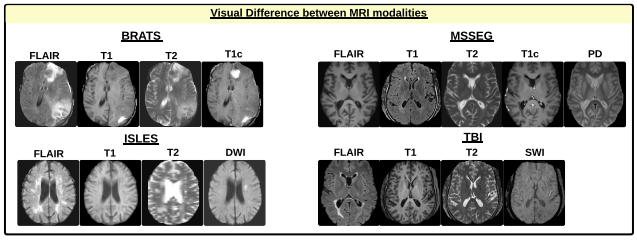

Figure C1. **MRI Databases with their different modalities.**

D. Statistical Analysis

The primary goal of this paper is to investigate the feasibility of training a single federated model, FedUniBrain, that can effectively segment multiple diseases across different databases, each with different MRI input modalities and brain pathologies. Since this approach has not been previously explored for federated training, showing that FedUni-

Brain does not overfit to a single dataset, disease, or input modality combination would be a great success. This would show the feasibility and benefits of training a unified model.

In some instances, our goal is to show that FedUniBrain performs at least as well as a centralised model through non-inferiority testing, which would confirm the feasibility of training a single federated model. In other cases, we aim to show that FedUniBrain outperforms single-center training through superiority tests, highlighting additional benefits of federated training a unified model.

Since we are working with segmentation models, each method generates a distribution of Dice scores per experiment (one Dice score per patient). This allows us to compute the Dice scores of the test data in each experiment for each method and conduct a one-sided t-test for non-inferiority or superiority, depending on our claim. Following, we explain the non-inferiority test and superiority test used in this analysis.

Non-inferiority test: To evaluate whether the federated approach, FedUniBrain, is not inferior to a specific method (based on our claim), we conduct statistical non-inferiority tests. In these tests, the null hypothesis (H_0) states that the mean difference in Dice performance between FedUniBrain and the specific method is less than or equal to a specified margin, indicating that FedUniBrain is inferior. The alternative hypothesis (H_1) states that the mean difference in Dice performance is greater than $-\Delta$, suggesting that FedUniBrain is not worse than the specific method by more than the margin Δ.

$$H_0 : \delta_d \leq -\Delta$$
$$H_1 : \delta_d > -\Delta,$$

where δ_d represents the mean paired differences of the samples' Dice scores, and Δ represents a pre-specified margin, which we set to 5%, a commonly used value [1]. To conduct the non-inferiority tests, we use a **paired one-sided t-test**.

Superiority test: To evaluate whether the federated approach, FedUniBrain, is better than a specific method (based on our claim) in terms of segmentation performance, we employ statistical superiority tests. We also use a **paired one-sided t-test** to evaluate whether the mean performance of FedUniBrain is superior to the mean performance of a specific method. The null hypothesis (H_0) states that the mean difference in Dice performance is less than or equal to zero, indicating that FedUniBrain is not superior. The alternative hypothesis (H_1) states that the mean difference in performance is greater than zero, demonstrating that FedUniBrain's performance is superior compared to the specific method:

$$H_0 : \delta_d \leq 0$$
$$H_1 : \delta_d > 0.$$

Reporting of the statistical results: For all our tests, we report the number of samples in our test (N), the difference in means, the 95% confidence intervals, and the p-value.

Note, we are reporting the 95% confidence intervals for the difference in means of the Dice performance (from the paired t-test). For successful **superiority** testing, positive confidence intervals are expected, and a confidence interval's lower limit above 0 demonstrates statistical superiority. If the 95% confidence interval's lower limit is below zero, it indicates no statistical superiority. However, for non-inferiority testing, the interpretation is different. Here, we want to show that the mean performance difference is not worse than our predefined margin Δ. This means that negative confidence intervals are acceptable and indicate statistical significance, as long as the lower bound does not fall below the margin $-\Delta$. Note that since we use a one-sided test, it is expected that the 95% confidence interval has an upper bound of infinity.

We reject the null hypothesis if the p-value falls below 0.05 (our significance level), which would indicate that FedUniBrain is statistically not-inferior (= not worse than) or superior (= better than), depending on the test.

D.1. Statistical Analysis of FedUniBrain and Single Center Training

In this section, we statistically test FedUniBrain against single-center training. Our goal is to demonstrate that FedUniBrain can match or even improve upon the performance of single-center training, which would be a major success as it would show that our model does not overfit to a specific dataset, disease, or input modality combination. Achieving this would validate the main goal of our paper, proving that training a single model across multiple databases with different diseases and input modality combinations is feasible. We show that FedUniBrain is either statistically superior or non-inferior compared to single-center training for all datasets.

The results are presented in Table D.1, corresponding to the results of Tab. 2 of the main paper. We compare FedUniBrain with client-specific batch normalization (BN) parameters and modality drop against single-center training **without** modality drop (because these two are the best performing approaches). The results indicate that FedUniBrain consistently performs non-inferior to single-center training and, for the ATLAS and WMH datasets, superior.

D.2. Statistical Analysis of Zero-Shot Generalization

This section shows a statistical analysis comparing FedUniBrain with different normalization techniques to the centralized MultiUnet approach (all of them **with** Modality Drop). Demonstrating that FedUniBrain is statistically non-inferior to the centralized MultiUNet method would be a major accomplishment, indicating that FedUniBrain

Table D.1. Statistical comparison of FedUniBrain with single-center training for segmentation performance

Client joining	Stat. Test	Samples (N)	Mean Federated Dice	Mean Single-Center Dice	Mean of Diff.	95% CI	p-Value
ATLAS	Superiority	195	54.5	52.8	1.73	(0.27, ∞)	**0.0260**
BRATS	Non-Inferiority	40	91.8	91.9	-0.13	(-0.71, ∞)	**2.2e-16**
MSSEG	Non-Inferiority	16	69.1	68.3	0.79	(-3.16, ∞)	**0.0106**
TBI	Non-Inferiority	125	56.2	56.1	0.11	(-1.03, ∞)	**8.538e-12**
WMH	Superiority	18	73.7	71.5	2.21	(1.19, ∞)	**7.335e-06**

Table D.2. Statistical comparison of FedUniBrain with the centralized method for segmentation performance on ISLES and Tumor2

		ISLES						Tumor2					
Model	Norm	N	Cent. Mean Dice	Fed. Mean Dice	Mean of Diff.	95% CI	p-Value	N	Cent. Mean Dice	Fed. Mean Dice	Mean of Diff.	95% CI	p-Value
FedUniBrain	IN	28	55.5	55.3	0.26	(-2.71, ∞)	**0.0027**	57	72.2	72.7	0.12	(-1.45, ∞)	**5.443e-07**
FedUniBrain	NF	28	55.5	52.8	-2.18	(-6.01, ∞)	0.1105	57	72.2	72.1	-0.53	(-1.65, ∞)	**6.181e-09**
FedUniBrain (avg. BN params)	BN	28	55.5	54.5	-0.49	(-4.91, ∞)	**0.0469**	57	72.2	68.0	-4.62	(-6.76, ∞)	0.3848
FedUniBrain (client spec. BN params)	BN	28	55.5	49.9	-5.1	(-11.31, ∞)	0.5079	57	72.2	70.1	-2.52	(-3.91, ∞)	**0.0021**

Table D.3. Statistical comparison of FedUniBrain models with single center training when a new client joins in for segmentation performance

Client joining	Stat. Test	Samples (N)	Mean Single-Center Dice	Mean Federated Dice	Mean of Diff.	95% CI	p-Value
ISLES	Superiority	8	48.5	53.7	4.79	(0.24, ∞)	**0.0422**
Tumor2	Superiority	17	74.7	78.1	3.15	(0.07, ∞)	**0.0466**
TBI	Non-Inferiority	125	53.9	54.4	0.51	(-1.06, ∞)	**2.399e-08**
MSSEG	Non-Inferiority	16	66.7	67.6	0.91	(-2.72, ∞)	**0.0061**

is on par with the centralized method. We perform non-inferiority tests between FedUniBrain with different normalization methods and the centralized MultiUNet method. These tests correspond to the results from Tab. 4 of the main paper.

The results in Tab. D.2 show that FedUniBrain with Instance Normalization (IN) is statistically non-inferior to the centralized method for both the ISLES and Tumor2 datasets. Comparing different normalization techniques for FedUniBrain, IN is the only one that is statistically non-inferior on both datasets and therefore comes closest to matching the performance of the centralized method. This strengthens our argument that the choice of normalization is important depending on whether the goal is a personalized model or a model that generalizes well.

D.3. Statistical Analysis of a New Client Joining the Federation

In this section, we want to evaluate whether FedUniBrain does not perform worse than or even surpass single-center training in the challenging scenario where a new client joins during training. Specifically, for scenarios where a new client with a known pathology joins, we perform statistical superiority tests. The results of these tests are presented in Tab. D.3 and correspond to the results shown in Fig. 4 of the main paper. From the results, we can see that when a new client with an already-seen pathology joins, FedUniBrain performs significantly better than single-center training (Tumor2 and ISLES). In the more challenging scenario where a new client with a previously unseen pathology **and** unseen modality joins, we perform statistical non-inferiority tests. The results of these tests demonstrate that FedUniBrain is statistically non-inferior to single-center training, which is an important result. This indicates that even in the highly challenging setting of continual learning, FedUniBrain does not overfit to specific databases and is capable of learning from a completely new database with a new brain pathology, including a new modality, as effectively as single-center training.

References

[1] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. International evaluation of an ai system for breast cancer screening. *Nature*, 2020. 2

[2] Wentian Xu, Matthew Moffat, Thalia Seale, Ziyun Liang, Felix Wagner, Daniel Whitehouse, David Menon, Virginia Newcombe, Natalie Voets, Abhirup Banerjee, et al. Feasibility and benefits of joint learning from mri databases with different brain diseases and modalities for segmentation. In *Medical Imaging with Deep Learning*, 2024. 1