2024 IEEE International Symposium on Biomedical Imaging (ISBI) | 979-8-3503-1333-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISBI56570.2024.10635565

ARIA: ON THE INTERACTION BETWEEN ARCHITECTURES, INITIALIZATION AND AGGREGATION METHODS FOR FEDERATED VISUAL CLASSIFICATION

Vasilis Siomos[1], *Sergio Naval-Marimont*[1], *Jonathan Passerat-Palmbach*[1,2], *Giacomo Tarroni*[1,2]

[1] CitAI Research Centre, Department of Computer Science City, University of London
[2] BioMedIA, Department of Computing, Imperial College London

ABSTRACT

Federated Learning (FL) is a collaborative training paradigm that allows for privacy-preserving learning of cross-institutional models by eliminating the exchange of sensitive data and instead relying on the exchange of model parameters between the clients and a server. Despite individual studies on how client models are aggregated, and, more recently, on the benefits of ImageNet pre-training, there is a lack of understanding of the effect the architecture chosen for the federation has, and of how the aforementioned elements interconnect. To this end, we conduct the first joint ARchitecture-Initialization-Aggregation study and benchmark ARIAs across a range of medical image classification tasks. We find that, contrary to current practices, ARIA elements have to be chosen together to achieve the best possible performance. Our results also shed light on good choices for each element depending on the task, the effect of normalization layers, and the utility of SSL pre-training, pointing to potential directions for designing FL-specific architectures and training pipelines.

***Index Terms*—** Federated Learning, Self-Supervised Pre-training

1. INTRODUCTION

Federated learning (FL) for healthcare [1] has emerged as a promising approach that enables collaborative machine learning without direct access to raw patient data. The typical scenario for medical imaging data is the cross-silo setting, where a small number of data owners/stakeholders fully participate in a round of federated training by training their local/client models and sending the parameters to a central server, which aggregates the client models into a server/global model. The global model is then broadcast to all clients to start the next round, until training stops, and the final model is delivered to the stakeholders for deployment.

Since the seminal FedAvg paper [2], progress in cross-silo visual classification has been hard to determine, with innovation often focusing on improving the aggregation strategy for the frequent scenario where the client datasets are heterogeneous [3, 4]. Unfortunately, proposed methods most commonly use randomly initialized model weights, small/toy models, or both [5]. This makes comparing and drawing conclusions for real-world medical settings difficult.

Recent studies [6, 7] have been exploring the value of using ImageNet (IN) pre-trained networks for FL, showcasing improvements in closing the gap to centralized performance, improving overall performance, and reducing the effect of data heterogeneity. Another study by Qu et al. [8] highlighted the benefits of using IN pre-trained transformers for FL. Very recently, Pieri et al. [5] focused on the interaction of aggregation methods and architectures, but only examined IN pre-trained weights.

It's important to note IN pre-training restricts the input to 224x224 RGB images. When up-sampling of the original images is required to achieve that, it leads to a bigger than necessary computational and memory load, and the introduction of aliasing artifacts (e.g. Fig. 1). When down-sampling is required instead, it can degrade performance. Hence, IN pre-training is not a silver bullet, and benchmarking architectures and aggregation strategies without pre-training is also important. Furthermore, task-relevant pre-training through self-supervised learning (SSL) has recently emerged as a highly-effective alternative to IN pre-training [9], but its usefulness in the FL setting remains largely unexplored.

Motivated by the above, we conduct what, to the best of our knowledge, is the first study to jointly examine ARIAs: Architecture-Initialization-Aggregation combinations: we select 9 architectures, with the weights initialized from 3 starting points (Random, ImageNet, SSL on a relevant dataset), and use 3 of the most common methods (FedAvg, FedOpt, SCAFFOLD) to aggregate the models. We focus on perhaps the most beneficial domain for FL, medical imaging, and evaluate the resulting ARIAs on 3 different medical imaging datasets, namely Fed-ISIC, and two versions of OrganAMNIST (with and without simulated heterogeneity).

Our results after training more than 300 ARIAs indicate to researchers and practitioners designing FL pipelines for medical imaging data that all elements of an ARIA have to be evaluated together, but also shed light on the individual effects of network size, normalization methods, architecture choice, and utility of SSL pre-training.

2. METHODS

2.1. (AR)chitectures

We aim to compare popular architectures from both the convolutional and transformer families, while also pinpointing architectural block that boost FL performance. All models are of reasonable size and throughput for our target tasks. For comparison, rows in tables 1,2,3 are listed in decreasing training throughput, and we include model parameter counts here. From the family of residual networks [10], we choose a ResNet-18 (11.7M parameters), a ResNet-50 (25.6M), and a Wide-ResNet-50-2 [11] (68.9M), to examine the effect of depth and width. A DenseNet-121 [12] (8M), shows how the density of residual connections and feature re-use affect performance.

These architectures employ Batch Normalisation (BN), which is known to degrade FL performance in heterogeneous settings, due to the BN statistics being averaged across heterogeneous image distributions[13]. There is no clear solution to this, with replacing BN layers in a ResNet with Group or Layer Normalization, or not sharing the BN layers, having been proposed before [13]. Our approach to providing insight into alternatives to BN is benchmarking

Table 1. Average balanced accuracy across 6 clients on Fed-ISIC. IN top-1 accuracy reported next to model name. Models listed in decreasing measured training throughput (using AMP). Difference from average balanced accuracy of centrally trained model in parentheses.

Initialization	Random			ImageNet Pre-Training			DINO on Skin SSL dataset		
Agg. Method	FedAvg	FedOpt	SCAFFOLD	FedAvg	FedOpt	SCAFFOLD	FedAvg	FedOpt	SCAFFOLD
ResNet-18 (69.76)	51.65 (↓ 9.8)	46.7 (↓ 14.7)	52.45 (↓ 9)	65.87 (↓ 4.3)	67.55 (↓ 2.6)	68.66 (↓ 1.5)	66.57 (↓ 5.7)	62.36 (↓ 10)	66.87 (↓ 5.4)
NF-ResNet-50 (80.64)	55.93 (↓ 6.1)	56.25 (↓ 5.8)	**59.64 (↓ 2.4)**	71.88 (↑ 0.9)	68.75 (↓ 2.2)	71.53 (↑ 0.5)	67.83 (↓ 0.7)	67.92 (↓ 0.6)	70.11 (↑ 1.6)
ResNet-50 (80.86)	49.11 (↓ 12)	46.91 (↓ 14.2)	48.13 (↓ 13)	67.97 (↓ 6.3)	66.16 (↓ 8.1)	68.48 (↓ 5.8)	65.16 (↓ 7.2)	66.46 (↓ 5.9)	66.34 (↓ 6)
WRN-50-2 (81.6)	50.53 (↓ 8)	50.12 (↓ 8.4)	51.03 (↓ 7.5)	69.54 (↓ 5.3)	67.68 (↓ 7.2)	70.34 (↓ 4.5)	65.56 (↓ 6.9)	64.22 (↓ 8.3)	66.66 (↓ 5.8)
DenseNet-121 (74.43)	49.42 (↓ 13.3)	45.95 (↓ 16.8)	52.79 (↓ 9.9)	67.34 (↓ 5.8)	68.03 (↓ 5)	68.52 (↓ 4.6)	66.28 (↓ 5.3)	64.94 (↓ 6.6)	67.38 (↓ 4.2)
SWIN-T (81.47)	45.73 (↑ 23.2)	44.13 (↑ 21.6)	45.00 (↑ 22.5)	71.19 (↓ 1.3)	71.81 (↓ 0.6)	73.13 (↑ 0.7)	72.13 (↑ 1.7)	71.40 (↑ 0.9)	73.77 (↑ 3.3)
EfficientNetV2-S (84.22)	46.59 (↓ 10.8)	46.59 (↓ 10.8)	47.51 (↓ 9.8)	70.00 (↓ 9.6)	71.48 (↓ 8.1)	73.18 (↓ 6.4)	57.99 (↓ 14.9)	59.74 (↓ 13.1)	64.98 (↓ 7.9)
ViT-B-16 (81.07)	47.84 (↑ 7.2)	49.52 (↑ 8.9)	48.44 (↑ 7.8)	65.86 (↑ 1.6)	65.18 (↑ 0.9)	68.09 (↓ 3.8)	71.06 (↓ 2.9)	71.52 (↓ 2.5)	69.49 (↓ 4.5)
ConvNext-S (83.61)	48.10 (↓ 7.9)	49.93 (↓ 6.1)	48.56 (↓ 7.5)	**75.08 (↓ 0.1)**	73.40 (↓ 1.7)	74.28 (↓ 0.8)	72.07 (↓ 3)	73.57 (↓ 1.5)	**74.56 (↓ 0.5)**

Table 2. Average accuracy across 4 clients on OrganAMNIST with $\alpha = 0.1$. IN top-1 accuracy reported next to model name. Models listed in decreasing measured training throughput (using AMP). Difference from the accuracy of the centrally trained model in parentheses.

Initialization	Random			ImageNet Pre-Training			DINO on Abdomen-SSL		
Agg. Method	FedAvg	FedOpt	SCAFFOLD	FedAvg	FedOpt	SCAFFOLD	FedAvg	FedOpt	SCAFFOLD
ResNet-18 (69.76)	88.8 (↓5.6)	90.76 (↓3.6)	89.16 (↓5.2)	94.02 (↓1.9)	94.78 (↓1.2)	94.33 (↓1.6)	83.54 (↓9.8)	87.89 (↓5.5)	84.76 (↓8.6)
NF-ResNet-50 (80.64)	71.6 (↓16.3)	78.84 (↓9.1)	73.8 (↓14.1)	94.39 (↓1.4)	95.26 (↓0.5)	95.2 (↓0.6)	84.58 (↓7.9)	87.93 (↓4.5)	86.92 (↓5.5)
ResNet-50 (80.86)	83.32 (↓10.5)	86.6 (↓7.2)	84.82 (↓9.0)	91.98 (↓3.5)	92.98 (↓2.5)	92.32 (↓3.1)	81.33 (↓12.9)	85.69 (↓8.5)	81.49 (↓12.8)
WRN-50-2 (81.6)	84.52 (↓9.6)	85.58 (↓8.5)	83.82 (↓10.3)	90.56 (↓4.3)	91.71 (↓3.2)	90.4 (↓4.5)	79.98 (↓13.7)	85.02 (↓8.6)	77.09 (↓16.5)
DenseNet-121 (74.43)	86.01 (↓8.6)	89.12 (↓5.5)	85.06 (↓9.6)	94.72 (↓2.2)	95.1 (↓1.9)	94.68 (↓2.3)	85.26 (↓9.2)	89.21 (↓5.3)	84.94 (↓9.5)
SWIN-T (81.474)	83.03 (↓8.6)	85.17 (↓6.4)	83.16 (↓8.4)	95.64 (↓0.6)	95.83 (↓0.4)	95.83 (↓0.4)	83.4 (↓8.2)	86.4 (↓5.2)	84.8 (↓6.8)
EfficientNetV2-S (84.22)	88.8 (↓6.2)	**91.46 (↓3.6)**	89.19 (↓5.9)	94.0 (↓2.7)	94.26 (↓2.4)	93.46 (↓3.2)	61.19 (↓31.6)	67.54 (↓25.3)	56.2 (↓36.6)
ViT-B-16 (81.072)	83.14 (↓4.2)	83.52 (↓3.9)	83.85 (↓3.5)	95.3 (↓1.5)	95.96 (↓0.9)	**96.01 (↓0.8)**	81.34 (↓6.8)	83.76 (↓4.4)	81.99 (↓6.2)
ConvNext-S (83.61)	53.76 (↓35.4)	56.07 (↓33.1)	55.34 (↓33.8)	94.12 (↓2.6)	94.92 (↓1.8)	94.84 (↓1.9)	87.31 (↓6.0)	**89.68 (↓3.7)**	87.64 (↓5.7)

networks that altogether do not use BN in their original form. To this end, we use a Normalization-Free (NF) ResNet-50 [14] (25.6M). NF architectures rely on Scaled Weight Standardization (SWS), i.e. careful scaling of weights, instead of normalization, to achieve correct signal propagation during learning.

Since the emergence of vision transformers, more advanced convolutional architectures have been introduced with the goal of outperforming them, and we pick EfficientNetV2-S [15] (21.5M, uses BN), an evolution of ResNets guided by neural architecture search, and ConvNext-S [16] (50.2M, uses LN), which borrows design principles from SWIN transformers, as modern CNN benchmarks.

From the transformer family, we benchmark a ViT-B-16 [17] (86.6M) and a SWIN-T [18] (28.3M), to compare convolution with self-attention. Both employ Layer Normalization (LN).

2.2. (I)initialization

Both random and IN initializations are of interest depending on the application, as explained in section 1; all rows in rows in tables 1,2,3 list the pre-trained model's top-1 accuracy on IN. However, in medical imaging scenarios, it is often the case that i) the target task images are dissimilar to the natural ones of IN and ii) medical datasets with similar images are publicly available. This leads us to examine whether training the models using self-supervised learning (SSL) as a pre-cursor task can be a beneficial initialization strategy for FL. We construct two task-relevant pre-training datasets, Abdomen-SSL and Skin-SSL (Fig 1), and train all models using DINO [19]for 100/300 epochs on the two datasets respectively, with the length chosen based on the loss plateauing.

2.3. (A)ggregation methods

We limit our scope to methods that produce a global model w_g, and not a personalized model for each client. We select three of the most common aggregation strategies, namely FedAvg, FedOpt, and SCAFFOLD, which share in common their ease/lack of hyper-parameters to be tuned, allowing for more universal insights. **FedAvg** [2] is the seminal FL parameter averaging method, which uses as the sample-weighted average of client models w_i to produce $w_g = N_i/N \cdot \sum_{i=1}^{C} w_i$.

The **FedOpt** [20] family of methods de-couples server and client-side optimization, and the server can employ any optimizer like SGD, Adam, etc. We use SGD with momentum at the server, similar to FedAvgM [3], with the addition of a cosine annealing to the server learning rate. The server learning rate is 1.0, tuned from $\{0.5, 1\}$, and the momentum to 0.6, tuned from $\{0.6, 0.9\}$.

SCAFFOLD [4] utilizes control variates to correct local model updates against client-drift (divergence from the global model). These parameters are of equal size to the model and there is one set being stored locally and another exchanged alongside the model, leading to twice the communication and triple the local storage cost.

3. EXPERIMENTS

3.1. Datasets

We conduct experiments on abdominal CT with OrganAMNIST [21] and skin lesions with Fed-ISIC [22]. The latter is naturally federated with multi-center data, and for the former we construct a federated version of 4 clients, by following convention and using the Dirichlet partitioning strategy [3], which induces size and label distribution heterogeneity based on the controllable concentration parameter α. We examine an IID setting by setting $\alpha = 100$, and a highly heterogeneous one by setting $\alpha = 0.1$. This leads to a wide range of difficulty to benchmark the chosen models, from the grayscale IID OrganAMNIST to the highly imbalanced, both in label distribution and data size RGB, Fed-ISIC. Moreover, the distance between

Table 3. Average accuracy across 4 clients on OrganAMNIST with $\alpha = 100$. IN top-1 accuracy reported next to model name. Models listed in decreasing measured training throughput (using AMP). Difference from accuracy of centrally trained model in parentheses.

Initialization	Random			ImageNet Pre-Training			DINO on Abdomen-SSL		
Agg. Method	FedAvg	FedOpt	SCAFFOLD	FedAvg	FedOpt	SCAFFOLD	FedAvg	FedOpt	SCAFFOLD
ResNet-18 (69.76)	93.8 (↓0.6)	94.3 (↓0.1)	93.97 (↓0.4)	96.05 (↑0.1)	96.38 (↑0.4)	95.99 (↓0.0)	92.06 (↓1.3)	93.47 (↑0.1)	92.14 (↓1.2)
NF-ResNet-50 (80.64)	84.28 (↓3.6)	88.09 (↑0.2)	84.4 (↓3.5)	95.5 (↓0.3)	95.64 (↓0.1)	95.6 (↓0.2)	92.08 (↓0.4)	92.74 (↑0.3)	92.09 (↓0.4)
ResNet-50 (80.86)	93.39 (↓0.5)	94.0 (↑0.2)	93.54 (↓0.3)	94.98 (↓0.5)	95.56 (↑0.1)	95.34 (↓0.1)	92.8 (↓1.4)	**93.51 (↓0.7)**	92.69 (↓1.5)
WRN-50-2 (81.6)	93.74 (↓0.3)	93.99 (↓0.1)	93.72 (↓0.4)	94.7 (↓0.2)	95.42 (↑0.5)	94.76 (↓0.1)	92.24 (↓1.4)	93.1 (↓0.5)	92.52 (↓1.1)
DenseNet-121 (74.43)	93.95 (↓0.7)	94.28 (↓0.3)	93.66 (↓1.0)	96.53 (↓0.4)	**97.0 (↓0.0)**	96.66 (↓0.3)	93.4 (↓1.1)	94.11 (↓0.4)	93.38 (↓1.1)
SWIN-T (81.47)	90.64 (↓1.0)	90.89 (↓0.7)	90.27 (↓1.3)	96.61 (↑0.4)	96.82 (↑0.6)	96.6 (↑0.4)	89.66 (↓2.0)	90.86 (↓0.8)	89.68 (↓1.9)
EfficientNetV2-S (84.22)	94.84 (↓0.2)	**95.13 (↑0.1)**	94.96 (↓0.1)	96.22 (↓0.5)	96.48 (↓0.2)	96.28 (↓0.4)	89.18 (↓3.6)	92.03 (↓0.8)	89.02 (↓3.8)
ViT-B-16 (81.07)	86.42 (↓1.0)	86.34 (↓1.0)	86.54 (↓0.8)	96.12 (↓0.7)	96.3 (↓0.5)	96.25 (↓0.6)	86.67 (↓1.5)	87.61 (↓0.6)	86.94 (↓1.2)
ConvNext-S (83.61)	84.56 (↓4.6)	87.29 (↓1.9)	78.48 (↓10.7)	96.3 (↓0.4)	96.24 (↓0.5)	96.18 (↓0.5)	92.15 (↓1.2)	92.87 (↓0.5)	92.24 (↓1.1)

the domains and ImageNet provides more insight into learning dynamics for the medical community compared to testing on natural images.

OrganAMNIST [21] consists of 58,850 28x28 grayscale images with 11 organ labels segmented from axial slices of abdominal CT scans. We upscale the images to 224x224 and copy the channel over 3 times for compatibility with IN pre-trained models. Each client has a training and validation set, with the local validation set used to determine good local training hyper-parameters. After that, clients train on the union of their two sets, and accuracy is reported on the original, pooled, test dataset of 17,778 images.

The **Abdomen-SSL** dataset was created by extracting 20 slices around the center of each volume in 4 abdominal CT datasets [23, 24, 25], cropping around the subject, resizing to 224x224 and copying the channel over, resulting in $\sim 21,000$ whole abdomen images. As seen in Fig.1, Abdomen-SSL is quite different to OrganAMNIST. Due to OrganAMNIST's uniqueness, it is difficult to design a more similar source dataset. However, SSL pre-training can still help the models learn general organ structures and channel redundancy.

Fed-ISIC [22] consists of 23,247 RGB skin lesion images with 8 classes, split across 6 clients representing different datacenters and imaging technologies. Fed-ISIC exhibits very high heterogeneity in size and label imbalance, so performance is measured through balanced accuracy, defined as the average recall on each class. We follow the pre-processing in [22], applying color constancy, and centre-cropping while maintaining the aspect ratio.

Skin-SSL was created from 3 skin lesion datasets [26, 27, 28], with the largest contributor being ISIC-2020, which has no overlap with Fed-ISIC, and consists predominantly of benign samples.

For all settings, besides federated training, we also train a central model on the pooled datasets to compare the FL models against, and tables 1,2,3 present each model's difference from its centrally trained counterpart in parentheses.

3.2. Hyper-parameters

In order to concentrate on the ARIA effects, we limit our scope to shared hyper-parameters between the clients (no client-level tuning), and across aggregation methods. For Fed-ISIC, we follow [22] and train for 20 rounds, using Weighted Focal Loss, a batch size of 64, Adam with $lr = 5 \cdot 10^{-4}$, and a cosine annealer. Instead of local epochs, each client performs 200 local steps, tuned from [100,200,600], which allows the biggest client to iterate through all of its data, but keeps client drift to a minimum. While 200 steps worked best for SSL and IN pre-training, we use 600 local steps when training from scratch as a *parity measure* since these models have not seen any data prior, and this improved performance. Adam, surprisingly, worked well for all networks, outperforming both SGD with momentum (favors CNNs) and AdamW (favors transformers) in our tests. The learning rate was tuned in the range $[10^{-4}, 10^{-}3]$; the combination of adaptive momentum buffers at each client and cosine annealing led to different initial learning rates having minimal effect. For OrganAMNIST, we used the local validation sets of the IID partition and majority voting to decide on the use of momentum SGD with ($lr = 0.01$, $m = 0.9$), a cosine annealing schedule, a batch size of 128, and 50 local steps. We transfer these settings to the heterogeneous case, since the heterogeneity is typically not known about in advance. Results are averaged across two seeds. We open source our code[1], which uses NVFlare [29].

4. RESULTS AND DISCUSSION

4.1. Comparing initializations

In the IID OrganAMNIST experiment (Table 3) IN pretrained networks virtually solve the task, and achieve very low gaps compared to centralized training (max 0.6%). This gap increases as heterogeneity in the other two datasets increases, as expected, but overall the IN Initialization outperformed the others. This leads to our first important finding: **ImageNet is generally the best initialization for federated learning on medical datasets**. In SSL initialization, Skin-SSL pre-training is predictably more useful (Table 1) than Abdomen-SSL due to the source and target images being much more similar (Figure 1). Abdomen SSL pre-training reduce performance on average, but helped "prime" ConvNext-S and NF-ResNet-50 compared to random initialization, indicating that SSL can counteract the reduction in regularization due to not using BN. Overall, **Skin-SSL greatly increases the performance of all models compared to random initialization**. Moreover, despite the much shorter pre-training time compared to IN, the SSL initialized ConvNext-S with SCAFFOLD nearly achieves the best overall performance. In summary, our findings suggest SSL can be extremely beneficial in medical FL in multiple scenarios, from task-specific architectures that have no public IN pre-trained weights, to tasks that cannot adhere to IN image size, and potentially even tasks beyond visual classification, such as segmentation.

4.2. ResNet depth, width, and connection density

Despite deepening and widening generally improving the centrally trained model, the increased central training accuracy was not transferred to FL training. Hence, in our findings, **ResNets do not scale**

[1] https://github.com/siomvas/ARIA

well in FL tasks, as ResNet-18 outperforms its deeper and wider counterparts in all settings except for Fed-ISIC with IN weights (Table 1), where the much larger and slower WRN-50-2 is modestly better. If a low memory footprint is a priority, DenseNet-121, which has much fewer parameters than all other networks but lower throughput than other residual networks, performs just as well or better depending on the task, suggesting that its salient characteristic, **feature reuse, is beneficial for FL.**

4.3. Comparing normalization methods

It has been widely discussed, most recently in [13], that BN impedes FL performance under heterogeneous settings due to the local clients calculating statistics that are not representative of each other's datasets. Simultaneously, BN is reliant on the batch size being sufficiently big to accurately approximate the mean and variance, in contrast to LN and SWS. For OrganAMNIST we use a batch size of 128; as a result, we observe (Tables 2, 3) that **randomly initialized BN models outperform LN and SWS ones**, under both IID and non-IID distributions, but there is no difference when the models are pre-trained. For Fed-ISIC, where the batch size was 64, BN could not help random models as much, and **when using IN weights the top three models all use LN or SWS.**

Compared to the (generally bigger) LN networks, the performance of NF-ResNet (which uses SWS) does not suffer as much for the random initialization, and the model even performs the best out of all random models on Fed-ISIC. **This makes NF networks and SWS even more promising for FL applications.**

4.4. Transformers vs CNNs for FL

Randomly initialized transformers perform poorly in our experiments, a finding that is perhaps due their lack of inductive bias compared to CNNs, and one that we cannot attribute to model size or speed, as SWIN-T often outperformed ViT-B while being similar to ResNet-50 in size. **Performance between IN initialized transformers and CNNs was very similar**, with the latter being, on average, marginally better when using IN weights. Thus, we find the extra space and time to train a VIT-B-16 model compared to ResNet-50 mostly fruitless. **SSL initialization greatly increases transformer performance**, and outperforms the IN one in Fed-ISIC. This is despite our SSL pipeline not being tuned to its full extent, which can likely further increase SSL performance. Hence, we argue that transformer models have a place in medical FL applications where the target domain is dissimilar to ImageNet, and suitable datasets to conduct SSL are available.

4.5. Comparing aggregation methods

In OrganAMNIST, FedOpt, on average, increases test accuracy by 0.68% and 2.4% compared to FedAvg for the IID and non-IID case respectively, while the difference between SCAFFOLD and FedAvg is negligible. In Fed-ISIC, FedOpt led, on average, to a loss of 0.59% balanced accuracy, but SCAFFOLD consistently improved performance (1.32% on average), meaning that **if the extra memory and bandwidth are not an issue, SCAFFOLD is worth considering**. This is in line with its design being for heterogeneous cross-silo, full participation settings, like ours. Despite that, a very important result is that **the best ARIA uses FedAvg** (Table 1, IN pre-training). Overall, **we found the benefits of switching architectures greater than those of switching aggregation methods**, suggesting we need to re-examine how much we have progressed on the algorithmic front since the introduction of FedAvg.

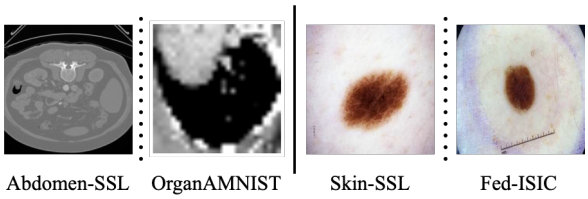

Fig. 1. Samples from the SSL and respective target datasets.

5. CONCLUSION

We conduct the first comprehensive study on ARIAs for federated cross-silo visual classification, giving answers to which parts of an ARIA are most important, and how choices for each compare between them. We find and present evidence that shows FedAvg is still not definitively surpassed, that transformers are not better than CNNs despite recent claims, that IN initialization is beneficial, and, in its absence/non-applicability, SSL also improves performance, as well as the interconnection between these elements. Our work can inform practitioners in the cross-silo setting on which ARIA to employ in real-world scenarios.

6. COMPLIANCE WITH ETHICAL STANDARDS

IEEE-ISBI supports the standard requirements on the use of animal This research study was conducted retrospectively using human subject data made available in open access by [21, 22, 23, 24, 25, 26, 27, 28]. Ethical approval was not required as confirmed by the license attached with the open access data.

7. ACKNOWLEDGMENTS

The authors have no relevant interests to disclose.

8. REFERENCES

[1] Micah J Sheller, Brandon Edwards, G Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R Colen, et al., "Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data," *Scientific reports*, vol. 10, no. 1, pp. 12598, 2020.

[2] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[3] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown, "Measuring the effects of non-identical data distribution for federated visual classification," *arXiv preprint arXiv:1909.06335*, 2019.

[4] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *International conference on machine learning*. PMLR, 2020, pp. 5132–5143.

[5] Sara Pieri, Jose Renato Restom, Samuel Horváth, and Hisham Cholakkal, "Handling data heterogeneity via architectural design for federated visual recognition," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[6] Hong-You Chen, Cheng-Hao Tu, Ziwei Li, Han Wei Shen, and Wei-Lun Chao, "On the importance and applicability of pre-training for federated learning," in *The Eleventh International Conference on Learning Representations*, 2022.

[7] John Nguyen, Jianyu Wang, Kshitiz Malik, Maziar Sanjabi, and Michael Rabbat, "Where to begin? on the impact of pre-training and initialization in federated learning," *arXiv preprint arXiv:2210.08090*, 2022.

[8] Liangqiong Qu, Yuyin Zhou, Paul Pu Liang, Yingda Xia, Feifei Wang, Ehsan Adeli, Li Fei-Fei, and Daniel Rubin, "Rethinking architecture design for tackling data heterogeneity in federated learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10061–10071.

[9] Micah Goldblum, Hossein Souri, Renkun Ni, Manli Shu, Viraj Prabhu, Gowthami Somepalli, Prithvijit Chattopadhyay, Mark Ibrahim, Adrien Bardes, Judy Hoffman, et al., "Battle of the backbones: A large-scale comparison of pre-trained models across computer vision tasks," *arXiv preprint arXiv:2310.19909*, 2023.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Identity mappings in deep residual networks," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, 2016, pp. 630–645.

[11] Sergey Zagoruyko and Nikos Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[12] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[13] Jike Zhong, Hong-You Chen, and Wei-Lun Chao, "Making batch normalization great in federated deep learning," *arXiv preprint arXiv:2303.06530*, 2023.

[14] Andrew Brock, Soham De, and Samuel L Smith, "Characterizing signal propagation to close the performance gap in unnormalized resnets," *arXiv preprint arXiv:2101.08692*, 2021.

[15] Mingxing Tan and Quoc Le, "Efficientnetv2: Smaller models and faster training," in *International conference on machine learning*. PMLR, 2021, pp. 10096–10106.

[16] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11976–11986.

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[18] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10012–10022.

[19] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[20] Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and H Brendan McMahan, "Adaptive federated optimization," *arXiv preprint arXiv:2003.00295*, 2020.

[21] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni, "Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification," *Scientific Data*, vol. 10, no. 1, pp. 41, 2023.

[22] Jean Ogier du Terrail, Samy-Safwan Ayed, Edwige Cyffers, Felix Grimberg, Chaoyang He, Regis Loeb, Paul Mangold, Tanguy Marchand, Othmane Marfoq, Erum Mushtaq, et al., "Flamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings," *Advances in Neural Information Processing Systems*, vol. 35, pp. 5315–5334, 2022.

[23] Holger R Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim B Turkbey, and Ronald M Summers, "Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I 18*. Springer, 2015, pp. 556–564.

[24] Nicholas Heller, Fabian Isensee, Dasha Trofimova, Resha Tejpaul, Zhongchen Zhao, Huai Chen, Lisheng Wang, Alex Golts, Daniel Khapun, Daniel Shats, et al., "The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct," *arXiv preprint arXiv:2307.01984*, 2023.

[25] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al., "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," *arXiv preprint arXiv:1902.09063*, 2019.

[26] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh, "Seven-point checklist and skin lesion classification using multitask multimodal neural nets," *IEEE journal of biomedical and health informatics*, vol. 23, no. 2, pp. 538–546, 2018.

[27] Andre GC Pacheco, Gustavo R Lima, Amanda S Salomao, Breno Krohling, Igor P Biral, Gabriel G de Angelo, Fábio CR Alves Jr, José GM Esgario, Alana C Simora, Pedro BC Castro, et al., "Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones," *Data in brief*, vol. 32, pp. 106221, 2020.

[28] Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, et al., "A patient-centric dataset of images and metadata for identifying melanomas using clinical context," *Scientific data*, vol. 8, no. 1, pp. 34, 2021.

[29] Holger R Roth, Yan Cheng, Yuhong Wen, Isaac Yang, Ziyue Xu, Yuan-Ting Hsieh, Kristopher Kersten, Ahmed Harouni, Can Zhao, Kevin Lu, et al., "Nvidia flare: Federated learning from simulation to real-world," *arXiv preprint arXiv:2210.13291*, 2022.