# Looking Beyond Aggregation for Medical Federated Learning: From Analysis to Novel Architecture Design

**Anonymous Author(s)**
Affiliation
Address
email

## 1 Introduction & Related Work

Federated Learning (FL) [1] offers a privacy-preserving pathway for collaborative model development across medical institutions, making it particularly valuable for medical imaging applications where data cannot be shared [2]. However, the statistical heterogeneity of multi-center medical data, where each institution's images vary in acquisition protocols, patient populations, and disease prevalence, severely challenges FL performance [3]. Medical FL research has extensively focused on developing improved aggregation algorithms to combat this heterogeneity, meaning two other aspects of the pipeline, namely the initialization strategy and the model architecture, have remained an under-explored frontier. For initialization, we know task-relevant pre-training through self-supervised learning (SSL) is a highly-effective alternative to ImageNet (IN) pre-training [4], even more so in the data-scarce and costly annotation medical landscape, but the potential of SSL in the FL setting remains largely unexplored. In terms of architectures, it is common for FL papers to present novel aggregation methods tested on shallow/toy networks [5], which do not mirror the deep and complex architectures deployed in real-world applications. In this paper, we present a two-stage investigation to fill this gap.

First, we conducted a large-scale empirical study systematically evaluating the interplay between Architectures, Initialization strategies, and Aggregation methods (ARIAs). This study, among other findings, conclusively demonstrated that architectural choice, particularly the use of BatchNorm (BN) [6], is a dominant factor in FL performance, often outweighing the benefits of advanced aggregation algorithms. This echoes prior work which has showed BN hinders performance in heterogeneous FL due to mismatched client-specific statistics and inconsistent parameter averaging [7, 8]. In response, using other feature normalization methods like Group Normalization (GN) [9] and Layer Normalization (LN) [10] has been frequent in FL research [11, 12, 13, 14]. These alternatives slow convergence and reduce performance compared to BN [15, 16, 17].

Second, guided by these findings, we designed a BN-free architecture that combines weight standardization [18] with channel attention [19] to directly tackle the challenges posed by non-IID data. Weight standardization normalizes convolutional layer weights instead of activations, avoiding reliance on mini-batch statistics, which is problematic in FL. Channel attention generates learnable scaling factors for feature maps, suppressing features that are inconsistent across clients due to heterogeneity, and emphasizing consistent ones. By integrating channel attention with weight-standardized models, we enhance the model's ability to focus on shared, informative features across clients. Our architecture, which we name ANFR (Adaptive Normalization-free Feature Recalibration), is designed from first principles for the statistical realities of FL, providing a versatile and powerful backbone for federated medical imaging applications.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

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) |

## 2 Methodology

To quantitatively understand architectural impacts in FL, we conducted an exhaustive benchmark evaluating 9 modern architectures, spanning convolutional networks (ResNet-18/50 [20], Wide-ResNet-50-2 [21], DenseNet-121[22], NF-ResNet-50[23], EfficientNetV2-S[24], ConvNext-S[25]) and transformers (ViT-B-16[26], SWIN-T[27]) across three initialization strategies (random weights, IN pre-training, domain-specific SSL) and three aggregation methods (FedAvg[1], FedOpt[12], SCAFFOLD[28]). We evaluated these ARIAs on the tasks of skin lesion classification on Fed-ISIC2019[29] and abdominal organ classification on OrganAMNIST[30]. For the SSL component, our *Skin-SSL* pretraining dataset was created from 3 skin lesion datasets [31, 32, 33], while the *Abdomen-SSL* dataset was created by extracting 20 slices around the center of each volume in 4 abdominal CT datasets[34, 35, 36], cropping around the subject, resizing to 224x224 and copying the channel over, resulting in 21,000 whole abdomen images.

Next, guided by the ARIA findings, we developed ANFR as an architectural solution specifically designed for FL's statistical challenges. ANFR eliminates dependency on batch-specific statistics through two synergistic components, Scaled Weight Standardization and Adaptive Feature Recalibration. 1) Scaled Weight Standardization (SWS)[23]: instead of normalizing activations, ANFR normalizes convolutional weights themselves using carefully scaled standardization that maintains signal propagation stability. This ensures consistent forward passes regardless of client-specific data distributions, removing the statistical conflicts inherent in BN during federation. 2) Adaptive Feature Recalibration: to actively combat heterogeneity and compensate for lost regularization, we integrate channel attention mechanisms after weight-standardized layers. This enables dynamic suppression of client-specific noisy features while amplifying universally informative patterns. We validate ANFR on Fed-ISIC2019, FedChest (our own multi-label chest X-ray dataset with 4 clients and covariate shift), and CIFAR-10 [37]. To benchmark ANFR, we perform focused, ablated studies against strong baselines: BN-ResNet, GN-ResNet, SE-ResNet [19], and NF-ResNet[23]. All models are evaluated under multiple aggregation methods (FedAvg, FedProx, SCAFFOLD, FedAdam).

## 3 Results

**Finding 1: Architecture Dominates Performance.** Architecture choice yielded up to 30% performance differences while aggregation methods typically changed results by <2%. On Fed-ISIC with ImageNet initialization, ConvNeXt-S achieved 75.08% while ResNet-50 reached only 67.97%. The

Table 3: Performance comparison across all architectures under different global FL aggregation methods and different datasets. Best in bold, second best underlined. ANFR consistently outperforms the baselines, often by a wide margin.

| Dataset | Method | Architecture | | | | |
|---|---|---|---|---|---|---|
| | | BN-ResNet | GN-ResNet | SE-ResNet | NF-ResNet | ANFR (Ours) |
| Fed-ISIC2019 | FedAvg | 66.01±0.73 | 65.09±0.42 | 65.29±1.32 | 72.49±0.60 | **74.78±0.16** |
| | FedProx | 66.49±0.41 | 66.51±1.21 | 66.29±0.63 | 71.28±2.14 | **75.61±0.71** |
| | FedAdam | 65.88±0.67 | 64.60±0.39 | 65.18±1.90 | 69.96±0.14 | **73.02±0.93** |
| | SCAFFOLD | 65.41±0.72 | 68.84±0.46 | 68.99±0.18 | 73.30±0.50 | **76.52±0.60** |
| FedChest | FedAvg | 82.80±0.13 | 83.40±0.25 | 82.14±0.18 | 83.40±0.11 | **83.49±0.14** |
| | FedProx | **82.14±0.10** | 82.04±0.08 | 81.50±0.26 | 81.26±0.58 | **82.14±0.10** |
| | FedAdam | 83.02±0.11 | 82.11±0.10 | 82.72±0.16 | 83.10±0.09 | **83.33±0.07** |
| | SCAFFOLD | 83.52±0.14 | 83.95±0.05 | 83.50±0.08 | 84.06±0.02 | **84.26±0.10** |
| CIFAR-10 | FedAvg | 91.71±0.74 | 96.60±0.11 | 94.07±0.04 | 96.72±0.05 | **97.42±0.01** |
| | FedProx | 95.03±0.04 | 96.05±0.04 | 94.60±0.07 | **96.82±0.04** | 96.33±0.09 |
| | FedAdam | 91.23±0.29 | 95.80±0.24 | 94.09±0.17 | 95.54±0.10 | **96.93±0.06** |
| | SCAFFOLD | 92.51±0.99 | 96.78±0.01 | 94.30±0.03 | 96.84±0.01 | **97.38±0.03** |

implication is clear: selecting the right architecture provides much larger gains than sophisticated aggregation algorithms.

**Finding 2: ImageNet Initialization Generally Wins, But SSL Shows Promise.** ImageNet consistently provided best results but medical SSL demonstrated domain-specific value: Skin-SSL achieved 74.56% on Fed-ISIC, nearly matching ImageNet's 75.08%. Critically, SSL enables FL for non-standard medical images (non-RGB, varying resolutions) without introducing aliasing artifacts from forced resizing. For transformers, pre-training proved essential: random initialization yielded 45-48% on Fed-ISIC versus 65-73% with proper initialization.

**Finding 3: Normalization Layers Create FL Bottlenecks.** Batch Normalization's limitations were stark. Normalization-Free ResNet-50 consistently outperformed standard ResNet-50: on Fed-ISIC with random initialization, 55.93% versus 49.11% . Under heterogeneity, BN models suffered 15-20% performance drops while NF models remained stable. This robustness stems from avoiding client-specific statistics that become meaningless when averaged across heterogeneous distributions.

**Finding 4: Model Scaling Paradox.** Contrary to centralized learning intuitions, deeper/wider models underperformed in FL. ResNet-18 beat ResNet-50 in 7/9 Fed-ISIC experiments despite fewer parameters. Wide-ResNet-50-2 (2.7× parameters) showed minimal gains over ResNet-50. DenseNet-121 achieved competitive performance with 68% fewer parameters than ResNet-50. This suggests FL favors architectural efficiency over raw capacity—likely because larger models are harder to optimize under non-IID conditions.

**Finding 5: Aggregation Methods Have Limited but Consistent Effects.** SCAFFOLD provided modest improvements ($\sim$1.3% on Fed-ISIC) but required 3× communication overhead. FedOpt showed mixed results: helping heterogeneous OrganAMNIST (+2.4%) but hurting Fed-ISIC (-0.59%), with high sensitivity to server learning rate. Remarkably, simple FedAvg remained highly competitive, often achieving the best results. The marginal gains from complex aggregation methods further emphasize that architectural improvements offer more promising returns.

**Finding 6: ANFR provides universal performance benefits.** As see on 3, our proposed architecture manages to consistently outperform the baselines across all datasets and aggregation methods. This serves as clear evidence more research into architectures for FL can advance the field just as much, if not more than, aggregation methods.

The results yielded a clear verdict: architectural selection consistently outweighed aggregation improvements, with networks employing Batch Normalization suffering dramatic performance drops up to 14% in balanced accuracy under heterogeneous conditions. This systematic analysis identified BN's dependency on consistent batch statistics as a fundamental architectural weakness in the FL setting, motivating our architectural redesign.

## 4 Potential Negative Societal Impact

While our study provides valuable insights, there are potential negative impacts to consider. The emphasis on performance metrics could overshadow critical considerations such as model fairness across different patient demographics, interpretability for clinical decision-making, and robustness to distribution shifts in real-world medical settings. Finally, while SSL pre-training shows promise for non-ImageNet domains, the requirement to create large-scale medical SSL datasets may be infeasible for rare diseases or under-represented populations, potentially widening healthcare disparities .

## References


[1] 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*, pages 1273–1282. PMLR, 2017.

[2] 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*, 10(1):12598, 2020.

[3] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletarì, Holger R. Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N. Galtier, Bennett A. Landman, Klaus Maier-Hein, Sébastien Ourselin, Micah Sheller, Ronald M. Summers, Andrew Trask, Daguang Xu, Maximilian Baust, and M. Jorge Cardoso. The future of digital health with federated learning. *npj Digital Medicine*, 3(1):1–7, September 2020. Number: 1 Publisher: Nature Publishing Group.

[4] 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 pretrained models across computer vision tasks. *arXiv preprint arXiv:2310.19909*, 2023.

[5] Sara Pieri, Jose Restom, Samuel Horvath, and Hisham Cholakkal. Handling data heterogeneity via architectural design for federated visual recognition. *Advances in Neural Information Processing Systems*, 36:4115–4136, 2023.

[6] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

[7] Yanmeng Wang, Qingjiang Shi, and Tsung-Hui Chang. Why batch normalization damage federated learning on non-iid data? *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[8] Rachid Guerraoui, Rafael Pinot, Geovani Rizk, John Stephan, and François Taiani. Overcoming the challenges of batch normalization in federated learning. *arXiv preprint arXiv:2405.14670*, 2024.

[9] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[10] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016.

[11] Kevin Hsieh, Amar Phanishayee, Onur Mutlu, and Phillip Gibbons. The non-iid data quagmire of decentralized machine learning. In *International Conference on Machine Learning*, pages 4387–4398. PMLR, 2020.

[12] Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and H. Brendan McMahan. Adaptive federated optimization, 2021.

[13] Jianyu Wang, Zachary Charles, Zheng Xu, Gauri Joshi, H. Brendan McMahan, Blaise Aguera y Arcas, Maruan Al-Shedivat, Galen Andrew, Salman Avestimehr, Katharine Daly, Deepesh Data, Suhas Diggavi, Hubert Eichner, Advait Gadhikar, Zachary Garrett, Antonious M. Girgis,

Filip Hanzely, Andrew Hard, Chaoyang He, Samuel Horvath, Zhouyuan Huo, Alex Ingerman, Martin Jaggi, Tara Javidi, Peter Kairouz, Satyen Kale, Sai Praneeth Karimireddy, Jakub Konecny, Sanmi Koyejo, Tian Li, Luyang Liu, Mehryar Mohri, Hang Qi, Sashank J. Reddi, Peter Richtarik, Karan Singhal, Virginia Smith, Mahdi Soltanolkotabi, Weikang Song, Ananda Theertha Suresh, Sebastian U. Stich, Ameet Talwalkar, Hongyi Wang, Blake Woodworth, Shanshan Wu, Felix X. Yu, Honglin Yuan, Manzil Zaheer, Mi Zhang, Tong Zhang, Chunxiang Zheng, Chen Zhu, and Wennan Zhu. A field guide to federated optimization, 2021.

[14] Zhixu Du, Jingwei Sun, Ang Li, Pin-Yu Chen, Jianyi Zhang, Hai" Helen" Li, and Yiran Chen. Rethinking normalization methods in federated learning. In *Proceedings of the 3rd International Workshop on Distributed Machine Learning*, pages 16–22, 2022.

[15] Hong-You Chen and Wei-Lun Chao. Fedbe: Making bayesian model ensemble applicable to federated learning, 2021.

[16] Irene Tenison, Sai Aravind Sreeramadas, Vaikkunth Mugunthan, Edouard Oyallon, Irina Rish, and Eugene Belilovsky. Gradient masked averaging for federated learning, 2023.

[17] Jike Zhong, Hong-You Chen, and Wei-Lun Chao. Making batch normalization great in federated deep learning, 2024.

[18] Siyuan Qiao, Huiyu Wang, Chenxi Liu, Wei Shen, and Alan Yuille. Micro-batch training with batch-channel normalization and weight standardization, 2020.

[19] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[20] 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*, pages 630–645. Springer, 2016.

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

[22] 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*, pages 4700–4708, 2017.

[23] Andrew Brock, Soham De, and Samuel L. Smith. Characterizing signal propagation to close the performance gap in unnormalized resnets, 2021.

[24] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021.

[25] 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*, pages 11976–11986, 2022.

[26] 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.

[27] 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*, pages 10012–10022, 2021.

[28] 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*, pages 5132–5143. PMLR, 2020.

[29] 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*, 35:5315–5334, 2022.

[30] 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*, 10(1):41, 2023.

[31] 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*, 23(2):538–546, 2018.

[32] 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*, 32:106221, 2020.

[33] 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*, 8(1):34, 2021.

[34] 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*, pages 556–564. Springer, 2015.

[35] 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.

[36] 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.

[37] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *University of Toronto*, 2009.

# NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [NA]

   Justification: Short format paper where there are no traditional abstract and introduction sections.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [No]

   Justification: Short format paper, no space for limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: No theoretical results were presented.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [No]

Justification: Most hyper-parameters and implementation dtails had to be left out due to space constraints.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification:

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [No]

   Justification:

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification:

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification:

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification:

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification:

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification:

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification:

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification:

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: