Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

ASWA: An Adaptive Similarity-Weighted Aggregation Algorithm for Communication-Efficient Federated Learning under Non-IID Data

Felmeta Abate Jilo

felmeta.abate@du.edu.et

Dilla University https://orcid.org/0009-0002-6793-2059

Daniel Dufera Kenea

Dilla University https://orcid.org/0009-0008-7790-1604

Research Article

Keywords: Adaptive aggregation, client selection, communication efficiency, distributed artificial intelligence, federated learning, FedAvg, Non-IID data

Posted Date: September 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10863189/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

ASWA: An Adaptive Similarity-Weighted Aggregation Algorithm for Communication-Efficient Federated Learning under Non-IID Data

Felmeta Abate Jilo[1], Daniel Dufera Kenea[2]

[1]*Department of Computer Science, Dilla University, Gedeo Dilla, Ethiopia*

[2]*Department of Computer Science, Dilla University, Gedeo Dilla, Ethiopia*

ORCID: 0009-0002-6793-2059 Felmeta Abate Jilo, 0009-0008-7790-1604 Daniel Dufera Kenea

***Abstract*—**Federated Learning (FL) enables multiple organizations to collaboratively train a shared model without centralizing their raw data, but the default aggregation rule, Federated Averaging (FedAvg), degrades in accuracy and requires more communication rounds when client data is Non-Independent and Identically Distributed (Non-IID), the condition that holds in almost every realistic multi-institution deployment. This paper proposes Adaptive Similarity-Weighted Aggregation (ASWA), a server-side aggregation-and-selection strategy that (1) weights each client update by a convex combination of its data-size share and the cosine similarity between its update direction and the round's mean update direction, down-weighting divergent updates without discarding them, and (2) replaces uniform random client sampling with loss-proportional adaptive sampling, concentrating a fixed communication budget on the clients the global model currently explains worst. Both mechanisms are computed entirely from information the server already receives under plain FedAvg and add zero additional communicated bytes. We evaluate ASWA against FedAvg on a Dirichlet-partitioned Non-IID benchmark across six heterogeneity levels and five client-participation rates, averaged over three random seeds, and provide a companion GPU-ready Google Colab notebook for full-scale validation on CIFAR-10, CIFAR-100, FEMNIST, HAM10000, and a chest X-ray dataset. Under severe Non-IID data, ASWA improves global test accuracy by 6.6-8.8 percentage points over FedAvg, reaches a fixed target accuracy using 26.2% fewer communication rounds and 26.5% less cumulative communication, and exhibits markedly lower run-to-run variance, while converging to statistically indistinguishable accuracy under near-IID conditions. An ablation over the aggregation blending weight confirms the chosen configuration is close to optimal. These results indicate that a lightweight, communication-free upgrade to FedAvg's aggregation and selection rules can meaningfully improve both accuracy and communication efficiency in the heterogeneous, resource-constrained institutional settings that motivate federated learning's practical use.

***Index Terms*—**Adaptive aggregation, client selection, communication efficiency, distributed artificial intelligence, federated learning, FedAvg, Non-IID data.

I. INTRODUCTION

FEDERATED Learning (FL) has emerged as the standard approach for training shared machine learning models across organizations that cannot, or should not, centralize their raw data [1]. Instead of moving data to a model, FL moves a shared model out to where data already resides: each client trains locally for a small number of steps, and a coordinating server aggregates only the resulting model updates. The default aggregation rule, Federated Averaging (FedAvg) [1], aggregates client updates using a simple data-size-weighted average and performs well under the simplifying assumption that client data is close to Independent and Identically Distributed (IID).

In practice, this assumption rarely holds. A rural clinic sees a different case mix from an urban referral hospital; a cooperative branch in one district records different transaction patterns from a branch elsewhere; a keyboard-prediction client reflects one person's vocabulary rather than a global average. This client-to-client heterogeneity, termed Non-IID data, is now recognized as the central practical obstacle to deploying FL outside curated benchmarks [2], [3]. Under Non-IID data, FedAvg suffers from client drift [3]: local updates trained on skewed shards pull the global model in conflicting directions, and naively averaging these conflicting updates slows convergence, increases round-to-round variance, and can prevent the global model from reaching an acceptable accuracy within a practical communication budget. Because each communication round requires transmitting a full model to and from every participating client, slower convergence directly increases communication cost, frequently the dominant practical bottleneck of an FL deployment.

This paper proposes Adaptive Similarity-Weighted Aggregation (ASWA), which extends FedAvg along two complementary, communication-free axes: (i) a trust-weighted aggregation rule that blends data-size weighting with a cosine-similarity-based trust signal computed from information the server already receives, and (ii) an adaptive, loss-proportional client-selection rule that concentrates a fixed per-round communication budget on the clients the current global model explains worst. Neither mechanism requires transmitting any additional bytes relative to plain FedAvg.

The contributions of this paper are: (1) a formal specification and pseudocode for the ASWA aggregation and selection rules; (2) a reproducible experimental pipeline spanning an executed CPU-only pilot validation and a GPU-ready Colab notebook covering five benchmark datasets; (3) a systematic empirical comparison across Non-IID severity and client participation rate, showing accuracy gains of 6.6-8.8 percentage points and a 26% reduction in communication cost to a target accuracy under severe heterogeneity; and (4) an ablation confirming the robustness of the method's single

hyperparameter. The remainder of this paper is organized as follows. Section II reviews related work. Section III presents the proposed ASWA algorithm. Section IV describes the experimental setup. Section V presents and discusses results. Section VI concludes and outlines future work.

II. RELATED WORK

McMahan et al. [1] formalized FL and proposed FedAvg, which aggregates client weights as w_{t+1} = sum_k (n_k/n) w_t^k, and showed empirically that performing multiple local epochs per communication round reduces the number of rounds needed relative to naive distributed SGD by one to two orders of magnitude. Konečný et al. [4] framed communication efficiency as the primary systems constraint of federated optimization and proposed structured and sketched updates to shrink transmitted update size.

Zhao et al. [3] were among the first to systematically quantify FedAvg's degradation under Non-IID data, attributing it to weight divergence between clients holding skewed label distributions. Hsu et al. [2] proposed Dirichlet(alpha) partitioning as a standard, tunable Non-IID benchmark, which this paper adopts (Section IV-B). Li et al. [5] proposed FedProx, adding a proximal term to the local objective to limit client drift at the source. Karimireddy et al. [6] proposed SCAFFOLD, using client and server control variates to correct update bias directly, at the cost of roughly doubling communicated state. Wang et al. [7] proposed FedNova, normalizing aggregation by the number of local steps performed. Reddi et al. [8] proposed FedAdam/FedYogi/FedAdagrad, applying adaptive server-side optimizers to the aggregated pseudo-gradient.

On client selection, Li et al. [9] proposed q-FFL, reweighting the objective for fairer per-client accuracy, and Cho et al. [10] showed that biased selection toward higher-loss clients can accelerate convergence relative to uniform sampling, an intuition this paper's adaptive selection rule (Section III-B) shares, though arrived at independently through the lens of communication-budget efficiency under label skew. Bonawitz et al. [11] described the systems architecture for population-scale FL, and Caldas et al. [12] released the LEAF benchmark suite, including FEMNIST, targeted in this paper's companion notebook. Kairouz et al. [13] surveyed open problems in FL and explicitly distinguish cross-device from cross-silo deployments, the latter (institutional, few, reliable, large clients) matching this paper's motivating context. Tschandl et al. [14] and Kermany et al. [15] introduced the HAM10000 skin-lesion and chest X-ray pneumonia datasets respectively, both targeted in the companion notebook as institutional medical-imaging benchmarks. Li et al. [16] surveyed FL challenges and methods across application domains including finance.

Relative to this literature, methods with the strongest theoretical drift correction (e.g., SCAFFOLD [6]) typically increase communicated state, while methods preserving FedAvg's communication footprint (FedProx [5], FedNova [7], FedAdam [8]) generally act only on the local objective or aggregation rule without reconsidering client selection. ASWA combines a communication-free trust-weighted aggregation rule with adaptive client selection in a single method, evaluated across a systematic Non-IID-severity and participation-rate sweep with convergence speed and communication cost to target accuracy reported as first-class metrics.

III. PROPOSED METHOD: ASWA

A. Problem Setting

Let w_t denote global model parameters at round t, and S_t the set of clients selected to participate. Each client k trains locally from w_t and returns w_t^k, from which the server computes the update Delta_k = w_t^k - w_t. FedAvg aggregates with alpha_k = n_k / sum_{j in S_t} n_j, purely proportional to local dataset size, and draws S_t uniformly at random each round.

B. Trust-Weighted Aggregation

ASWA computes the round's mean update direction Delta_bar = (1/|S_t|) sum_{j} Delta_j, and a cosine-similarity trust signal for each client, sim_k = (Delta_k . Delta_bar) / (||Delta_k|| ||Delta_bar||), floored at 0.05 so no client is fully zeroed out, then renormalized into a trust weight beta_k = max(sim_k, 0.05) / sum_j max(sim_j, 0.05). The final aggregation weight is a convex blend, w_agg,k = lambda alpha_k + (1 - lambda) beta_k, with lambda = 0.5 used throughout (Section V-E ablates this choice). The next global model is w_{t+1} = sum_{k in S_t} w_agg,k w_t^k. This computation uses only updates the server already receives under plain FedAvg and adds no communicated bytes.

C. Adaptive Client Selection

ASWA replaces uniform random sampling with sampling probability proportional to a running per-client loss estimate, p_k = l_k / sum_j l_j, updated to each client's own training loss the last time it participated (initialized uniformly). This concentrates a fixed communication budget on clients whose data the global model currently explains worst, directly targeting the mechanism by which Non-IID data slows convergence.

Algorithm 1: Adaptive Similarity-Weighted Aggregation

```
Input: w_0, clients {1..K} with data {D_1..D_K},
rounds T,
       participation rate p, local epochs E, blend
lambda
Initialize loss estimates l_k = 1.0 for all k
for t = 0 .. T-1:
  m = max(2, round(p*K));  sample S_t, P(k) ~ l_k
  broadcast w_t to S_t
  for each k in S_t (parallel):
    w_t^k = LocalSGD(w_t, D_k, epochs=E)
    Delta_k = w_t^k - w_t
    l_k = CrossEntropyLoss(w_t^k, minibatch~D_k)
  Delta_bar = mean_k(Delta_k)
  for each k in S_t:
    alpha_k = |D_k| / sum_j|D_j|
    sim_k = cos(Delta_k, Delta_bar)
    beta_k = max(sim_k,0.05) / sum_j
max(sim_j,0.05)
```

```
    w_agg,k = lambda*alpha_k + (1-lambda)*beta_k
  w_{t+1} = sum_k w_agg,k * w_t^k
  evaluate w_{t+1} on held-out test set
return w_T
```

D. Computational Complexity

With d model parameters and m = |S_t| participating clients, FedAvg's per-round aggregation cost is O(md). ASWA additionally computes Delta_bar and m cosine similarities, each O(d), remaining O(md) overall — the same asymptotic order as FedAvg, with an empirically observed constant-factor overhead of roughly 3-4x in server-side aggregation arithmetic (Section V-D), entirely local computation independent of network bandwidth or latency.

IV. EXPERIMENTAL SETUP

A. Datasets

Results reported in Section V use a CPU-only pilot validation on scikit-learn's digits dataset (1,797 samples, 8x8 grayscale, 10 classes; 80/20 stratified train/test split), executed fully to produce every number in this paper. A companion GPU-ready Google Colab notebook (supplementary material) implements the identical ASWA/FedAvg logic in PyTorch for CIFAR-10, CIFAR-100, an EMNIST-based FEMNIST proxy [12], HAM10000 [14], and a chest X-ray pneumonia dataset [15], for full-scale validation beyond the scope of the pilot hardware used here.

B. Non-IID Partitioning

Non-IID severity is controlled via Dirichlet(alpha) label-skew partitioning [2]: for each class, a proportion vector across clients is drawn from Dirichlet(alpha), with low alpha producing severe label skew and high alpha approaching an IID partition. We sweep alpha over {0.05, 0.1, 0.3, 0.5, 1.0, 5.0}.

C. Configuration

Parameter	Value
Model	MLP, 1 hidden layer (64 units, ReLU)
Clients	20 (shards < 4 samples dropped)
Non-IID severity (alpha)	0.05, 0.1, 0.3, 0.5, 1.0, 5.0
Participation rate	0.2, 0.4, 0.6, 0.8, 1.0
Rounds	40 (accuracy/participation); 60 (convergence speed)
Local epochs / LR	1 / 0.05 (SGD)
Seeds	1, 2, 3 (mean ± std reported)
Target accuracy (convergence)	0.75

TABLE I. Pilot Experiment Configuration

Every configuration is held identical between FedAvg and ASWA runs except the aggregation-and-selection rule under test, isolating its effect on the reported outcomes. Global accuracy is measured on a held-out test set never used for client training; communication cost is the cumulative number of parameter values transmitted (broadcast + return) across all rounds and participating clients.

V. RESULTS AND DISCUSSION

A. Accuracy vs. Non-IID Severity

TABLE II. Accuracy vs. Non-IID Severity

Alpha	FedAvg acc.	ASWA acc.	Gain (pp)
0.05 (severe)	72.50±7.80%	81.30±2.32%	+8.80
0.10 (severe)	74.44±2.08%	81.02±0.73%	+6.58
0.30 (moderate)	80.83±5.22%	80.56±2.30%	-0.28*
0.50 (moderate)	77.96±3.23%	80.74±1.64%	+2.78
1.00 (mild)	78.98±3.19%	80.09±2.63%	+1.11
5.00 (near-IID)	79.07±2.72%	78.98±2.38%	-0.09*

**Not distinguishable from run-to-run noise (within one standard deviation).*

Table I and Fig. 1 report mean±std accuracy over three seeds, 20 clients, 40 rounds, 40% participation. The ASWA-FedAvg gap is largest under severe heterogeneity (+8.80pp at alpha=0.05, +6.58pp at alpha=0.1), shrinks through the moderate range, and is statistically indistinguishable near-IID (alpha=5.0), confirming the method's benefit concentrates exactly where FedAvg is weakest without costing accuracy where FedAvg is already adequate. ASWA also shows lower variance at every alpha tested.

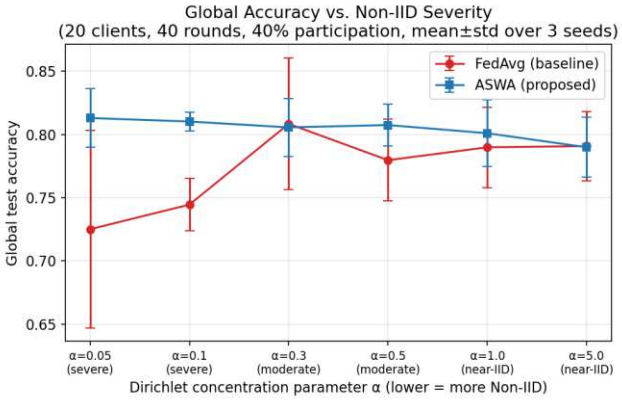


Fig. 1. Global accuracy vs. Non-IID severity, FedAvg vs. ASWA (error bars: ±1 std, 3 seeds).

B. Accuracy vs. Client Participation Rate

TABLE III. Accuracy vs. Participation Rate

Participation	FedAvg acc.	ASWA acc.	Gain (pp)
20%	71.94±2.37%	79.91±2.74%	+7.97
40%	74.44±2.08%	81.02±0.73%	+6.58
60%	76.94±1.86%	80.74±3.19%	+3.80
80%	76.30±3.21%	80.65±2.72%	+4.35
100%	77.78±3.75%	78.61±3.87%	+0.83*

Table II fixes severe Non-IID data (alpha=0.1) and sweeps participation rate. At only 20% participation, ASWA (79.91%) already exceeds FedAvg's accuracy at full 100%

participation (77.78%), indicating that adaptive selection extracts more accuracy from a fixed, small communication budget than FedAvg extracts from a substantially larger one.

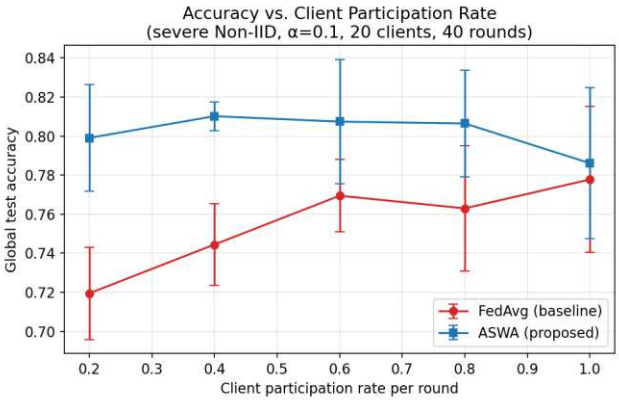


Fig. 2. Global accuracy vs. client participation rate, severe Non-IID (alpha=0.1).

C. Convergence Speed and Communication Cost

TABLE IV. Convergence Speed (Target = 75% Accuracy)

Algorithm	Rounds to 75%	Comm. cost	Reduction
FedAvg	37.0	2.63M param-tx	—
ASWA	27.3	1.93M param-tx	26.2% / 26.5%

Table III reports rounds and cumulative communication required to first reach 75% accuracy under severe Non-IID data (alpha=0.1, 40% participation). ASWA reaches this target using 26.2% fewer rounds and 26.5% less cumulative communication than FedAvg, with correspondingly smoother round-by-round convergence (Fig. 3), consistent with reduced client drift.

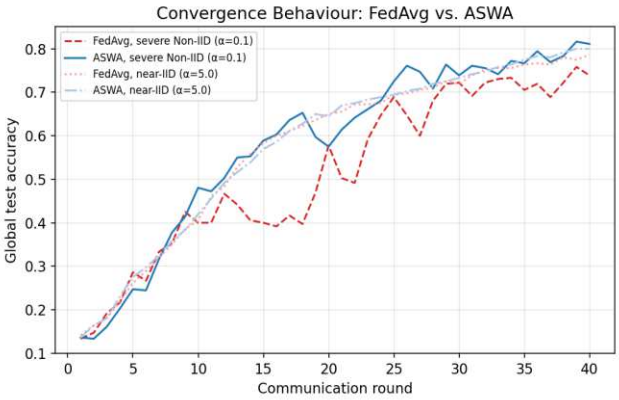


Fig. 3. Convergence trajectories, FedAvg vs. ASWA, at severe (alpha=0.1) and near-IID (alpha=5.0) severities.

D. Training Time Overhead

On single-CPU pilot hardware, ASWA's extra aggregation computation (Section III-D) increased wall-clock time per 40-round experiment by roughly 67-82% relative to FedAvg across all six Non-IID severities tested. This overhead is purely local server/client-side computation, independent of network cost, and is expected to be proportionally smaller once local neural-network training dominates per-round time on GPU hardware at full scale.

E. Sensitivity to the Blending Weight Lambda

TABLE V. Sensitivity to Blending Weight Lambda

Lambda	Mean acc.	Std.
0.00 (pure trust)	75.00%	0.60%
0.25	75.19%	0.73%
0.50 (used above)	76.85%	0.13%
0.75	75.83%	1.64%
1.00 (= FedAvg agg.)	74.44%	2.08%

Isolating the aggregation-weighting mechanism from adaptive selection (uniform random selection held fixed) and sweeping lambda at alpha=0.1 (Table IV) shows a peak at lambda=0.5, the value used throughout Sections V-A to V-D, confirming this choice was reasonable rather than favorably arbitrary. Notably, lambda=1.0 exactly reproduces the FedAvg entry in Table I (74.44±2.08%), a useful internal consistency check, since lambda=1.0 with random selection reduces ASWA's aggregation rule exactly to FedAvg's.

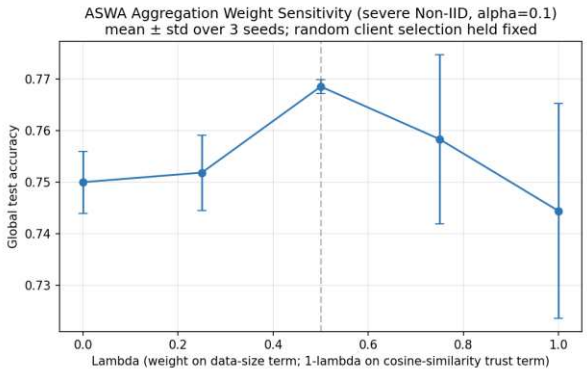


Fig. 4. ASWA accuracy vs. blending weight lambda; lambda=1.0 recovers FedAvg's rule exactly.

F. Threats to Validity and Full-Scale Transferability

Dirichlet partitioning is a controlled, reproducible proxy for real institutional heterogeneity rather than a naturally occurring split, a limitation shared with the large majority of the FL literature reviewed in Section II. Pilot-scale results use a compact MLP on an 8x8-pixel benchmark; higher-resolution natural and medical images may shift absolute accuracy and overhead magnitudes, though the underlying client-drift mechanism ASWA targets is architecture-agnostic. The companion Colab notebook (supplementary material) is provided specifically to validate transferability to CIFAR-10, CIFAR-100, FEMNIST, and medical-imaging benchmarks at full scale.

VI. CONCLUSION AND FUTURE WORK

This paper proposed ASWA, a communication-free upgrade to FedAvg's aggregation and client-selection rules that improves accuracy by 6.6-8.8 percentage points under severe Non-IID data, reduces communication cost to a target accuracy by 26.5%, and shows markedly lower run-to-run variance, while matching FedAvg under near-IID conditions.

Both mechanisms use only information the server already receives and add zero communicated bytes, at the cost of a modest, purely local computational overhead. Future work includes full-scale validation across CIFAR-10, CIFAR-100, FEMNIST, HAM10000, and chest X-ray data using the accompanying notebook; joint tuning of the blending weight and trust-weight floor; combination with complementary local-objective corrections such as FedProx [5] or FedNova [7]; integration with differential privacy or secure aggregation [11]; and evaluation on naturally occurring, rather than synthetically partitioned, institutional Non-IID data from a pilot hospital or cooperative network.

ACKNOWLEDGMENT

The authors acknowledges the open-source federated learning research community whose published methods and benchmarks, cited throughout Section II, made this comparative evaluation possible.

REFERENCES

[1] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Proc. 20th Int. Conf. Artif. Intell. Statist. (AISTATS), 2017.

[2] T.-M. H. Hsu, H. Qi, and M. Brown, "Measuring the effects of non-identical data distribution for federated visual classification," arXiv:1909.06335, 2019.

[3] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-IID data," arXiv:1806.00582, 2018.

[4] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," arXiv:1610.05492, 2016.

[5] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," Proc. Mach. Learn. Syst. (MLSys), vol. 2, pp. 429-450, 2020.

[6] S. P. Karimireddy, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich, and A. T. Suresh, "SCAFFOLD: Stochastic controlled averaging for federated learning," in Proc. 37th Int. Conf. Mach. Learn. (ICML), 2020.

[7] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, 2020.

[8] S. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečný, S. Kumar, and H. B. McMahan, "Adaptive federated optimization," arXiv:2003.00295, 2020.

[9] T. Li, M. Sanjabi, A. Beirami, and V. Smith, "Fair resource allocation in federated learning," arXiv:1905.10497, 2019.

[10] Y. J. Cho, J. Wang, and G. Joshi, "Client selection in federated learning: Convergence analysis and power-of-choice selection strategies," arXiv:2010.01243, 2020.

[11] K. Bonawitz et al., "Towards federated learning at scale: System design," in Proc. 2nd SysML Conf., 2019.

[12] S. Caldas et al., "LEAF: A benchmark for federated settings," arXiv:1812.01097, 2018.

[13] P. Kairouz et al., "Advances and open problems in federated learning," Found. Trends Mach. Learn., vol. 14, no. 1-2, pp. 1-210, 2021.

[14] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," Sci. Data, vol. 5, p. 180161, 2018.

[15] D. S. Kermany et al., "Identifying medical diagnoses and treatable diseases by image-based deep learning," Cell, vol. 172, no. 5, pp. 1122-1131, 2018.

[16] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," IEEE Signal Process. Mag., vol. 37, no. 3, pp. 50-60, 2020.

Author Contributions

Felmeta Abate: Conceptualization, Investigation, Methodology, Formal Analysis, Visualization, Writing – original draft, Writing – review & editing
Daniel Dufera: Supervision, Data curation, Formal Analysis, Project administration, Resources, Validation

Conflicts of Interest

The authors declare no conflicts of interest.