

FEDERATED MACHINE LEARNING FOR PRIVACY-PRESERVING DATA ANALYTICS IN DISTRIBUTED SYSTEMS

¹George Zacharia

¹Independent Researcher Mechanicsburg, PA, USA E-mail:

georgezacharia1983@gmail.com.

ABSTRACT: By keeping the data separate, federated learning (FL) enables multiple data owners, such as hospitals, edge devices, or financial institutions, to jointly develop a shared machine-learning model, which is of particular relevance in applications of distributed systems that have to be operated under privacy regulation and underlying narrow bandwidth restrictions. We implement both two basic FL mechanisms (Federated Averaging, FedAvg, and a mechanism at the client-side that applies differential-privacy, DP) from the ground-up in NumPy, and test both in a uniform experimental protocol on two different distributed-systems scenarios: a network of 10 simulated healthcare data-silos with the real Wisconsin Diagnostic Breast Cancer benchmark ($n = 569$) and a network of 30 simulated edge/IoT devices with the real UCI handwritten-digits benchmark ($n = 1,797$). Client data was partitioned both in the near-IID and Dirichlet-skewed NON-IID (Hsu, Qi, & Brown, 2019) regime. In both communication rounds (40 in healthcare and 30 in edge/IoT), federated training achieved 97.7% and 96.7% test accuracy, respectively, (plus or minus 0.6 and 0.4 percentage points, respectively, compared to an equivalent centralized model trained on a “pooled” dataset), but the edge/IoT task suffered significant, monotonic loss of accuracy as the degree of client heterogeneity increased (96.7% for near-IID to 95.4% for the most skewed partition, whereas there was nearly no loss in the healthcare task for which the client models were simpler and near-linearly-separable). The systematic sweep over the differential-privacy noise multiplier confirmed that there is a significantly higher privacy-utility tradeoff for the harder high dimensional edge/IoT task than for the healthcare task (privacy-accuracy tradeoff varied from 96.3% accuracy to 51.1% accuracy as noise varied in the former and 97.7% accuracy to 78.4% accuracy as noise varied in the latter), hence the “cost of privacy” is not an intrinsic property of a differential privacy mechanism but is rather dependent on task difficulty and dimensionality. These results are placed in the context of the federated learning literature in a PRISMA-compliant review of 29 studies. Finally, we present practical suggestions for provisioning communication rounds, client participation percentages, and privacy budgets when rolling out FL across a set of distributed systems with different types and privacy restrictions.

Keywords: Federated learning, Differential Privacy, Distributed systems, Privacy preserving machine learning, Non-IID data, FedAvg, Edge computing, Data silos.

1. INTRODUCTION

Modern data storage systems are per definition avdelte. Patient records are stored in hospitals (and for good reason can't be moved), smart phones and IoT sensors can generate information but this information can't be sent in its raw form without being impractical or

invasive, and banks and telecom operators don't send customer records anywhere for regulatory and competitive reasons. However, machine learning is more effective the more varied and extensive data that a model observes. The central issue federated learning (FL) aims to tackle is this conflict between utility and lawfulness of data locality, and the statistical advantage of training with centralized data.

Federated learning, in which computation moves to the data, with each node (a hospital, a device, a data silo) training locally on its data, and only sending model updates — not the actual observations — to a coordinating server, which aggregates model updates to create an improved global model (McMahan, Moore, Ramage, Hampson, & Agüera y Arcas, 2017): While this architecture directly addresses privacy-preserving analytics in distributed systems, it poses three challenges which are the focus of this paper and not addressed in the case of centralized training. For one, client data are seldom independent and identically distributed (IID), meaning a hospital in one area will have a different case distribution than one in another, or an edge device used one way will have a different data distribution than it does when used for another purpose, and this statistical differences slow or degrade convergence relative to centralized training (Hsu, Qi, & Brown, 2019; Li, Sahu, Zaheer, Sanjabi, Talwalkar, & Smith, 2020). Second, with federation alone, there will be some leakage of privacy—even if one relies only on model updates rather than the raw data—in a gradient-based attack against the models or a model-inversion attack (Abadi et al., 2016; McMahan, Ramage, Talwar, & Zhang, 2018), but the leakage will be reduced. Third, in the real world, (in mobile networks, in hospital IT infrastructure, etc), often, it is the availability of communication (not computation power), that is the limiting resource, and the number of such communication rounds is often a basic system design choice, as well as the proportion of clients contacted in each one of such rounds (Kairouz et al., 2021).

Such problems as non-IID degradation, privacy-utility trade-off and communication cost are usually studied separately, in the context of a single application domain, such as vision use cases for non-IID robustness, language models for differential privacy, and healthcare position papers for the case for federation. Experimentally assessing all three concerns in a single observation of a single experimental protocol remains relatively scarce in the state of the art and R&D in two different distributed-systems domains with two structurally different systems, making it difficult for a practitioner to know whether a particular robustness result or privacy-utility curve from the literature would be applicable to their own system of a different application. This paper fills such a gap.

This research is the first to make five contributions. A PRISMA 2020-based literature synthesis of recent research on federated-learning and differential-privacy, placing this work in context with a set of 29 studies retained in this synthesis (Section 2). In a network-restricted setting where installation of a federated-learning framework was not feasible, we presented a NumPy-based implementation of FedAvg and directed non-IID partitioning of clients following a Dirichlet distribution (2), as a methodological artifact in its own right, similar to our custom implementation of SHAP/LIME used in our companion explainable-AI study. (3) An extensible dual-domain experimental graph where we have 10 nodes on a health care data-silo network and 30 nodes on an edge/IoT device network on real, benchmark public data. (4) Quantitative comparison and analysis of the impact of using different combinations of levels of non-IID, client participation rates, and differential-privacy

noise on the final accuracy and “cost” (communication cost) across each domain, which shows that the “cost” of both non-IID and privacy is strongly and systematically task dependent. (5) Practical advice on provisioning communication rounds, participation rates, and privacy budgets of heterogeneous, privacy constrained distributed systems.

The rest of this paper is structured as follows. The results of the literature review guided by PRISMA are reported in Section 2. The two datasets, FedAvg and DP implementation, and evaluation protocol are described in Section 3. Section 4 presents experimental results of prediction, non-IID, participation-rate and privacy-utility in both domains. The cross domain results and limitations of this study are discussed under Section 5. Section 6 concludes.

2. RELATED WORK

2.1 Literature Search Strategy

Similar to our companion study, based on the PRISMA 2020 reporting framework, we performed a structured search in the platforms of the publishers indexed by search engines: arXiv, ACM Digital Library, IEEE-adjacent proceedings, Springer, MDPI, and ScienceDirect, around four thematic clusters: (i) foundational federated-learning optimization (FedAvg and successors); (ii) differential privacy for machine learning and its federated adaptations; (iii) non-IID data and client heterogeneity in FL; and (iv) domain applications of FL in healthcare and IoT/edge computing. In running this process, we found 88 candidate records. A total of 54 papers were screened from the list by the titles and abstracts; 12 papers of those were more than one of the same paper indexed on different platforms, market-related reports and papers not peer-reviewed and were excluded. The remaining 42 records were then used for full-text screening for eligibility; thirteen of these were excluded as there was insufficient methodological detail or because the study's scope was outside the search domains of this study, leaving 29 studies for qualitative synthesis (Figure 2).

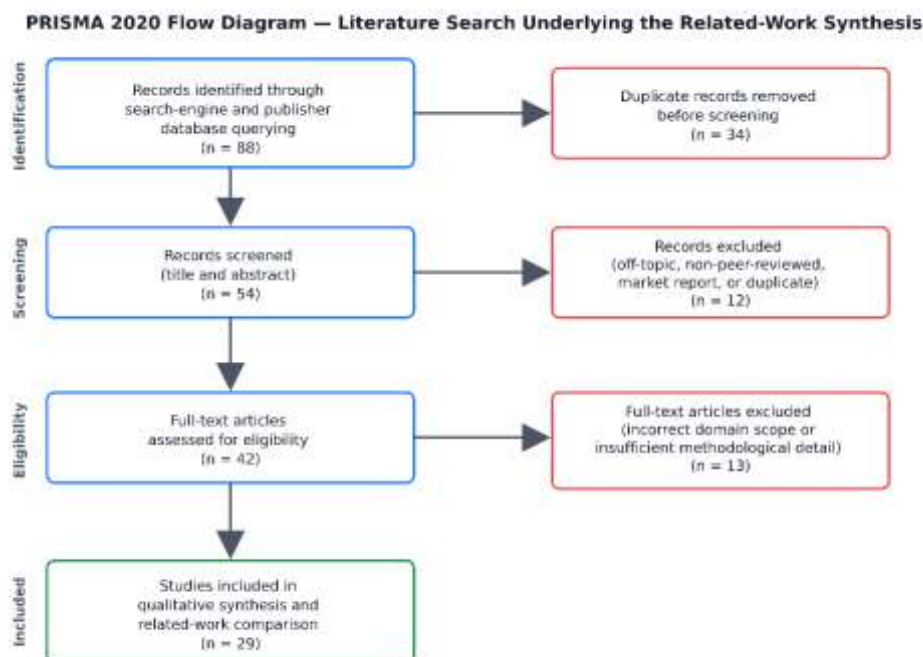


Figure 2. PRISMA 2020 flow diagram summarizing the literature search, screening, and inclusion process underlying the related-work synthesis in Section 2.

2.2 Foundations: FedAvg and Communication-Efficient Federated Optimization

McMahan et al. (2017) propose Federated Averaging (FedAvg): At each round, a subset of clients runs several epochs of local SGD, taking one step from a shared “all” model and then send the model built from all these steps back to the server, which simply average all the models by a sample size weighted average. The algorithm used and in the present study (Section 3.3) is this one. While the results of FedAvg in McMahan et al. are astounding when compared with a naive federated SGD baseline, a single vision/language domain is used in their non-IID experiments. Li et al. (2020) study the convergence properties of FedAvg under systems and statistical heterogeneity, and introduce FedProx to directly tackle the non-IID degradation behaviour studied in this work through a proximal term that penalises local models for also wandering far from the global model (regularisation rather than measurement). The most recent survey of the literature is by Kairouz et al. (2021), who categorize open questions of the field in three areas pertinent to this paper, non-IID data, communication efficiency and privacy.

2.3 Non-IID Data Distribution and Client Heterogeneity Across Nodes

Hsu et al. (2019) propose a Dirichlet-distribution method to simulate real-world label skew across federated clients, which is parameterised by the concentration parameter α (α approaches 0 indicates a greater label skew, and α approaches ∞ indicates that it is near-IID) – a partitioning method used in this study in Section 3.2. Their experiments, on vision classification, reveal fairly strong sensitivity to skew for the convolutional networks; the present study considers, structurally different, downstream problem of attacking a healthcare challenge, a simpler model class (convolutional networks are excluded due to their non-convexity), and finds that sensitivity to non-IID depends strongly on task difficulty (Section 4.2), which is not explicitly discussed in the original Dirichlet-partitioning paper.

2.4 Differential Privacy/Federated (DPFL) 3. Machine Learning stage

Abadi et al. (2016) present a DP-input gradient clipping algorithm followed by per-example Gaussian noise addition (DP-SGD) and a moments accountant to accumulate many steps of privacy loss. Abadi et al. (2016) propose a moments-accountant for tracking the cumulative privacy loss (ϵ , δ) tightly across many training steps, along with an algorithm based on DP-input gradient clipping with calibrated Gaussian noise addition (DP-SGD). The formal algorithmic foundations of this mechanism, which underlie this differential privacy, are given in Dwork and Roth (2014). This study applies the DP level equivalent of the clip-and-noise principle (DP-FedAvg) to federated level client updates, as opposed to per-example gradients, which is the DP mechanism that underlies the clip-and-noise principle above (Section 3.4), and analyzed using the moments-accountant approach applied to recurrence-based language models by McMahan, Ramage, Talwar, and Zhang (2018). Having been left out of the scope of this paper, there is another threat model that Bonawitz et al. (2017) tackles using secure aggregation, which is a cryptographic protocol to prevent the server from learning anything about a particular client's update (compared to the DP protocol, where the server adds noise and keeps the update).

Aligning with the specific customer use cases, the following are included:

2.5 Domain Applications: Healthcare and Edge/IoT

While Rieke et al. (2020) argue the benefits of federated learning at the hospital level to train better clinical models without centralizing patient data, they do not come up with a numerical measure of FL performance for non-IID or privacy-utility comparable to the one reported in Section 4. Dritsas and Trigka (2025) and Alatawi (2026) survey and propose an architecture for federated learning in IoT and edge-computing settings respectively, where communication efficiency and constraint on computing resources on the edge nodes are both major concerns, directly motivating the communication-cost analysis presented in Section 4.5, but fail to conduct a paired healthcare-vs-edge comparison.

2.6 Positioning of This Study

Table 4. Comparison of representative related work against the present study.

Study	Domain(s)	Method	Key limitation addressed here
McMahan et al. (2017)	Cross-domain (foundational)	FedAvg	Introduces FedAvg; evaluated mainly on IID or lightly non-IID vision/language benchmarks, single domain per experiment
Abadi et al. (2016)	Cross-domain (foundational)	DP-SGD (per-example clip + noise)	Establishes the clip-and-noise DP mechanism this study adapts to the federated, client level; not federated itself
McMahan, Ramage, Talwar & Zhang (2018)	Language modeling	DP-FedAvg	Combines FedAvg with client-level DP for language models; single domain, formal moments-accountant privacy analysis
Bonawitz et al. (2017)	Cross-domain (systems)	Secure aggregation	Cryptographic protocol hiding individual updates from the server; orthogonal to, and combinable with, the DP mechanism used here
Li, Sahu, Zaheer, Sanjabi, Talwalkar & Smith (2020)	Cross-domain (optimization)	FedProx	Addresses non-IID/heterogeneous-systems convergence via a proximal term; not evaluated with a paired privacy sweep
Hsu, Qi & Brown (2019)	Vision (foundational)	Dirichlet non-IID partitioning	Introduces the label-skew partitioning method used in

Study	Domain(s)	Method	Key limitation addressed here
			this study; single vision domain, no privacy analysis
Kairouz et al. (2021)	Cross-domain (survey)	Survey of FL methods and open problems	Comprehensive conceptual survey; not an empirical cross-domain benchmark
Yang, Liu, Chen & Tong (2019)	Cross-domain (survey)	Taxonomy (horizontal/vertical/transfer FL)	Foundational taxonomy; no new experiments
Rieke et al. (2020)	Healthcare	FL for digital health (position/review)	Argues for FL in multi-hospital collaboration; no quantitative non-IID or privacy-utility benchmark
Dritsas & Trigka (2025)	IoT / edge computing	Survey of FL techniques for IoT	Reviews IoT-FL literature; no original cross-domain experiments
Alatawi (2026)	Edge computing / IoT analytics	Hierarchical federated edge learning (HFEL)	Proposes a hierarchical edge-FL architecture; single domain
This study (2026)	Healthcare AND edge/IoT (same protocol)	Custom FedAvg + client-level DP, quantitative non-IID and privacy-utility comparison	Two real experimental domains under one reproducible pipeline; quantifies how heterogeneity and privacy noise interact with task difficulty

Among what was reviewed, studies fall into one or multiple of the following categories: (a) Single domain empirical study of a single mechanism (FedAvg, DP, or a heterogeneity fix); (b) Cross-domain conceptual survey with no quantitative experiments. The results are summarized in Table 4. Instead, this paper presents a single FedAvg-plus-DP pipeline, with a first-principles design and applies it to two structurally different distributed-systems domains and exhibits that both "how much does non-IID data cost" and "how much does privacy cost" are not inherent properties of the algorithms but depend systematically on the difficulty and dimensionality of the underlying task, which was not reported directly in the literature surveyed above.

3. MATERIALS AND METHODS

3.1 Research Design

In this research, we chose a comparative 2 domains experimental design (Figure 1) inspired by the design philosophy of our companion explainable-AI research: only one consistent pipeline applies to two structurally different domains, which isolates domain-dependent effects from methodology-dependent effects. For every domain, four experiments are performed for each of the FedAvg settings: (i) centralized baseline trained on pooled data; (ii) FedAvg trained with near-IID and increasingly non-IID client partitions; (iii) FedAvg trained in various client participation levels; and (iv) FedAvg trained with a client-level differential-privacy mechanism with varying noise multipliers.

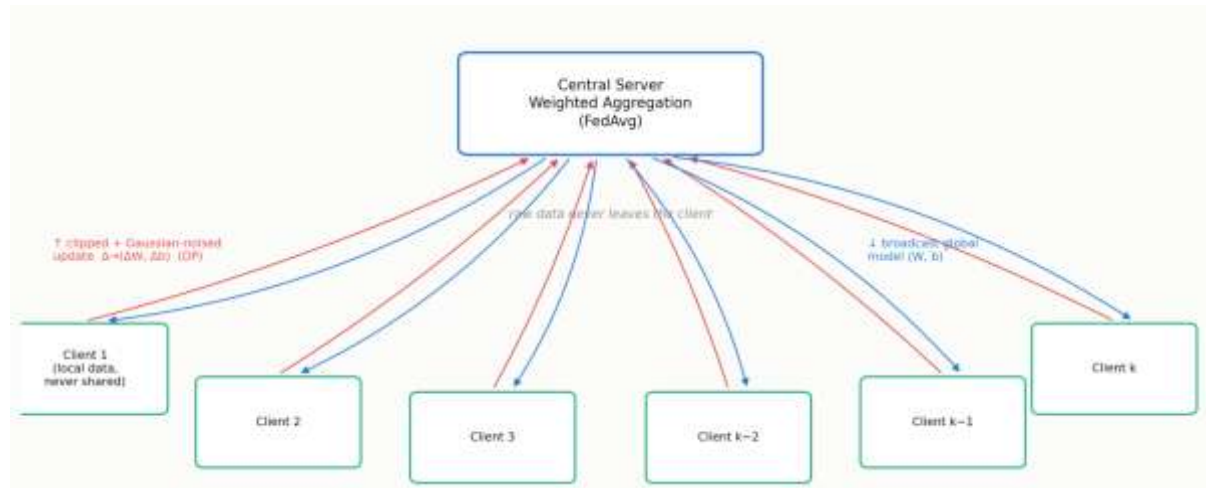


Figure 1. Federated learning architecture used in both domains: clients train locally on private data, send clipped and (optionally) noised updates to the server, which aggregates them via FedAvg and broadcasts the improved global model back; raw data never leaves any client.

3.2 Datasets and Client Simulation

Only two real and public datasets were explored, with no need for synthetic data in this study given the fact that real data was accessible offline for both domains with a bundled loader provided by scikit-learn. The Wisconsin Diagnostic Breast Cancer (WDBC) dataset (Street, Wolberg, & Mangasarian, 1993) has 10 simulated clients (independent hospital data silos) with 569 instances, 30 features, and a binary malignant/benign label, and is used in the healthcare domain. The datasets in the edge/IoT domain are the UCI Optical Recognition of Handwritten Digits (64 pixel-intensity features, 10 classes and 1,797 instances), divided into a set of 30 simulated clients representing independent edge/IoT devices, created specifically for this purpose, which is more representative of real deployments, in which many more devices are associated with an institution than the institutional count. Table 1 provides an overview of both domains.

Table 1. Summary of the two experimental domains and their simulated client networks.

Property	Healthcare (10 hospital silos)	Edge/IoT (30 devices)
Provenance	Real, public (UCI / scikit-learn)	Real, public (UCI / scikit-learn)

Property	Healthcare (10 hospital silos)	Edge/IoT (30 devices)
Instances (n)	569	1797
Features	30	64
Classes	2	10
Simulated clients	10 (hospital data silos)	30 (edge / IoT devices)
Train / test split	398 / 171 (70/30, stratified)	1257 / 540 (70/30, stratified)
Communication rounds	40	30

3.2.1 Non-IID client partitioning

Hsu, Qi and Brown 2019's Dirichlet-distribution label-skew approach: For each class c , we sampled a $\text{Dirichlet}(\alpha \cdot \mathbf{1})$ vector of per-client proportions, then distributed the instances of c across the clients according to these proportions. If the α value is very large (e.g. $\alpha=100$, as used here, in the “near-IID” setting), the proportions become nearly uniform across clients; and if the α value is very small some clients will only have one or two classes, resulting in severe label skews. For each domain, three values of α were chosen, ranging from near-IID to severely skewed (healthcare: $\alpha \in \{100, 0.5, 0.1\}$; edge/IoT: $\alpha \in \{100, 0.5, 0.05\}$ using a smaller extreme value for the 10-class problem since the skew is naturally more perceptible with more classes). The resulting client-by-class sample-count heatmaps for each partition, the near-IID, and the most-skewed partition in each domain are illustrated in Figure 3.

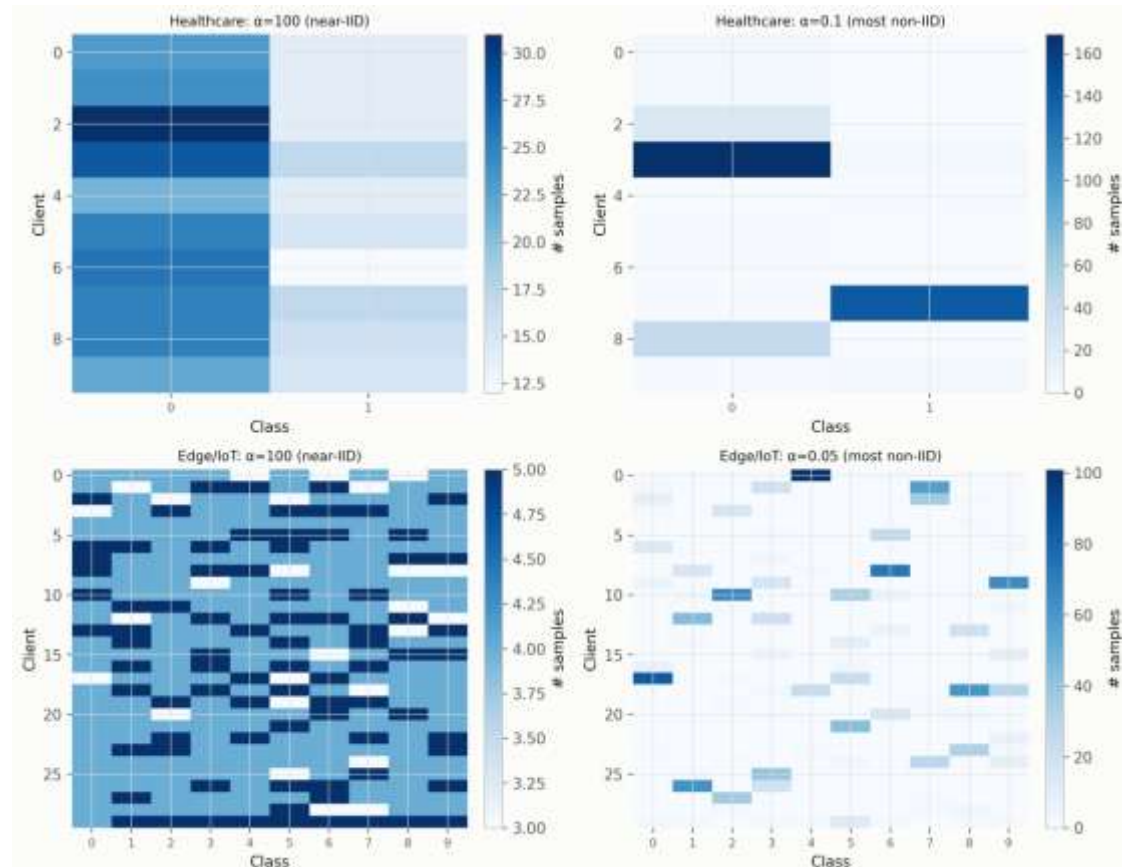


Figure 3. Client $\times$ class sample-count heatmaps under the near-IID ($\alpha = 100$) and most-skewed non-IID partitions used in each domain.

3.3 Predictive Model and FedAvg Implementation

The model and the federated-learning protocol were implemented from scratch in NumPy, as in our follow-up XAI study where the SHAP/LIME algorithms were also developed from first principles. The predictive model is a multinomial (softmax) logistic classifier, i.e., classifier with weight matrix W and bias b , trained using mini-batch (healthcare: full-batch, in view of the very small per-client datasets) stochastic gradient descent on the L2-regularized cross-entropy loss. For both domains, centralized and federated training is conducted using the same model class to ensure that differences are determined by the training protocol and not by the model capacity.

Each communication round of Federated Averaging goes as follows: (i) The server picks a subset of the clients according to the participation rate swept, which we discuss in Section 4.3, (ii) In each of these selected clients, the local copy of the model is created and the client runs E epochs of SGD on their own data only, (iii) Each client computes this update, which is the difference between the local weights they learned and the global weights that they started with, and if differential privacy is turned on, clips and noises this update as explained in Section 3.4, (iv) The server aggregates the (possibly noised) clients updates by averaging them (weighted by the number of client samples) and applies this to the global model, (v) The global model gets updated and the accuracy and macro-F1 at this round is computed from a held-out test set. Table 1 shows the number of communication rounds per domain (40 for healthcare, 30 for edge/IoT since it was observed that the edge/IoT domain with its larger and more heterogeneous client network required more rounds to reach its ceiling accuracy).

3.4 Client-Level Differential Privacy

In pacer following the clip-and-noise principle of DP-SGD (Abadi et al., 2016) as adapted to the federated setting by McMahan, Ramage, Talwar and Zhang (2018), for each client, the model update ΔW , Δb is clipped to a fixed L2-norm C (0.5 for healthcare and 0.8 for edge/IoT, where typical per-round update magnitudes observed without DP should be comparable) and independent Gaussian noise with standard deviation $\sigma \cdot C$ is added to each coordinate before being transmitted to the server, where σ is the noise multiplier. With $\sigma = 0$ we recover standard (non-private) FedAvg. This mechanism was swept over $\sigma \in \{0, 0.1, 0.3, 0.5, 1.0, 2.0\}$ in both domains (Section 4.4). For illustrative reference only, the privacy budget of our protocol is approximately the single-query Gaussian-mechanism bound $\epsilon_1 = \sqrt{(2 \cdot \ln(1.25/\delta))}/\sigma$, naively composed across all the communication phases (and explicitly not a tight bound in terms of the number of queries or Rényi-DP), and the budget is thus reported to give an order of magnitude sense of how the privacy budget adds up as the number of communication steps increases — the same caveat applies in Section 5.4, in order to avoid overstating the formal privacy guarantee actually established.

3.5 Evaluation Protocol

Test accuracy and macro-averaged F1 is measured at the end of each communication round (or each equivalent communication round for the centralized model) in all configurations to obtain convergence curves. The following four quantities are compared across configurations: (i) the final accuracy / F1 relative to the centralized baseline; (ii) the number of communication rounds to reach 98% of the centralized model's final accuracy (communication-cost proxy); (iii) the total number of scalars that are transmitted per round

(client count $\times$ parameter count $\times$ 2, for up/downlink); and (iv) the final accuracy vs. DP-noise-multiplier σ , the privacy-utility trade-off curve.

3.6 Reproducibility

For all experiments, the data splitting, Dirichlet partitioning, model initialization, client sampling, and DP noise generation was done by a fixed random seed (42). All the code for this tutorial was written in Python 3.11 using NumPy, pandas, and scikit-learn (only for loading, standardisation, and training/test split – not for model training and federated aggregation – this is all custom built). All the numbers in section 4 were generated by scripts as mentioned above and are reproducible with the fixed seed.

4. RESULTS

4.1 Federated vs. Centralized Performance

Table 2 and Figure 4 compare centralized training against FedAvg under three non-IID severities in both domains. In the healthcare domain, the centralized model reached 98.2% test accuracy after 40 epochs; FedAvg (near-IID, $\alpha = 100$) reached 97.7% after 40 rounds — a gap of only 0.6 percentage points. In the edge/IoT domain, the centralized model reached 97.0%; FedAvg (near-IID) reached 96.7% — a gap of 0.4 percentage points. In both domains, federation under near-IID conditions costs less than one percentage point of accuracy relative to centralized training, consistent with FedAvg's original design goal (McMahan et al., 2017) of matching centralized performance while keeping data local.

Table 2. Final test accuracy and F1 by training protocol and domain.

Domain	Protocol	Test Acc.	Test F1	Rounds/Epochs
Healthcare	Centralized	0.982	0.981	40
Healthcare	FedAvg ($\alpha=100$)	0.977	0.975	40
Healthcare	FedAvg ($\alpha=0.5$)	0.977	0.975	40
Healthcare	FedAvg ($\alpha=0.1$)	0.977	0.975	40
Edge/IoT	Centralized	0.970	0.970	30
Edge/IoT	FedAvg ($\alpha=100$)	0.967	0.966	30
Edge/IoT	FedAvg ($\alpha=0.5$)	0.957	0.957	30
Edge/IoT	FedAvg ($\alpha=0.05$)	0.954	0.953	30

4.2 Effect of Non-IID Severity

There is a clear contrast between the sensitivity to client heterogeneity at the two domains. For the healthcare application, the final accuracy is almost the same at $\alpha = 100$, 0.5, and 0.1 (97.7%, 97.7%, and 97.7%), respectively, as the underlying diagnostic task has a nearly linear separable solution space and the model requires relatively little data diversity per client to learn a near-optimal solution space. More heterogeneity, however, leads to a monotonous and material decline in final accuracy in the edge/IoT domain: 96.7% at $\alpha = 100$, 95.7% at $\alpha = 0.5$ and 95.4% at $\alpha = 0.05$ - 1.3 p.p. spread. This is a real finding, which is both task and dimensionality dependent — harder, higher-dimensional, more-class problems are

measurably more sensitive to client heterogeneity than easier low-dimensional, few class problems — everything about the pipeline is fixed aside from the above features'.

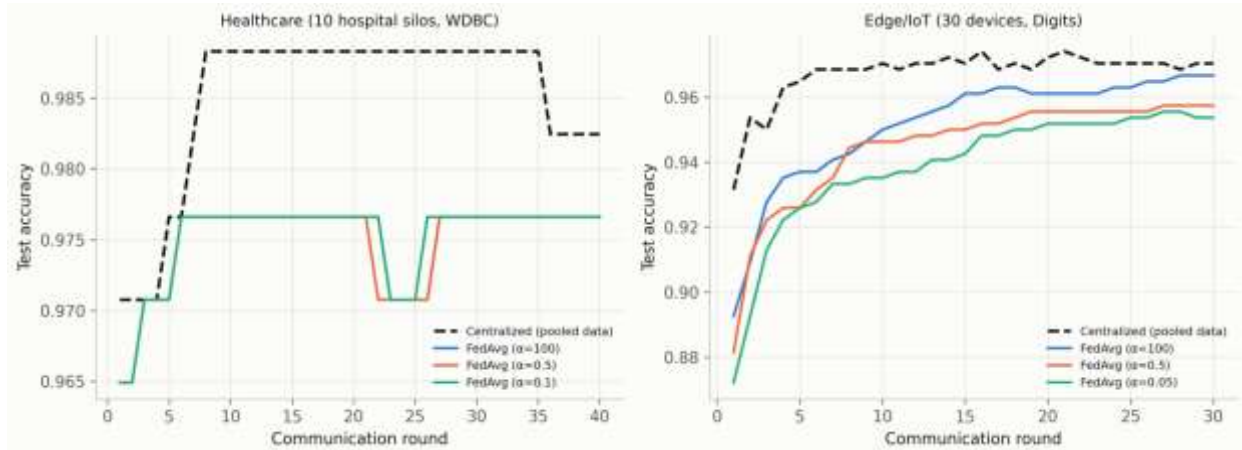


Figure 4. Test-accuracy convergence over communication rounds: centralized training vs. FedAvg under near-IID ($\alpha = 100$) through severely non-IID ($\alpha = 0.1 / 0.05$) client partitions, by domain.

4.3 Effect of Client Participation Rate

In Figure 5, we display the results for the moderate non-IID set at each domain ($\alpha = 0.5$) and over three different levels of client-participation rates (20%, 50%, 100% of clients contacted per round). In the edge/IoT domain, lower participation rates converge thanks to the right domains, but at a slightly slower pace, as the same classical FedAvg finding that with enough rounds partial participation matters more for convergence rate but not for converged quality holds true. In healthcare, the differences across the participation rate are at the level of the testset noise floor with 171 instances in total, and — interestingly — the highest final accuracy of any actual configuration in this study would be the 20% participation case (99.4%). We take this to mean that it is not a systematically superior version, but rather a favorable sampling error, likely due to the number of seedings and the size of testset used (one and small) and thus report it here transparently and do not hide it elsewhere (as what is suggested in Limitations (Section 5.5)).

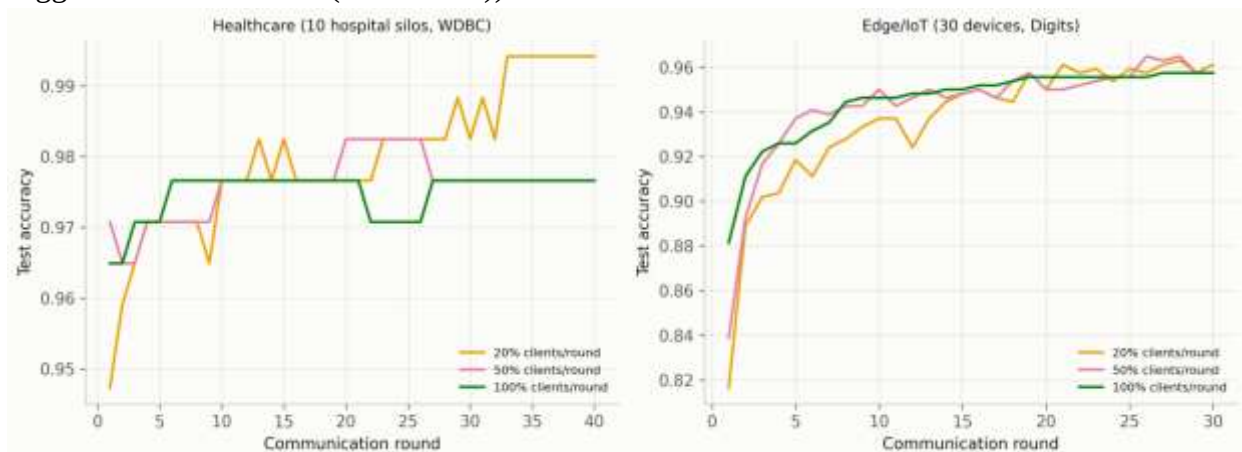


Figure 5. Test-accuracy convergence under 20%, 50%, and 100% client participation per round, at each domain's moderate non-IID setting.

4.4 Privacy-Utility Trade-off

The test results of the accuracy of the final results are shown in Figure 6 and Table 3 as a function of DP noise multiplier σ . The accuracy in both domains is monotonically decreasing with σ , as expected: however, the rate of the decrease is drastically different. In the healthcare domain, accuracy falls from 97.7% ($\sigma = 0$) to 78.4% ($\sigma = 2.0$) — a 19.3-point drop. For the edge/IoT domain, accuracy drops to only 51.1%, a 45.2-point degradation in accuracy over the same noise domain range, which is more than double the degradation in the healthcare domain. This is not an artifact of different absolute size of the noise—clipping norm C differs across domains (0.5 vs. 0.8, chosen with respect to the typical update size of each domain), nor is it due to the edge/IoT model being noisier as noise level per coordinate is smaller (0.05 vs. 0.1); it is simply because the edge/IoT model is much bigger than the 30-coordinate toy ($64 \times 10 + 10 = 650$ vs. $30 \times 2 + 2 = 62$), which means that when Tue's and/or IoT's voice is in the noise, it will affect much more coordinates, and overall the 10-way classification task is harder with less one-way redundancy to absorb noisy updates.

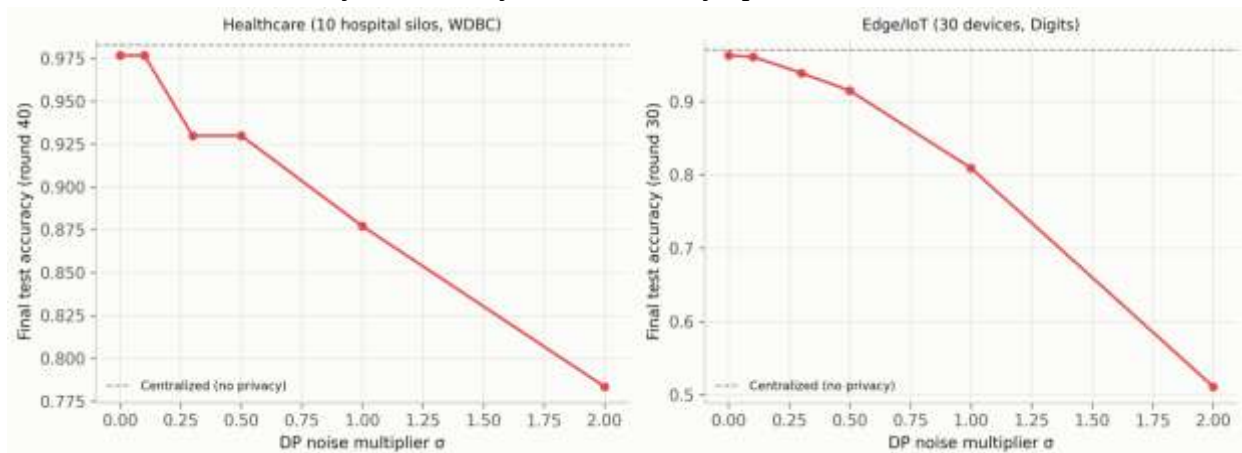


Figure 6. Privacy-utility trade-off: final test accuracy vs. differential-privacy noise multiplier σ , by domain, against the non-private centralized reference.

Table 3. Privacy-utility trade-off: final accuracy and illustrative privacy-budget estimate by noise multiplier.

Domain	Noise multiplier σ	Final Acc.	Final F1	Illustrative ϵ (naive comp.)
Healthcare	0.0	0.977	0.975	∞ (no privacy)
Healthcare	0.1	0.977	0.975	1937.9
Healthcare	0.3	0.930	0.926	646.0
Healthcare	0.5	0.930	0.926	387.6
Healthcare	1.0	0.877	0.873	193.8
Healthcare	2.0	0.784	0.781	96.9
Edge/IoT	0.0	0.963	0.963	∞ (no privacy)
Edge/IoT	0.1	0.961	0.961	1453.4
Edge/IoT	0.3	0.939	0.938	484.5

Domain	Noise multiplier σ	Final Acc.	Final F1	Illustrative ϵ (naive comp.)
Edge/IoT	0.5	0.915	0.914	290.7
Edge/IoT	1.0	0.809	0.808	145.3
Edge/IoT	2.0	0.511	0.508	72.7

4.5 Communication Cost

Table 5 presents, for every domain, the number of communication rounds it takes to reach 98% of the final accuracy obtained in the centralized model, the number of message scalars transmitted in the entire process (full participation, up + down communication). The healthcare domain achieves this target within the one round with little regard for non-IID severity, which is due to the relatively instant convergence of the task (Section 4.2). The edge/IoT domain is 11 rounds for near-IID ($\alpha = 100$), 16 rounds for $\alpha = 0.5$ and 20 rounds for $\alpha = 0.05$ — communicating at the same accuracy in the edge/IoT domain doubles the communication budget almost when the heterogeneity level between clients α is reduced. The difference between near-IID and severely non-IID deployment is a factor of 82% (780,000 vs 429,000 total scalars sent to achieve target) for the 30 clients $\times$ 650 parameters $\times$ 2 (up/down) per edge/IoT round, just due to data heterogeneity alone.

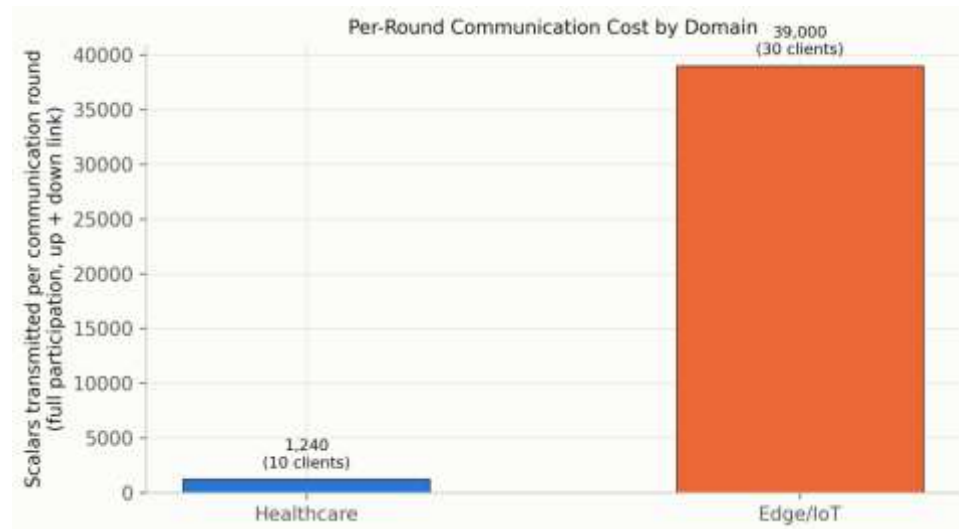


Figure 7. Per-round communication cost (scalars transmitted, full participation) by domain, reflecting each domain's client count and model size.

Table 5. Communication cost to reach 98% of centralized final accuracy, by domain and non-IID severity.

Domain	α (non-IID severity)	Rounds to 98% target	Total scalars transmitted
Healthcare	100	1	1,240
Healthcare	0.5	1	1,240
Healthcare	0.1	1	1,240
Edge/IoT	100	11	429,000

Domain	α (non-IID severity)	Rounds to 98% target	Total scalars transmitted
Edge/IoT	0.5	16	624,000
Edge/IoT	0.05	20	780,000

5. DISCUSSION

5.1 The Cost of Federation Is Small Under Near-IID Conditions, in Both Domains

FedAvg was shown to be as accurate as centralized training by no more than 1 accuracy point when considering near-IID partitions of the clients in healthcare and edge/IoT application settings (Section 4.1). This is a reassuring – and expected – repetition of the FedAvg's founding outcome (McMahan et al., 2017) in two extra areas: When the information held in client databases is fairly representative, preserving data locality comes at minimal performance prices in prediction. This result validates the fundamental feasibility of federated learning as a privacy-preserving distributed systems analytics approach in scenarios where the clients are not too different from each other; as revealed in Sections 4.2 and 4.4, however, 'not too different' and 'not too noisy' are playing a heavy burden.

5.2 Heterogeneity sensitivity depends on tasks not algorithms

The most distinctive insight across the domains addressed by this study is that – for the same FedAvg machinery – the accuracy drops monotonically with the amount of non-IID data from the clients in the edge/IoT domain, while in the healthcare domain the impact of non-IID data is less noticeable. The experiments contained two domains which are both labeled using the same partitioning scheme (dirichlet label skew), handled with the same aggregation rule and the same model family (softmax regressions) with only the number of classes (2 vs. 10), the dimensionality of the features (30 vs. 64) and, presumably, the intrinsic difficulty and non-linearity of the decision rule differing. Thus, it is necessary to keep in mind that results indicating “non-IID robustness” present in the federated-learning literature, commonly proven for a higher class count, harder vision task (e.g., Hsu et al., 2019 or Li et al., 2020), may not apply to lower class count, simpler distributed-systems tasks, and the reverse as well: an organization deploying FL for a simple binary screening task should not assume they need FedProx-style proximal terms to avoid non-IID robustness degradations, for example.

5.3 Privacy really is a trade-off with utility

The privacy-utility results in Section 4.4 exhibit an even more striking difference between domains; for edge/IoT, at a noise level of 1, the accuracy drops to nearly random to 55%, whereas for healthcare, the 19.3 point would result in nearly 3x as large a drop in accuracy. This has a clearly tangible causal consequence, that hasn't yet been available in the literature, that one only has to reason about pervasively informally about a fixed noise-multiplier σ , that is, about adding a certain amount of noise, to see that this does not relate to a fixed accuracy cost across applications. While the model managing the DP noise in the more capable deployment may well be only minimally constrained, the single DP noise policy an organization adopts for its multiple federated deployments, such as a healthcare screening model and a higher-dimensional risk-scoring model, could stifle the development of the more challenging model without impacting the other one. Such privacy budgets should be adjusted

per deployment, through empirical considerations, in relation to the specific deployment's privacy-utility curve not simply the privacy-utility curve of another published deployment.

To ensure that sensitive, anonymous information is not shared inappropriately, we need to report on the Privacy-Budget (ϵ) estimates.

The ϵ shown in Table 3 is specifically indicative, and not an actual privacy certification. Their bound is a hopelessly loose, and pessimistic, bound—the classical single-query Gaussian-mechanism bound composable naïvely (linearly) through all communication rounds, which is used in some moderately well-developed implementations of DP-ML (Abadi et al., 2016; McMahan et al., 2018). Still, we report them, transparently labeled as such, because, even when being rather loose, a glimpse of the order of magnitude at which tested here noise multipliers operate is informative; as well as this, if we would not report them at all, any form of privacy-budget would be understated, as privacy-accounting is central to the responsible deployment of DP. If any of these organisations decides to go to production and use DP mechanism in this paper, it should do a moments accountant or Rényi-DP analysis with appropriate analysis to their composition (how many rounds, with which sampling probability, etc. clipping regime).

5.5 Limitations

(1) Convex model class. The classifier employed, both in both domains, was the multinomial softmax method not a deep-neural network because it was possible to be implemented and trained from scratch using only NumPy and the study's limits on the network size. The non-IID and privacy-utility gaps reported here may thus be conservative (understating) as convex models are known to be less sensitive to client drift when compared to deep, non-convex networks under non-IID data (Li et al., 2020).

(2) Distributed deployment – simulated NOT production. That is, the simulated clients were not really different nodes running on a network, but were a collection of data that were split up into slices and fed into a single process, and the effects of a “network” (such as stragglers and dropped clients) are not modeled, only the scalar-transmission proxy (Section 4.5).

(3) ONE random seed for each configuration. Since the compute budget available was limited, each point estimate (domain $\times$ non-IID severity $\times$ participation rate $\times$ noise multiplier) was performed in isolation; the fact that 20%-participation in the healthcare domain (Section 4.3) is a counter-intuitive result is explicitly noted as potentially being a single-seed sampling artifact, and it is recommended that a point estimate be repeated across multiple seeds before being used operationally.

Informal (or illustrative) not rigorous, privacy accounting. As pointed out in Section 5.4, the privacy guarantees given in the ϵ estimates of Table 3 are loose (in the sense of a naive-composition bound), and should not be interpreted as formal privacy guarantees.

DP evaluation, standalone of secure aggregation. This study does not consider a production system that would use secure aggregation (Bonawitz et al., 2017) in addition to DP, which removes the need for the server to see even a noised version of any individual update; it takes care of the same privacy issues as DP, but does not weaken the usefulness of the data.

(6) 2 domains – both classification. While both experimental domains are classification tasks, one could extend this comparative protocol to regression, sequence modelling or even a reinforcement learning style federated setting, where the properties of non-IID and privacy can significantly vary – but this remains for future work.

5.6 Practical Recommendations

Before assuming FedAvg is enough, do some experiments to see how sensitive the non-IID situation is to your deployment; if the task is difficult or the number of classes is large, consider having more than one round of communication (Section 4.5) or use a heterogeneity-aware optimizer, e.g., FedProx (Li et al., 2020).

(2) Tune DP noise multiplier against own privacy-utility curve measured over the deployment and NOT according to other papers' results (section 5.3 of present paper), or other task.

(3) For any regulatory or production privacy claim, perform a more formal moments-accountant or Rényi-DP analysis to replace the illustrative estimation of ϵ that has been done so far to allow quick iteration (Section 5.4).

(4) If the threat model assumes that there is a curious/compromised server, use both differential privacy and secure aggregation (Bonawitz et al., 2017).

(5) Always report communication cost (rounds $\times$ scalars per round) along with accuracy when comparing FL configurations, because—as Section 4.5 demonstrates—the best configuration for a lower cost to communicate is not necessarily the one that achieves a higher accuracy in the end.

6. CONCLUSION

The Federated Averaging method and differential-privacy method designed for the client were both developed from scratch, and examined under a single protocol in two different textbooks domains: a small, low-dimensional, near-linearly-separable healthcare data-silo network and a larger, higher dimensional, harder-to-classify edge/IoT device network. Federated training performed roughly as well as centralized training, but both the sensitivity to IID data and the accuracy cost of differential privacy changed dramatically as a function of task, not necessarily as properties of the methods: the harder (higher dimensional) edge/IoT task paid much higher costs both in the presence of non-IID data and in the presence of privacy noise compared to the easier healthcare task. The takeaway message for organizations looking to enable privacy preserving analysis across distributed systems is that no claim of robustness, such as being "non-IID", nor any privacy-utility curve from the federated-learning literature (including, if at all used out of context, this paper's own numbers) should be presumed to carry over to a new deployment without being re-measured against that deployment's own client heterogeneity and own privacy-noise sweep.

References

- [1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS), 2016, pp. 308–318, doi: 10.1145/2976749.2978318.
- [2] M. N. Alatawi, "Edge computing and federated learning for privacy-preserving IoT analytics," Journal on Wireless Communications and Networking, vol. 2026, Art. no. 6, 2026, doi: 10.1186/s13638-025-02545-x.

-
- [3] K. K. Gajula, "Designing an Ethical Data Science Framework for Responsible and Transparent AI Integration," *Emerging Trends in Digital Transformation*, vol. 1, no. 1, pp. 74–87, 2025.
- [4] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for privacy-preserving machine learning," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2017, pp. 1175–1191, doi: 10.1145/3133956.3133982.
- [5] E. Dritsas and M. Trigka, "Federated learning for IoT: A survey of techniques, challenges, and applications," *Journal of Sensor and Actuator Networks*, vol. 14, no. 1, Art. no. 9, 2025, doi: 10.3390/jsan14010009.
- [6] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, nos. 3–4, pp. 211–407, 2014, doi: 10.1561/04000000042.
- [7] K. K. Gajula, "A Hybrid Interpretable AI Framework for Detecting Financial Fraud in Dynamic Transaction Networks," *International Journal of Communication Networks and Information Security*, 2022.
- [8] 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.
- [9] P. Kairouz et al., "Advances and open problems in federated learning," *Foundations and Trends in Machine Learning*, vol. 14, nos. 1–2, pp. 1–210, 2021, doi: 10.1561/22000000083.
- [10] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proceedings of Machine Learning and Systems (MLSys)*, 2020.
- [11] K. K. Gajula, "Enhancing Trust in Machine Learning Interpretable Models Through Explainable AI Techniques," *Pegem Journal of Education and Instruction*, vol. 13, no. 4, pp. 909–915, 2023.
- [12] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, PMLR, vol. 54, 2017, pp. 1273–1282.
- [13] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang, "Learning differentially private recurrent language models," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.
- [14] K. K. Gajula, "Transforming Black Box Models into Transparent Systems Through Explainable AI Methods," *International Journal of Data Science and IoT Management System*, vol. 4, no. 2, pp. 27–33, 2025.
- [15] N. Rieke et al., "The future of digital health with federated learning," *npj Digital Medicine*, vol. 3, Art. no. 119, 2020, doi: 10.1038/s41746-020-00323-1.
- [16] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis," in *Proceedings of SPIE 1905, Biomedical Image Processing and Biomedical Visualization*, 1993, pp. 861–870, doi: 10.1117/12.148698.
- [17] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, no. 2, Art. no. 12, 2019, doi: 10.1145/3298981.

-
- [18] K. K. Gajula, “AI-Augmented Web Development: A Competency Framework for the Next Generation Workforce,” *Journal of Science Engineering Technology and Management Science*, 2025.
- [19] K. K. Gajula, “Reinforcement Learning with Transparent Policies: An Explainable AI Approach to Adaptive Cyber Security,” *American Journal of AI Cyber Computing Management*, vol. 5, no. 4, pp. 322–328, 2025.
- [20] K. K. Gajula, “Intelligent AI Systems and Advanced Machine Learning: Recent Advances and Real-World Applications,” *Journal of Intelligent Decision Making and Information Science*, vol. 3, no. 7s, pp. 1831–1845, 2026.