

Keep Your Data, Share the Model: Federated Learning for Privacy-Preserving Hardware Anomaly Detection Across Distributed Fleets

Cynara Justine
Microsoft Corporation
cjustine@microsoft.com

Abstract—Modern semiconductor and datacenter hardware emit vast telemetry that is increasingly used to train AI models for anomaly detection, fault prediction, and yield analysis. Yet the most valuable training signal is fragmented across organizational and commercial boundaries—fabs, regions, and vendors each hold data that is sensitive, proprietary, or siloed—so centralizing it to train a single high-accuracy model is often legally or competitively infeasible. We present a federated learning (FL) approach to privacy-preserving hardware anomaly detection in which multiple sites collaboratively train a shared model while raw telemetry never leaves its owner: each site trains locally and transmits only model-weight updates to an aggregator that averages them into a global model. Our contribution is an open, reproducible benchmark and blueprint for FL in hardware reliability, together with a concrete finding—a fab that has never observed a given fault cannot detect it alone yet recovers that capability through federation. On a synthetic multivariate-telemetry benchmark with heterogeneous operating regimes, federation lifts detection ROC-AUC from 0.83 (local-only) to 0.99, matching the 1.00 centralized upper bound and recovering 94% of the gap. On the public UCI SECOM dataset partitioned across eight non-IID fabs, FedAvg improves minority-class F1 by 54% over local-only, and FedProx raises ROC-AUC from 0.54 to 0.62, near the 0.64 centralized bound—while exchanging only ~160 KB of parameters and adding optional client-level differential privacy. The results show that hardware-reliability collaboration is achievable without data sharing, lowering a key adoption barrier for cross-vendor cooperation in the semiconductor ecosystem.

Index Terms—federated learning, anomaly detection, semiconductor manufacturing, hardware reliability, privacy-preserving machine learning, differential privacy, non-IID data.

I. INTRODUCTION

Reliability is a first-class concern across the semiconductor ecosystem. Fabrication lines, test floors, and datacenter fleets continuously emit high-dimensional telemetry—process-sensor traces, environmental readings, and health counters—that is increasingly consumed by machine-learning models for anomaly detection, fault prediction, and yield analysis [1], [2]. The accuracy of such models is fundamentally limited by the diversity of failure examples they observe during training. In practice, however, the data needed to cover the space of faults is fragmented: a single fab sees only the recipes it runs, a single datacenter region sees only its own workloads, and competing vendors have no mechanism to pool observations of rare defects.

The obvious remedy—centralizing all telemetry to train one strong model—is usually infeasible. Process data encodes proprietary recipes and yield economics; fleet telemetry can expose customer workloads; and cross-vendor sharing raises antitrust and confidentiality concerns. As a result, each site trains a weaker model on its own partial

view, and rare faults that one site has never seen remain undetectable there even though a sibling site may have abundant examples.

Federated learning (FL) [1] offers a principled escape from this dilemma. Instead of moving data to a central model, FL moves the model to the data: each participating site trains locally and transmits only model-weight updates to an aggregator, which averages them into a shared global model that is redistributed for the next round. Raw telemetry never leaves its owner, so the privacy and confidentiality boundary is preserved while the collective still benefits from every site's observations.

This paper asks a focused, practical question: does federation meaningfully improve hardware anomaly detection under the data-fragmentation conditions that actually occur in the semiconductor ecosystem, and by how much? We answer it with an open, fully reproducible benchmark built entirely on non-confidential data. Our contributions are:

- 1) A reproducible FL benchmark for hardware reliability spanning two complementary detectors—an unsupervised autoencoder on multivariate telemetry with heterogeneous operating regimes, and a supervised yield classifier on the public UCI SECOM dataset [3].
- 2) A quantitative characterization of the collaboration value: we show that a fab which has never observed a fault cannot detect it in isolation, yet recovers that capability through federation—turning a previously qualitative argument into a measured effect.
- 3) An honest evaluation of communication cost, non-IID robustness (FedAvg vs. FedProx [4]), a fault-aware aggregation ablation, and the utility cost of client-level differential privacy [5], with all code and data released for independent reproduction.

II. RELATED WORK

A. Federated learning and heterogeneity

Federated averaging (FedAvg) [1] established communication-efficient training over decentralized data and remains the canonical aggregation rule. Its accuracy degrades when client data is non-identically distributed (non-IID) [6], [7], a regime that is the norm rather than the exception for fabs running different recipes. FedProx [4] addresses this by adding a proximal term that penalizes local drift from the global model, improving stability under

heterogeneity. Secure aggregation [8] and differential privacy [5], [9] further harden the protocol so that even the shared updates leak little about individual records. Our work applies these building blocks rather than proposing a new optimizer, and measures what they buy in a hardware-reliability setting.

B. Anomaly detection for hardware telemetry

Reconstruction-based detectors—autoencoders and their probabilistic variants [10], [11]—are a standard

unsupervised approach for multivariate telemetry: a model trained on normal data flags samples it cannot reconstruct. Supervised yield classification on semiconductor process data, exemplified by the SECOM dataset [3], is the complementary labeled setting. Both are widely studied in the centralized regime; what has been largely argued only qualitatively is whether these detectors can be trained collaboratively across owners who will not share raw data. We close that gap with direct local-vs-federated-vs-centralized measurements on both detector families.

Federated learning: collaborate without sharing telemetry

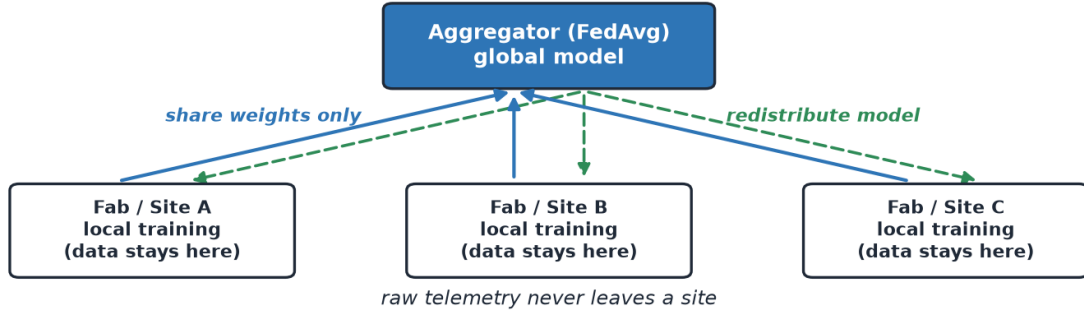


Fig. 1. Federated learning architecture: each site trains locally and shares only model weights; raw telemetry never leaves the site.

III. PROBLEM SETTING AND THREAT MODEL

We consider K sites (fabs, regions, or fleets), each holding a private dataset that it will not disclose. The sites share a common model architecture and wish to train a single anomaly detector that performs well on the union of their failure modes. A semi-honest aggregator coordinates rounds: it receives weight updates, averages them, and returns the global model. The privacy requirement is that raw telemetry never leaves a site; only model parameters (or clipped, noised parameter deltas under differential privacy) are exchanged.

Two properties make this setting hard. First, data is non-IID: the prevalence and even the presence of specific fault classes varies sharply across sites, so some sites hold almost no positive examples. Second, the minority (fault) class is rare in absolute terms—on SECOM roughly 6.6% of wafers fail—so detectors must extract signal from few positives without overfitting.

For measurement only, we additionally train a centralized model on pooled data. This is deliberately not a deployable configuration—pooling raw telemetry is exactly what confidentiality forbids—but it serves as a demanding performance ceiling: the accuracy attainable if privacy were a non-issue. Our objective is for federation, which never moves data, to approach this ceiling. Reporting it therefore strengthens rather than weakens the evaluation, because it

bounds how much (if any) accuracy is sacrificed to preserve privacy.

IV. METHOD

A. Federated averaging and FedProx

Each round r proceeds in three steps. (i) The aggregator broadcasts the current global weights θ to all sites. (ii) Each site k performs E epochs of local training on its private data to obtain θ_k . Under FedProx [4], local training minimizes the task loss plus a proximal term $(\mu/2)\|\theta_k - \theta\|^2$ that discourages drift from the global model; FedAvg is the special case $\mu = 0$. (iii) The aggregator forms the new global model as a weighted average $\theta \leftarrow \sum_k a_k \theta_k$, with weights a_k proportional to each site's sample count. Only the vectors θ_k cross the network; the datasets never do.

Here r is the communication-round index; K the number of participating sites, indexed by k ; θ the shared global weight vector; θ_k the weight vector after site k 's local training; E the number of local training epochs per round; $\mu \geq 0$ the FedProx proximal coefficient ($\mu = 0$ recovers FedAvg); and a_k the aggregation weight of site k , set to $a_k = n_k / \sum_j n_j$, i.e., proportional to its local sample count n_k . The proximal norm $\|\cdot\|$ is the Euclidean (L2) distance between the local and global weight vectors.

B. Fault-information-weighted aggregation (FIWA)

Sample-proportional averaging under-weights small fabs that are nonetheless rich in rare faults. We evaluate a fault-information-weighted aggregation (FIWA) ablation that upweights each site by its count of minority-class examples (with a floor to retain contributions from fault-free sites). FIWA tests whether steering the average toward fault-bearing sites recovers additional minority-class signal on top of FedProx.

C. Client-level differential privacy

To provide a formal privacy guarantee beyond “no raw data leaves the site,” each site can clip its weight delta to an L2 norm bound and add Gaussian noise before transmission [5], [9]. This bounds the influence of any single record on the shared update at a measurable cost to utility, which we quantify below.

D. Detectors

We deliberately implement both detectors from scratch in NumPy so the federated weight exchange is fully explicit and auditable, with no framework hiding the privacy boundary. For multivariate telemetry we use a small dense autoencoder and score samples by reconstruction error. For SECOM we use an L2-regularized logistic-regression yield classifier with class weighting to counter imbalance; features are mean-imputed, standardized, and reduced to 50 principal components. Detection thresholds are set from a held-out reference at the expected anomaly prevalence.

V. EXPERIMENTAL SETUP

A. Datasets and partitioning

We use two non-confidential datasets (Table I). The first is a synthetic multivariate-telemetry benchmark: each of five sites runs a distinct normal operating regime—a different low-rank correlated sensor manifold with its own mean offset—mirroring different fab recipes. Anomalies are injected as spikes, level shifts, and variance bursts. This models the realistic failure mode in which an autoencoder trained on one site’s normal data raises false alarms on another site’s perfectly normal telemetry. The second is the public UCI SECOM dataset [3]: 590 process-sensor features over 1,567 wafers with a ~6.6% fail rate. We partition SECOM across eight simulated fabs using a Dirichlet split ($\alpha = 0.3$) over the minority class, so failures are unevenly distributed and several fabs see almost no faults—an explicitly non-IID, label-skewed regime.

All results average over multiple random seeds (12 for SECOM, 5 for telemetry), each reshuffling the partition and initialization. The synthetic generator is procedural and seeded; SECOM downloads directly from the UCI repository. No Microsoft-internal data is used.

Table I. Benchmark datasets and partitioning.

Property	Telemetry	SECOM
Sites / fabs	5	8 (non-IID, $\alpha=0.3$)
Features	25	590 $\rightarrow$ PCA-50
Samples	2,000	1,567 wafers
Anomaly rate	injected	6.6% fail
Detector	Autoencoder	Logistic reg.

B. Protocol and metrics

We compare three configurations on a common held-out test set—a subset of wafers set aside before training and never seen during it, so reported scores reflect generalization to new data: (a) local-only, K independent models each trained on its own partition with no collaboration; (b) federated, FedAvg/FedProx with only weights shared; and (c) centralized, a single model on pooled data—an idealized, non-private upper bound. SECOM runs use 25 rounds of 25 local epochs; the centralized and local baselines are matched on total training budget so that any federated advantage cannot be attributed to more compute. We report threshold-independent ROC-AUC together with minority-class (fault) F1, recall, and precision, since accuracy is uninformative when only 6.6% of samples belong to the fault class (its prevalence)—a model that always predicts “pass” scores 93.4% accuracy while catching no faults.

VI. RESULTS

A. Heterogeneous telemetry: recovering the normal manifold

On the synthetic telemetry benchmark (Table III, Fig. 1 shows the architecture), a local-only autoencoder learns only its own site’s normal regime and misclassifies other sites’ normal telemetry as anomalous, capping detection at ROC-AUC 0.83. Federation exposes the shared model to every regime and lifts ROC-AUC to 0.99—statistically indistinguishable from the 1.00 centralized upper bound—recovering 94% of the local-to-centralized gap. Minority-class F1 rises from 0.69 to 0.79. Because federation matches the centralized ceiling, the missing performance was never a matter of model capacity or training budget: it was missing cross-site signal that only collaboration supplies.

Table III. Telemetry autoencoder (5 heterogeneous sites, 5 seeds). Bold marks the best privacy-preserving configuration (federated).

Setting	ROC-AUC	F1
Local-only	0.83 $\pm$ 0.01	0.69
Federated (FedAvg)	0.99 $\pm$ 0.00	0.79
Centralized (upper bound)	1.00 $\pm$ 0.00	0.79

Federated detection approaches the centralized upper bound, far above local-only

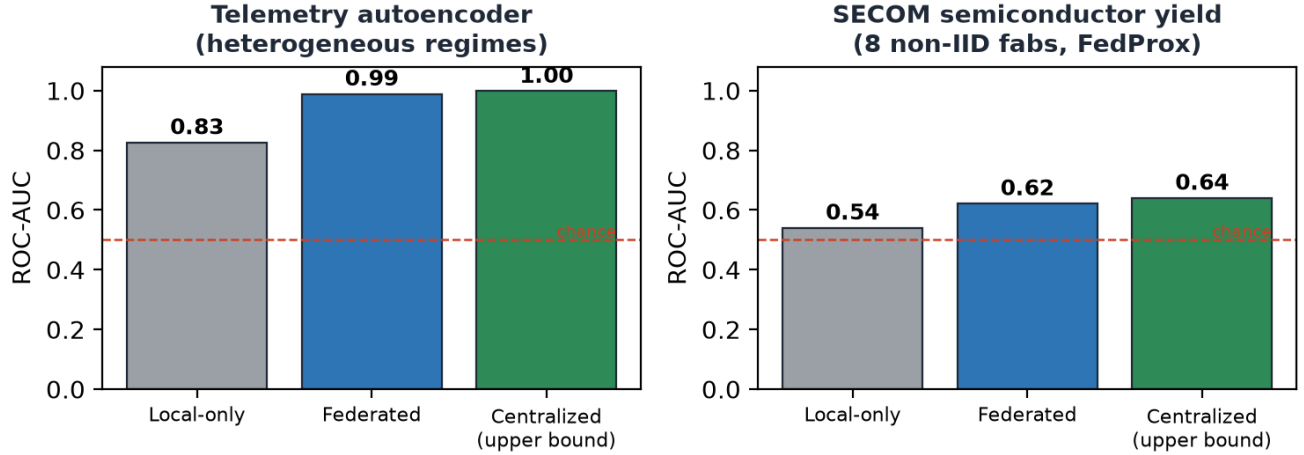


Fig. 2. Detection ROC-AUC: federation approaches the centralized upper bound and far exceeds local-only on both the telemetry autoencoder and SECOM (8 non-IID fabs; FedProx variant shown).

Table II. SECOM yield-anomaly detection across eight non-IID fabs ($\alpha = 0.3$, 12 seeds). Minority (fault) class metrics; centralized is an idealized, non-private upper bound. Bold marks the best privacy-preserving configuration (FedProx).

Configuration	ROC-AUC	F1 (fail)	Recall (fail)	Precision (fail)
Local-only	0.540 ± 0.014	0.093	0.162	0.069
Federated (FedAvg)	0.596 ± 0.052	0.144	0.141	0.150
Federated (FedProx)	0.622 ± 0.052	0.144	0.135	0.156
Federated (FedProx + FIWA)	0.618 ± 0.048	0.143	0.138	0.149
Federated (FedProx + DP)	0.554 ± 0.075	0.101	0.096	0.107
Centralized (upper bound)	0.640 ± 0.044	0.135	0.128	0.146

B. SECOM: the data-poor fab recovers rare-fault detection

The SECOM results (Table II, Fig. 2) sharpen the finding under realistic label skew. Several of the eight fabs hold essentially no failures under the $\alpha = 0.3$ Dirichlet split; in isolation they cannot detect a fault they have never seen, and local-only detection sits barely above chance at ROC-AUC 0.54. FedAvg raises minority-class F1 by 54% ($0.093 \rightarrow 0.144$) and ROC-AUC to 0.60 by pooling fault knowledge through weights alone. Under the heterogeneous split, FedProx's proximal term adds a further +4.4% ROC-AUC to 0.62—recovering roughly four-fifths of the way to the 0.64 centralized bound—confirming that curbing local drift matters most precisely when clients are most dissimilar.

C. Aggregation ablation, privacy, and communication cost

Fault-information-weighted aggregation (FIWA) did not beat FedProx on these linear detectors (-0.6% ROC-AUC). We report this as an honest negative result: FedProx's proximal term already captures the available minority-class signal, so upweighting fault-rich fabs adds nothing on a linear model—motivating fault-aware aggregation on nonlinear detectors and more extreme skew as future work, rather than claiming a new aggregator here.

Client-level differential privacy—clipping updates to unit L2 norm and adding Gaussian noise ($\sigma = 0.6$)—reduces

ROC-AUC from 0.62 to 0.55, a 10% utility cost in exchange for a formal, quantifiable privacy guarantee on top of the no-raw-data property. Crucially, the entire 25-round, 8-fab federated run exchanges only 160 KB of parameters in total; the protocol is bandwidth-trivial relative to moving telemetry, and moving telemetry is exactly what confidentiality forbids.

Two conclusions hold across both detector families. First, more local compute does not close the gap—the centralized ceiling is reached and then matched, so federation, not additional training, supplies the missing signal. Second, the benefit is largest for the sites that need it most: those starved of rare-fault examples gain the ability to detect faults they never locally observed.

VII. DISCUSSION

The practical implication for the semiconductor ecosystem is that reliability collaboration and data confidentiality are not mutually exclusive. A fab, a datacenter region, or even a set of competing vendors can jointly train a stronger anomaly detector while each keeps its raw telemetry entirely private. The quantified collaboration value—a data-poor site recovering rare-fault detection it cannot achieve alone—directly addresses the most common reason such collaborations are dismissed: the belief that useful pooling

requires sharing sensitive data. It does not; sharing model weights suffices, and the communication footprint is negligible.

We emphasize what our contribution is and is not. It is an open, reproducible benchmark and a measured effect, using detectors implemented transparently so reviewers can audit the privacy boundary in roughly a hundred lines of NumPy. It is not a new optimizer: we intentionally build on FedAvg, FedProx, and standard differential privacy so the results reflect what an adopter would actually obtain today with off-the-shelf FL.

VIII. LIMITATIONS AND FUTURE WORK

Our SECOM detectors are linear, chosen for auditability; nonlinear models may expose gains from fault-aware aggregation that FIWA could not surface on a linear classifier. The absolute ROC-AUC on SECOM is modest—a property of the dataset's weak, noisy yield signal—so we report relative improvements and a centralized bound rather than headline accuracy. The synthetic telemetry, while designed to reflect heterogeneous regimes, is not a substitute for real fleet data; extending the autoencoder experiment to the public ESA anomaly dataset [12], [13] of real spacecraft telemetry is a natural next step and is scaffolded in our released code. Finally, we assume a semi-honest aggregator; integrating secure aggregation [8] to defend against a curious server, and calibrating end-to-end (ϵ , δ) privacy budgets across rounds, are important directions for a deployable system.

IX. CONCLUSION

We presented an open, reproducible benchmark for federated anomaly detection on hardware telemetry and showed, on two non-confidential datasets, that federation closes most of the gap to a centralized upper bound while raw data never leaves its owner. Federation lifts telemetry detection from 0.83 to 0.99 ROC-AUC (94% of the gap) and, on non-IID SECOM, improves minority-class F1 by 54% and ROC-AUC from 0.54 to 0.62 with FedProx—exchanging only ~160 KB of parameters and supporting optional differential privacy. The central message for the semiconductor ecosystem is simple: keep your data, share the model. Cross-site and even cross-vendor reliability collaboration is achievable without data sharing, removing a key adoption barrier for AI in hardware.

ACKNOWLEDGMENT

The author thanks colleagues at Microsoft for reviewing the content, and the UCI Machine Learning Repository for the public SECOM dataset.

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. AISTATS**, 2017, pp. 1273–1282.
- [2] P. Kairouz et al., “Advances and open problems in federated learning,” **Foundations and Trends in Machine Learning**, vol. 14, no. 1–2, pp. 1–210, 2021.
- [3] M. McCann and A. Johnston, “SECOM data set,” UCI Machine Learning Repository, 2008. [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/SECOM>
- [4] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, “Federated optimization in heterogeneous networks,” in **Proc. MLSys**, 2020.
- [5] M. Abadi et al., “Deep learning with differential privacy,” in **Proc. ACM CCS**, 2016, pp. 308–318.
- [6] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, “Federated learning with non-IID data,” **arXiv:1806.00582**, 2018.
- [7] 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.
- [8] K. Bonawitz et al., “Practical secure aggregation for privacy-preserving machine learning,” in **Proc. ACM CCS**, 2017, pp. 1175–1191.
- [9] C. Dwork and A. Roth, “The algorithmic foundations of differential privacy,” **Foundations and Trends in Theoretical Computer Science**, vol. 9, no. 3–4, pp. 211–407, 2014.
- [10] J. An and S. Cho, “Variational autoencoder based anomaly detection using reconstruction probability,” SNU Data Mining Center, Tech. Rep., 2015.
- [11] M. Sakurada and T. Yairi, “Anomaly detection using autoencoders with nonlinear dimensionality reduction,” in **Proc. MLSDA Workshop**, 2014, pp. 4–11.
- [12] K. Kotowski et al., “European Space Agency benchmark for on-board anomaly detection in satellite telemetry (ESA-ADB),” 2024. [Online]. Available: <https://doi.org/10.5281/zenodo.12528695>
- [13] K. Kotowski et al., “European Space Agency benchmark for anomaly detection in satellite telemetry,” **arXiv:2406.17826**, 2024.