



The Price of Privacy

Fairness–Utility Trade-offs in Responsible Federated Learning

Jin-Hee Cho (Email: jicho@vt.edu)

Department of Computer Science, Virginia Tech

Joint work with Dawood Wasif, Dian Chen, and collaborators

FPAI-HPC '26

The 3rd Workshop on Federated and Privacy-Preserving AI for HPC, co-located with IEEE Cluster 2026

The Central Question

Federated learning keeps raw data local.

Privacy mechanisms may widen accuracy gaps across clients.

Can we protect privacy without widening performance gaps or sacrificing utility?



Responsible Federated Learning

Federated learning

- Organizations train a shared model while retaining local data.
- Only model updates move between clients and the server.
- This structure fits distributed healthcare, vehicles, finance, and HPC systems.

The responsible-AI challenge

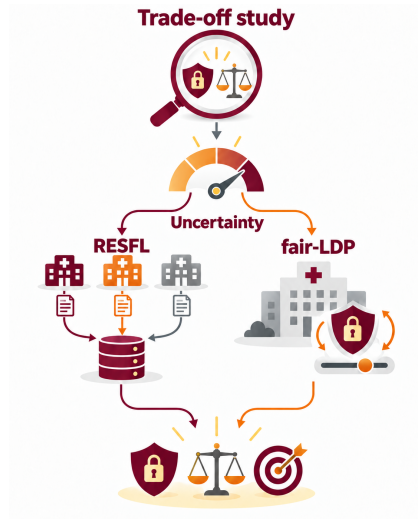
- Privacy controls add noise, encryption, or protocol constraints.
- Heterogeneous clients experience those constraints differently.
- Fairness, privacy, and accuracy must be evaluated together.

Talk structure

We first quantify how privacy protection affects fairness and utility, then use uncertainty to manage these trade-offs in autonomous driving and healthcare.

Research Roadmap

- **Tradeoff Analysis in Federated Learning (AAAI/ACM AIES 2025):** Compares privacy and fairness models, revealing trade-offs and costs.
- **Responsible Federated Learning (RESFL; ICLR 2026):** Uses uncertainty-guided aggregation to balance privacy, fairness, and accuracy.
- **Fairness-aware Local Differential Privacy (fair-LDP; IEEE ICDM 2025):** Adapts privacy budgets via uncertainty for fairer Alzheimer's diagnosis.



Study 1: Research Gap and Contribution (AAAI/ACM AIES 2025)

Research gap

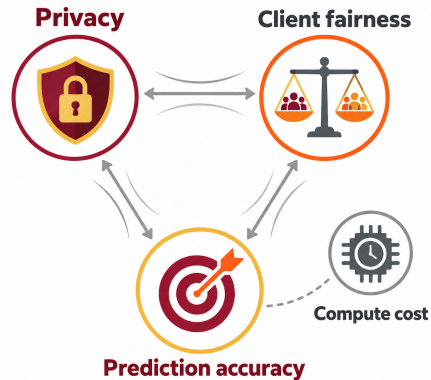
Prior work often examined privacy and fairness separately, with limited testing under heterogeneous data and privacy attacks.

Key contribution

We compare differential privacy (DP), homomorphic encryption (HE), and secure multiparty computation (SMC) with fairness-aware optimizers across benchmarks, Alzheimer's MRI, and fraud data.

Research questions

Which combinations best balance privacy leakage, client fairness, accuracy, and computational cost? How does tuning privacy or fairness change that balance?



Study 1: Experimental Coverage

13 privacy baselines	3 perturbation (local/global DP, gradient masking); 3 anonymization (k -anonymity, l -diversity, t -closeness); 2 cryptographic (HE, SMC); 5 hybrids (local or global DP with HE or SMC, plus HE+SMC).
5 fairness baselines	q-FedAvg, q-FedSGD, q-MAML, agnostic federated learning, and Ditto.
4 datasets and splits	MNIST, Fashion-MNIST, Alzheimer's MRI, and credit-card fraud; independent and identically distributed (IID) and heterogeneous (non-IID) clients.
Metrics	Privacy: membership inference attack success ratio (MSR), differential leakage rate (DLR). Fairness: loss disparity (LD), accuracy disparity (AD). Utility: global accuracy. Cost: computation overhead.
Threats	Privacy attacks: membership inference and differential leakage. Integrity attacks: data poisoning and backdoors. Fairness impact: disparities in attack effects across clients.

Study scale: 50 clients, 5 participating per round, 20 communication rounds, and 10 independent runs per setting.

Study 1: Privacy Mechanism Findings

Method	Observed pattern	Design implication
DP	Under heterogeneous (non-IID) client data, privacy noise can disproportionately degrade performance for clients with underrepresented data; we found up to 20% higher loss disparity .	Low computational cost, but privacy noise needs careful calibration.
HE	Loss and accuracy disparity stayed within 5% of baseline ; fairness was relatively stable as encryption parameters changed.	Preserves update fidelity, at 2–3× computation cost.
SMC	Often maintained fairer performance than DP under heterogeneous data.	Intermediate cost; the required number of participating clients affects security and dropout tolerance.

Takeaways: Data heterogeneity changes the cost of privacy for different clients. Compare client disparity alongside global accuracy and attack leakage.

Study 1: Fairness and Privacy Controls Interact

Lower LD = fairer client performance · Higher MSR = greater privacy leakage

Increase fairness weight: $q = 0 \rightarrow 10$

Local DP · non-IID Fashion-MNIST

Fairness improves

Loss disparity (LD)

0.0338  0.00392

High-loss clients receive more weight

Privacy weakens

Membership-inference success (MSR)

0.60  0.71

Fairness gain $\leftrightarrow$ greater privacy leakage

Increase DP budget: $\epsilon = 2 \rightarrow 8$

q-FedAvg · IID Fashion-MNIST · less privacy noise

Fairness improves

Loss disparity (LD)

0.00635  0.00220

A larger ϵ means weaker DP protection

Privacy weakens

Membership-inference success (MSR)

0.26  0.32

Fairness gain $\leftrightarrow$ reduced privacy protection

Optimizer choice is context-dependent

Ditto: Fashion-MNIST and MRI

Agnostic FL: non-IID MNIST and MRI **No universal best privacy–fairness balance**

Tune privacy and fairness jointly for the dataset and client distribution.

Study 2: Research Gap and Contribution (ICLR 2026)

Research gap

Existing federated learning approaches often protect privacy without reducing **pedestrian-detection disparities across skin-tone groups**. Adverse weather and lighting conditions can further widen these disparities.

Key contribution

Adversarial learning embeds a privacy attacker into training: the attacker tries to infer skin tone, while the feature extractor learns to prevent that inference and thereby reduce privacy leakage.

Research questions

Can this joint design reduce sensitive-attribute leakage and improve group fairness by narrowing pedestrian-detection gaps across skin-tone groups, without sacrificing overall detection accuracy? Does it remain reliable under adverse weather?

Driverless cars ‘more likely to run over black people’



A woman in Arizona was killed last year by a self-driving Uber vehicle

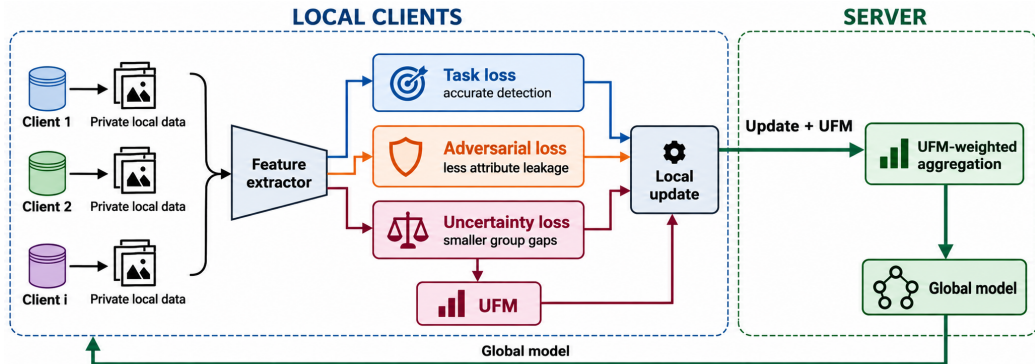
FUTURE PERFECT

A new study finds a potential risk with self-driving cars: failure to detect dark-skinned pedestrians

The findings speak to a bigger problem in the development of automated systems: algorithmic bias.



Study 2: RESFL: Joint Loss and UFM-Guided Aggregation



$$\mathcal{L}_{\text{local}} = \mathcal{L}_{\text{task}} + \lambda_{\text{priv}} \mathcal{L}_{\text{adv}} + \lambda_{\text{fair}} \mathcal{L}_{\text{uncertainty}}$$

Interpretation: Detection errors raise task loss. Gradient reversal weakens sensitive-attribute cues. Unequal group confidence raises uncertainty loss.

Weights: λ_{priv} and λ_{fair} tune privacy and fairness pressure during local training.

Study 2: RESFL: Privacy, Fairness, and Utility on FACET

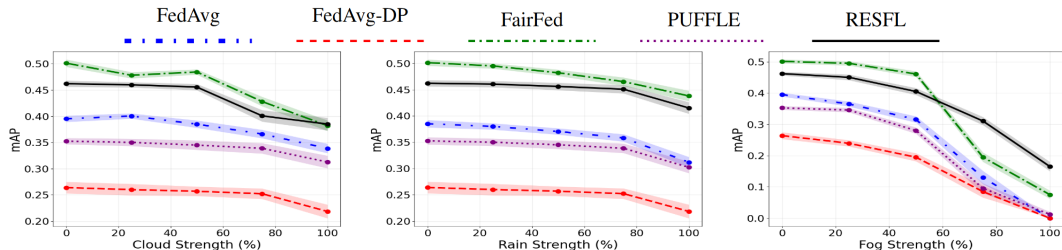
Uncertainty Fairness Metric (UFM): favors confident client updates with smaller group disparities.

Method	Prediction accuracy	Fairness	Privacy preservation	
	mAP ↑	EOP gap ↓	MIA SR ↓	AIA SR ↓
FedAvg	0.638	0.236	0.334	0.443
FedAvg-DP	0.461	0.288	0.236	0.263
FairFed	0.701	0.256	0.441	0.526
RESFL	0.665	0.196	0.209	0.183

Metric key: mAP = mean average precision (accuracy); EOP = equality of opportunity gap (group true-positive rates); MIA SR = membership-inference attack success rate; AIA SR = attribute-inference attack success rate. FACET means over three seeds.

Takeaways: Versus FedAvg, RESFL cuts membership-inference success by **37%** and the EOP gap by **17%**, while slightly improving mAP.

Study 2: RESFL: Robustness under Adverse Weather



At full cloud and rain intensity, RESFL retains **0.39** and **0.42 mAP**, respectively.

Takeaways: Severe fog hurts every method, but RESFL retains **0.17 mAP** at 100% fog; every plotted baseline falls below 0.10.

Study 3: fair-LDP for Federated Healthcare (IEEE ICDM 2025)

Research gap

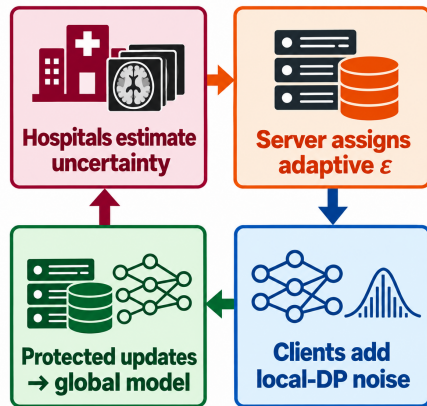
Uniform local differential privacy (LDP) noise can hurt hospitals unevenly when their datasets differ in size or quality. Few methods adapt privacy to client uncertainty and fairness together.

Key contribution

fair-LDP estimates client uncertainty with evidential neural networks, assigns adaptive privacy budgets (ϵ), and adds calibrated noise to local updates before sharing.

Research question

Can uncertainty-guided privacy budgets narrow client accuracy gaps while preserving useful Alzheimer's diagnosis under local DP?



Study 3: fair-LDP: Accuracy and Client Fairness on Alzheimer's MRI

Method	IID clients		Non-IID clients	
	Accuracy (%) $\uparrow$	LD $\downarrow$	Accuracy (%) $\uparrow$	LD $\downarrow$
LDP	72.96	6.49	75.14	0.440
q-FFL + LDP	≤ 50	0.0001	≤ 50	0.0022
fair-LDP	75.87	1.45	78.25	0.220

Metric key: Accuracy is the percentage of correct diagnoses. Loss disparity (LD) is the variance of client test losses; lower means fairer performance.

Takeaways: fair-LDP raises accuracy and lowers disparity; q-FFL + LDP falls to $\leq 50\%$ accuracy.

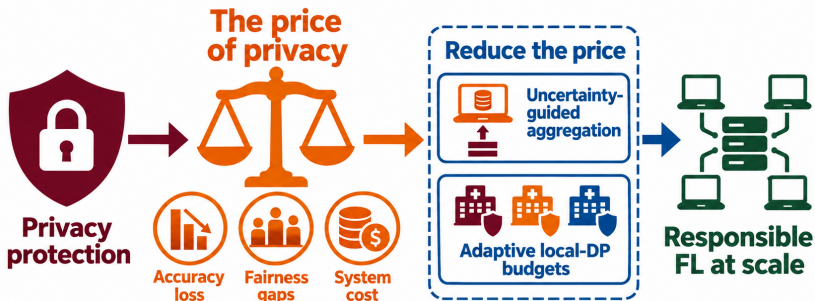
Study 3: fair-LDP: Privacy and Runtime on ADNI

Method	IID clients		Non-IID clients	
	MIA ASR ↓	RT ↓	MIA ASR ↓	RT ↓
LDP	0.3337	2.18	0.5532	2.15
q-FFL + LDP	0.5521	2.14	0.4338	2.13
fair-LDP	0.1155	2.16	0.3999	2.16
SMC	≈ 0	5.77	≈ 0	5.89
HE	≈ 0	130.00	≈ 0	137.93

Metric key: ADNI is the Alzheimer's Disease Neuroimaging Initiative dataset. MIA ASR is membership-inference attack success rate; RT is reported run time. Lower is better for both.

Takeaways: fair-LDP lowers attack success at LDP-like runtime; SMC and HE nearly eliminate attacks but cost much more.

Reducing the Price of Privacy



Measure the full price: utility, fairness, privacy, robustness, and cost.

Takeaways: Uncertainty-guided controls can reduce the trade-offs, but responsible FL must measure the full price at scale.

Thank you

Questions?

Jin-Hee Cho

Department of Computer Science, Virginia Tech

Email: jicho@vt.edu



Lab website

Research Project Website

Trustworthy and Responsible AI for Smart Critical Infrastructure