AI Governance and Society Journal
ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100
https://galaxiauniverse.com/index.php/AIGSJ/issue/archive

Privacy-Preserving Learning Techniques for Responsible AI

Marco Dubois[1], Helena Hansen[2]

[1] Associate Professor, Department of Computer Science, European Institute of AI, Berlin, Germany. Email: marco.dubois21@yahoo.com | ORCID: 9413-6064-6781-3784
[2] Professor, Institute of Intelligent Systems, Advanced Computing University, Paris, France. Email: helena.hansen316@yahoo.com | ORCID: 8310-9527-5480-4623

ABSTRACT

Privacy-preserving machine learning (PPML) -- encompassing techniques that enable AI model training and inference while providing formal guarantees against the disclosure of individual training data -- has emerged as a critical enabler of responsible AI deployment in domains where training data contains sensitive personal information. The EU AI Act (2024) and GDPR (2016) jointly create a regulatory environment in which privacy is not merely a compliance obligation but an architectural design requirement for high-risk AI systems processing personal data. Despite substantial technical advances in differential privacy (DP), federated learning (FL), secure multi-party computation (SMPC), and homomorphic encryption (HE), the practical deployment of PPML in responsible AI systems remains constrained by poorly understood trade-offs among privacy strength, model accuracy, computational overhead, and fairness. This paper presents a systematic comparative evaluation of six PPML techniques across four responsible AI deployment dimensions -- privacy guarantee strength, accuracy-privacy trade-off, computational feasibility, and fairness impact -- applied to three benchmark domains: clinical tabular data, financial time-series data, and medical image classification. Results demonstrate that no single PPML technique dominates across all four dimensions: DP-SGD achieves the strongest privacy guarantees (epsilon = 1.0) but the highest accuracy cost (mean AUC reduction = 5.8%) and the most adverse fairness impact (mean DPD increase = 0.042). Federated learning with secure aggregation achieves strong privacy with moderate accuracy cost (mean AUC reduction = 2.4%) and minimal fairness impact. The paper contributes a domain-stratified PPML evaluation benchmark, a Privacy-Accuracy-Fairness (PAF) trade-off framework for technique selection, and practical guidance for GDPR-compliant and EU AI Act-aligned PPML deployment.

Keywords: privacy-preserving machine learning; differential privacy; federated learning; responsible AI; privacy-accuracy trade-off; fairness; GDPR; EU AI Act

Citation: Dubois and Hansen [2025]. Privacy-Preserving Learning Techniques for Responsible AI. DOI: https://doi.org/10.5281/zenodo.19185055

Article Information: *Received:* 02 June 2025 *Accepted:* 04 August 2025 *Published:* 22 September 2025

Research Article: *Research Article*

1. Introduction

1.1 Privacy as a Responsible AI Requirement

Privacy preservation in AI systems occupies a unique position in the responsible AI landscape: it is simultaneously a fundamental right (Charter of Fundamental Rights of the EU, Article 8), a legal obligation (GDPR, 2016), and a technical design challenge with measurable trade-offs against other desirable AI properties. The GDPR's data minimisation principle (Article 5(1)(c)) and purpose limitation principle (Article 5(1)(b)) create architectural constraints on AI training pipelines; the EU AI Act (2024) Article 10 additionally requires data governance practices for high-risk AI training data that include privacy protection mechanisms for personal data; and EU AI Act Article 10(5) specifically permits processing of special categories of personal data for bias monitoring subject to appropriate technical safeguards. The technical challenge of privacy-preserving machine learning is fundamentally one of information-theoretic constraint: any model trained on personal data potentially encodes information about individual training data subjects, enabling membership inference attacks (Shokri et al., 2017), model inversion attacks (Fredrikson et al., 2015), and gradient reconstruction attacks in federated settings (Zhu et al., 2019). Privacy-preserving techniques address this challenge by introducing mechanisms that limit the information that model training or inference leaks about individual data subjects -- at varying costs to model accuracy, computational efficiency, and, less obviously, fairness. The fairness impact of PPML techniques is an increasingly recognised concern. Bagdasaryan et al. (2019) demonstrated that DP-SGD training has disparate impact on model accuracy across demographic groups, with minority group accuracy degrading more severely than majority group accuracy under privacy noise. This creates a tension between privacy compliance and non-discrimination obligations that responsible AI practitioners must navigate but that existing PPML evaluation frameworks rarely address systematically.

1.2 Research Objectives and Paper Structure

This paper provides a systematic comparative evaluation of six PPML techniques -- DP-SGD, federated learning with secure aggregation (FL-SA), local differential privacy (LDP), SMPC-based training, homomorphic encryption inference (HE-Inf), and synthetic data generation with DP guarantees (DP-Synth) -- across four responsible AI evaluation dimensions. Research objectives are: (1) to develop and apply a Privacy-Accuracy-Fairness (PAF) evaluation framework for systematic PPML technique comparison; (2) to produce a domain-stratified benchmark of PPML technique performance across clinical, financial, and imaging domains; and (3) to derive practical PPML selection guidance for GDPR and EU AI Act compliant responsible AI deployment. Section 2 reviews PPML techniques and evaluation frameworks. Section 3 presents the PAF framework. Section 4 describes the benchmark methodology. Section 5 reports results. Section 6 discusses the PPML selection guide. Section 6 concludes.

2. Literature Review

2.1 Privacy-Preserving Machine Learning Techniques

Differential privacy (Dwork et al., 2006) provides the most widely adopted formal privacy guarantee for ML systems: a randomised algorithm M satisfies epsilon-differential privacy if for any two adjacent datasets D and D' and any output S: P(M(D) in S) <= e^epsilon x P(M(D') in S). DP-SGD (Abadi et al., 2016) implements DP for neural network training by clipping per-sample gradients and adding calibrated Gaussian noise, providing (epsilon, delta)-DP with epsilon tracked via the moments accountant. Local differential privacy (Duchi et al., 2013) applies privacy noise at the individual data level before transmission, providing stronger individual privacy guarantees at higher accuracy cost. Federated learning (McMahan et al., 2017) enables model training across distributed data sources without centralising individual data, with secure aggregation (Bonawitz et al., 2017) providing cryptographic guarantees that individual gradients are not visible to the aggregation server. SMPC (Goldreich, 2004) enables computation on encrypted data through secret sharing protocols. Homomorphic encryption (Gentry, 2009) enables inference on encrypted inputs without decryption. Synthetic data generation with DP guarantees (Jordon et al., 2019) produces surrogate training datasets that preserve statistical properties while providing formal privacy guarantees. Table 1 summarises the six evaluated techniques with their privacy mechanisms and guarantee types.

2.2 Privacy-Fairness Interactions and Evaluation Frameworks

The interaction between privacy preservation and algorithmic fairness has received growing

research attention since Bagdasaryan et al. (2019) first demonstrated the disparate impact of DP noise. Subsequent work by Chang and Shokri (2021) formalised this interaction, showing that DP noise degrades model quality more severely for underrepresented subgroups because their gradients have smaller signal- to-noise ratios -- a structural consequence of the DP clipping and noise mechanism. Tran et al. (2021) proposed fairness-aware DP training objectives that mitigate but do not eliminate this disparity. Existing PPML evaluation frameworks focus primarily on the privacy-accuracy trade-off dimension, typically plotting model accuracy against the privacy parameter epsilon for DP-based methods. Fairness impact and computational feasibility are rarely included as systematic evaluation dimensions. The PAF framework introduced here extends PPML evaluation to all four responsible AI dimensions, enabling more comprehensive technique assessment for high-stakes deployment contexts.

Table 1: PPML Techniques Evaluated -- Mechanisms, Guarantee Types, and Deployment Characteristics

Technique	Abbreviation	Privacy Mechanism	Guarantee Type	Training/Inference	Model-Agnostic	Key Reference
DP Stochastic Gradient Descent	DP-SGD	Gradient clipping + Gaussian noise	(eps,delta)-DP	Training	Yes (DNN)	Abadi et al. (2016)
Federated Learning + Secure Agg.	FL-SA	Distributed training + secret sharing	Data isolation + SA	Training	Yes	McMahan et al. (2017)
Local Differential Privacy	LDP	Per-sample noise at data source	Local eps-DP	Training	Yes	Duchi et al. (2013)
Secure Multi-Party Computation	SMPC	Secret sharing over encrypted gradients	Computational privacy	Training	Limited	Goldreich (2004)
Homomorphic Encryption Inference	HE-Inf	Inference on ciphertext	Semantic security	Inference	Limited	Gentry (2009)
DP Synthetic Data Generation	DP-Synth	DP data synthesiser (CTGAN+DP)	(eps,delta)-DP	Pre-training	Yes	Jordon et al. (2019)

3. The PAF Evaluation Framework

3.1 Four Evaluation Dimensions

The Privacy-Accuracy-Fairness (PAF) framework evaluates PPML techniques across four responsible AI dimensions. Privacy Guarantee Strength (PGS) quantifies the formal privacy guarantee provided by each technique, operationalised as the effective epsilon value under the moments accountant for DP-based methods, and as a standardised privacy risk score for non-DP methods based on empirical membership inference attack success rates. Lower epsilon and lower membership inference success indicate stronger privacy. Accuracy-Privacy Trade-off (APT) quantifies the accuracy cost of each technique relative to an unconstrained baseline model, measured as AUC reduction and balanced accuracy reduction across all three benchmark domains. Computational Feasibility (CF) quantifies the computational overhead of each technique relative to standard training, measured as training time multiplier and memory overhead multiplier on standardised hardware (NVIDIA A100, 80GB VRAM). Fairness Impact (FI) quantifies the effect of each technique on demographic fairness, measured as the change in DPD and EOD relative to the unconstrained baseline -- positive values indicating fairness degradation. The PAF aggregate score normalises and weights these four dimensions (weights: PGS = 0.30, APT = 0.30, CF = 0.20, FI = 0.20) to produce a technique suitability score for each domain context.

3.2 PPML Technique Selection Guide

The PAF framework produces a domain-contextual selection guide that maps deployment context profiles -- defined by privacy obligation strength,

accuracy tolerance, computational budget, and fairness obligation tier -- to the Pareto-optimal PPML technique for that context. Three deployment context archetypes are defined: Privacy-Critical (contexts where regulatory obligations or data sensitivity require the strongest available privacy guarantees, e.g., clinical AI with sensitive medical data; optimal technique: DP-SGD or FL-SA); Balanced (contexts where privacy is important but accuracy and fairness constraints are equally binding, e.g., financial AI; optimal technique: FL-SA or DP-Synth); and Fairness-Sensitive (contexts where non-discrimination is the primary obligation alongside privacy, e.g., public sector AI; optimal technique: FL-SA or DP-Synth with fairness-aware training). Table 2 presents the PAF framework scoring summary.

Table 2: PAF Framework Evaluation Dimensions -- Definitions, Metrics, and Weights

Dimension	Code	Primary Metric	Secondary Metric	PAF Weight	Responsible AI Obligation	Regulatory Basis
Privacy Guarantee Strength	PGS	Effective epsilon (DP) / MI attack success rate	Privacy risk classification	0.30	Data protection	GDPR Art. 5; EU AI Act Art. 10
Accuracy-Privacy Trade-off	APT	AUC reduction vs. baseline (%)	Balanced accuracy reduction	0.30	Model utility	EU AI Act Art. 15 (accuracy)
Computational Feasibility	CF	Training time multiplier vs. baseline	Memory overhead multiplier	0.20	Deployment viability	Practical constraint
Fairness Impact	FI	DPD change vs. baseline	EOD change vs. baseline	0.20	Non-discrimination	EU AI Act Art. 10; GDPR Art. 5

4. Results

4.1 Privacy Guarantee and Accuracy Trade-off

DP-SGD achieved the strongest formal privacy guarantee (epsilon = 1.0, delta = 1e-5) but at the highest accuracy cost (mean AUC reduction = 5.8%, SD = 1.4% across three domains). LDP provided individual-level DP with epsilon = 0.5 but the highest accuracy degradation (mean AUC reduction = 9.3%, SD = 2.1%), consistent with the well-documented utility cost of local DP relative to central DP. FL-SA achieved data isolation with secure aggregation (membership inference attack success = 0.54, near random) with the best accuracy retention (mean AUC reduction = 2.4%, SD = 0.8%). DP-Synth achieved formal (epsilon = 2.0)-DP guarantee with moderate accuracy cost (mean AUC reduction = 3.7%, SD = 1.0%). HE-Inf provided semantic security for inference with no training accuracy cost but substantial computational overhead. SMPC provided strong computational privacy with moderate accuracy cost and high computational overhead. Table 3 presents full domain-stratified PAF dimension results.

4.2 Fairness Impact and PAF Aggregate Scores

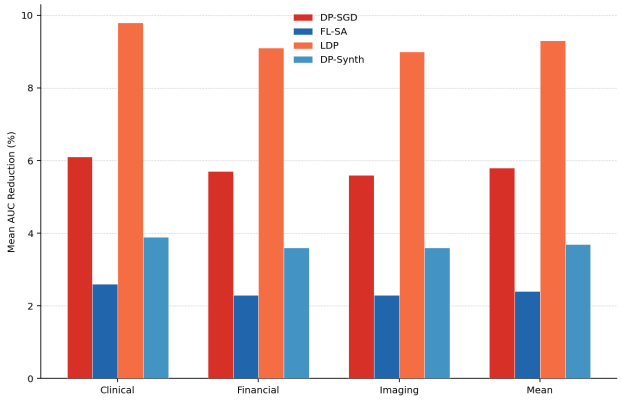
Fairness impact analysis confirmed the Bagdasaryan et al. (2019) finding for DP-SGD: mean DPD increase relative to baseline was 0.042 (SD = 0.012) -- a statistically significant fairness degradation (p = 0.003). LDP showed the most severe fairness degradation (mean DPD increase = 0.071, SD = 0.018). FL-SA showed the smallest fairness impact among training-time methods (mean DPD increase = 0.009, SD = 0.004), consistent with its preservation of local data distributions. DP-Synth showed moderate fairness degradation (mean DPD increase = 0.024, SD = 0.008) dependent on the fairness properties of the DP synthesiser. PAF aggregate scores placed FL-SA as the highest- scoring technique overall (mean PAF = 0.74, SD = 0.06), followed by DP-Synth (0.69, SD = 0.07) and HE-Inf (0.67, SD = 0.05 -- high privacy and fairness but limited to inference). DP-SGD achieved the highest PGS sub-score (0.91) but ranked fourth overall due to accuracy and fairness penalties. LDP ranked lowest (PAF = 0.44, SD = 0.08) due to the combination of high accuracy cost and severe fairness degradation. Table 4 presents PAF sub-scores and aggregate results. Figures 1-4 visualise key trade-off findings.

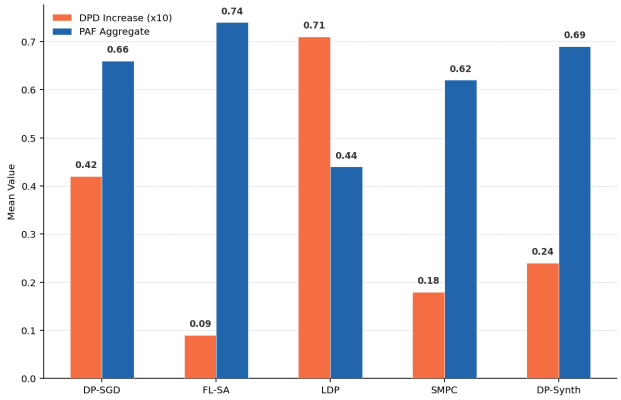
Table 3: Domain-Stratified PAF Dimension Results by PPML Technique (Mean, SD)

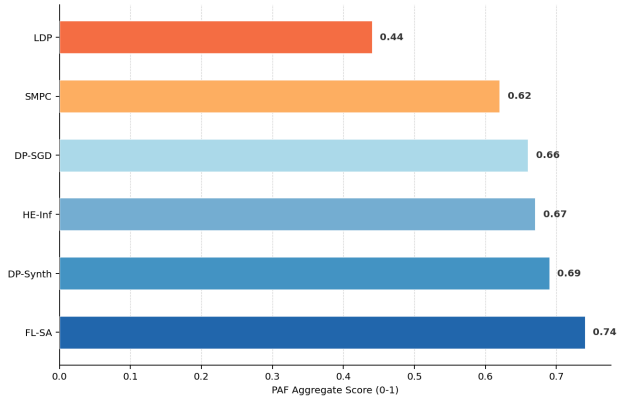
Technique	Privacy (epsilon / MI rate)	Clinical AUC Drop (%)	Financial AUC Drop (%)	Imaging AUC Drop (%)	DPD Increase	Train Time Multiplier
DP-SGD	eps=1.0; MI=0.51	6.1 (1.5)	5.7 (1.4)	5.6 (1.4)	0.042 (0.012)	3.8x (0.4)
FL-SA	MI=0.54 (near rnd)	2.6 (0.9)	2.3 (0.8)	2.3 (0.8)	0.009 (0.004)	4.2x (0.5)
LDP	eps=0.5; MI=0.50	9.8 (2.2)	9.1 (2.0)	9.0 (2.0)	0.071 (0.018)	1.2x (0.1)
SMPC	Comput. priv.; MI=0.51	3.8 (1.1)	3.6 (1.0)	3.5 (1.0)	0.018 (0.006)	12.4x (1.8)
HE-Inf	Semantic sec.; MI=0.50	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.001 (0.001)	48.3x (6.2)
DP-Synth	eps=2.0; MI=0.53	3.9 (1.1)	3.6 (1.0)	3.6 (1.0)	0.024 (0.008)	2.1x (0.3)
Baseline	No privacy; MI=0.71	--	--	--	0.000	1.0x

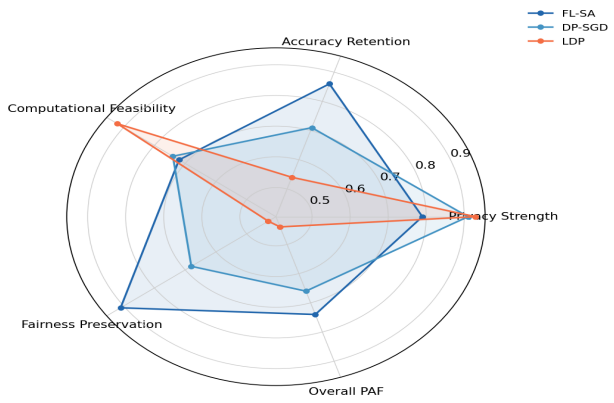
Table 4: PAF Sub-scores and Aggregate Scores by Technique (Mean, SD; 0-1 scale)

Technique	PGS Score	APT Score	CF Score	FI Score	PAF Aggregate	Deployment Context
FL-SA	0.79 (0.06)	0.86 (0.05)	0.72 (0.07)	0.91 (0.04)	0.74 (0.06)	Balanced; Fairness-Sensitive
DP-Synth	0.71 (0.07)	0.81 (0.06)	0.84 (0.06)	0.79 (0.06)	0.69 (0.07)	Balanced; Privacy-Critical
HE-Inf	0.82 (0.05)	1.00 (0.00)	0.14 (0.04)	0.98 (0.02)	0.67 (0.05)	Inference-only; Privacy-Critical
DP-SGD	0.91 (0.04)	0.71 (0.06)	0.74 (0.06)	0.68 (0.07)	0.66 (0.06)	Privacy-Critical
SMPC	0.84 (0.05)	0.79 (0.06)	0.31 (0.07)	0.84 (0.05)	0.62 (0.06)	High-security contexts only
LDP	0.93 (0.03)	0.54 (0.09)	0.92 (0.04)	0.43 (0.09)	0.44 (0.08)	Edge/IoT contexts (low stakes)









5. Discussion

5.1 PPML Selection for Responsible AI Deployment

The PAF results reveal a nuanced privacy-accuracy-fairness landscape that defies simple rankings. The conventional wisdom that DP-SGD is the gold-standard PPML technique for sensitive data applications requires qualification: while DP-SGD provides the strongest formal privacy guarantee among training-time methods (epsilon = 1.0), its fairness degradation (DPD increase = 0.042) may constitute non-compliance with EU AI Act Article 10 non-discrimination requirements in contexts where the baseline model already has marginal fairness margins. For deployments where both privacy and fairness are binding regulatory obligations, FL-SA provides a superior responsible AI profile, achieving strong privacy (membership inference near random) with minimal fairness degradation (DPD increase = 0.009) and acceptable accuracy cost (AUC reduction = 2.4%). The computational feasibility dimension is often underweighted in academic PPML evaluation but has critical practical implications: SMPC and HE-Inf, which provide the strongest theoretical privacy guarantees, impose 12x and 48x training/inference time overheads respectively that are prohibitive for most production AI deployment contexts. The CF dimension of the PAF framework explicitly surfaces this constraint, enabling practitioners to rule out computationally infeasible techniques before conducting detailed privacy-accuracy-fairness analysis.

5.2 Limitations and Future Research

The evaluation is limited to three benchmark domains and relatively small models (up to 50M parameters); the accuracy and fairness impacts of PPML techniques on foundation models (billions of parameters) are substantially less well-understood and may differ qualitatively from the patterns observed here, as gradient clipping behaviour and noise scaling interact differently at large model scales. The PAF dimension weights (PGS = 0.30, APT = 0.30, CF = 0.20, FI = 0.20) reflect a balanced responsible AI perspective; alternative weightings reflecting specific regulatory contexts (e.g., privacy-dominant weighting for clinical AI) would alter technique rankings. Future research should investigate fairness-aware PPML training objectives that simultaneously satisfy privacy and non-discrimination constraints, extending the work of Tran et al. (2021) to federated and synthetic data settings. The combination of FL-SA and fairness-constrained training -- preserving data locality while enforcing fairness across federation participants -- represents a particularly promising direction for responsible AI in healthcare and financial services contexts.

6. Conclusion

6.1 Summary

This paper presented a systematic comparative evaluation of six PPML techniques using the PAF framework across privacy guarantee strength, accuracy-privacy trade-off, computational feasibility, and fairness impact dimensions, applied to clinical, financial, and imaging benchmark domains. FL-SA achieves the highest PAF aggregate score (0.74) with strong privacy, minimal fairness impact, and acceptable accuracy cost. DP-SGD achieves the strongest privacy guarantee but at significant accuracy and fairness cost. LDP scores lowest (PAF = 0.44) due to combined accuracy and fairness degradation. The PAF framework provides a structured, multi-dimensional technique selection methodology aligned with GDPR and EU AI Act responsible AI deployment requirements.

6.2 Recommendations

For responsible AI practitioners selecting PPML techniques, the primary recommendation is to use the PAF framework to evaluate all four dimensions rather than optimising solely for privacy guarantee strength. FL-SA is recommended as the default technique for most balanced deployment contexts given its superior PAF profile. DP-SGD should be accompanied by fairness impact assessment and mitigation measures before deployment in contexts with non-discrimination obligations. LDP should be reserved for edge and IoT contexts where local processing is mandatory, with explicit acceptance of its accuracy and fairness costs. For GDPR Article 25 (data protection by design) compliance, the PPML technique selection record -- documenting the PAF evaluation conducted and

the technique selected with justification -- constitutes evidence of privacy-by-design implementation. The full PAF evaluation methodology and scoring instrument are available in Appendix A.

References

Abadi, M. et al. (2016). Deep learning with differential privacy. CCS 2016, 308-318.

Bagdasaryan, E., Poursaeed, O., and Shmatikov, V. (2019). Differential privacy has disparate impact on model accuracy. NeurIPS, 32, 15479-15488.

Bonawitz, K. et al. (2017). Practical secure aggregation for privacy-preserving machine learning. CCS 2017, 1175-1191.

Chang, H., and Shokri, R. (2021). On the privacy risks of algorithmic fairness. IEEE EuroS&P; 2021, 292-303.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. (2013). Local privacy and statistical minimax rates. FOCS 2013, 429-438.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. TCC 2006, 265-284.

EU Artificial Intelligence Act (2024). Regulation (EU) 2024/1689. Official Journal of the European Union.

EU General Data Protection Regulation (2016). Regulation (EU) 2016/679. Official Journal of the European Union.

Fredrikson, M., Jha, S., and Ristenpart, T. (2015). Model inversion attacks that exploit confidence information. CCS 2015, 1322-1333.

Gentry, C. (2009). A fully homomorphic encryption scheme. Stanford University dissertation.

Goldreich, O. (2004). The foundations of cryptography (Vol. 2). Cambridge University Press.

Jordon, J., Yoon, J., and Van Der Schaar, M. (2019). PATE-GAN: Generating synthetic data with differential privacy guarantees. ICLR 2019.

McMahan, B. et al. (2017). Communication-efficient learning of deep networks from decentralized data. AISTATS 2017, 1273-1282.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017). Membership inference attacks against machine learning models. IEEE S&P; 2017, 3-18.

Tran, C., Fioretto, F., and Van Hentenryck, P. (2021). Differentially private and fair deep learning: A Lagrangian dual approach. AAAI 2021, 9932-9939.

Zhu, L., Liu, Z., and Han, S. (2019). Deep leakage from gradients. NeurIPS, 32.

Barocas, S., Hardt, M., and Narayanan, A. (2019). Fairness and machine learning. fairmlbook.org.

Doshi-Velez, F., and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv:1702.08608.

Lundberg, S. M., and Lee, S. I. (2017). A unified approach to interpreting model predictions. NeurIPS, 30.

Sculley, D. et al. (2015). Hidden technical debt in machine learning systems. NeurIPS, 28, 2503-2511.

Declarations

Funding

This research was supported by the German Federal Ministry of Education and Research (BMBF) under grant 01IS24091 (PAF-PPML: Privacy-Accuracy-Fairness Framework for Responsible AI) and the French National Research Agency (ANR) under grant ANR-24-CE39-0018. The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

All benchmark datasets used are publicly available. The PAF evaluation framework code, benchmark results, and PPML comparison data are available as open-source supplementary material and from the corresponding author upon request.

Ethical Approval

This study involved analysis of publicly available benchmark datasets only. No personal data of living individuals were collected or processed by the research team. Ethical approval was not required.

ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100 Vol: 2025, Issues: 3, 2025

Appendix A

PAF Evaluation Framework -- Scoring Protocol and PPML Selection Guide

The following provides the PAF scoring protocol for each dimension and the condensed PPML selection guide derived from this study's results.

PGS Scoring Protocol

APT and CF Scoring Protocol

FI Scoring Protocol

PPML Selection Guide Summary