Research Summary

Privacy-Preserving Machine Learning (PPML)

Overview

This work presents a threat-driven Systematization of Knowledge (SoK) of Privacy-Preserving Machine Learning (PPML) from 2017–2025. Designed for practitioners building trustworthy AI under real adversarial constraints, the survey moves beyond algorithmic listings to analyze five concrete attack vectors—*Membership Inference, Model Inversion, LLM Data Extraction, Poisoning, and Gradient Leakage*. We map these threats to actionable defenses across Differential Privacy (DP), Federated Learning (FL), **Split Learning (SL)**, and SMPC, bridging the gap between theoretical guarantees and real-world deployment.

Threat-Driven Approach

- **Adversary Taxonomy:** We develop a taxonomy spanning white-box, black-box, and malicious settings, mapping attacks to specific stages of the ML lifecycle.
- **Differential Privacy (DP):** We evaluate DP as the primary mitigation for membership inference and memorization, supported by a **fully reproducible DP-SGD accounting experiment** (included in the companion repository).
- **Federated Learning (FL):** Evaluated for cross-institutional collaboration, with emphasis on **robust aggregation under Non-IID participation**.
- **Split Learning (SL):** Compared against FL, highlighting leakage risks in shared ("smashed") activations and mitigations such as **NoPeek** (Vepakomma et al.).

Key Contributions

To guide engineering decisions, the work provides:

- **Threat → Defense Map:** A structured mapping of attacks to frameworks (Opacus, Flower, PySyft, TenSEAL).
- **Deployment Decision Tree:** A logic flow for selecting hybrid stacks (e.g., *FL + Secure Aggregation* vs. *SMPC + DP*) based on data locality and latency constraints.
- **Indicative Micro-Benchmarks:** We provide provenance-logged performance indicators comparing the communication/compute overhead of DP-SGD vs. FL vs. SMPC pipelines.

Use-Cases & Deployment Reasoning

The framework is grounded in realistic deployments including:

- healthcare consortia (cross-silo FL)
- mobile personalization
- privacy-aware traffic analytics
- Collaborative Fraud Detection
- academic / education datasets

emphasizing trade-offs in **utility, latency, compute cost, and engineering complexity**.

Impact and Open Directions

The survey concludes with a targeted research agenda for 2026, focusing on **"Open Problems"** such as:

- poisoning-robust learning under Non-IID participation
- lowering cryptographic overhead in SMPC/HE
- integrating DP with robustness-aware aggregation
- standardized privacy auditing and certification

A key emerging area is the **Privacy–Hallucination Frontier in RAG pipelines**, where DP noise may degrade retrieval precision and increase hallucination risk — motivating ε-vs-hallucination evaluation as a future research direction.