Research Summary

Privacy-Preserving Machine Learning (PPML)

Overview

This work presents a threat-driven survey of Privacy-Preserving Machine Learning (PPML) from 2017–2025, designed for practitioners building trustworthy AI systems under real adversarial and regulatory constraints. Instead of organizing PPML defenses by algorithms, the survey analyzes them through the lens of five concrete attack vectors membership inference, model inversion, LLM data extraction, poisoning/backdoors, and gradient leakage. These threats are connected to actionable defenses that engineers can deploy today, bridging the gap between theoretical guarantees and real-world ML pipelines. The survey is motivated by the rise of privacy regulations such as GDPR, HIPAA, India's DPDP Act, and guidance frameworks like the NIST AI RMF, which make formal privacy controls essential in modern AI deployments.

Threat-Driven Approach

- The paper first develops an adversary taxonomy spanning white-box, black-box, semi honest, and malicious settings, clarifying how privacy attacks arise at different points in the ML lifecycle from data ingestion to inference. Each attack vector is mapped to suitable defenses and the practical conditions under which they work.
- Differential Privacy (DP) mitigates membership inference and memorization-driven extraction. A fully reproducible DP-SGD experiment demonstrates ε–accuracy trade-offs using a privacy accountant.
- Federated Learning (FL) enables cross-institutional training without centralizing raw data but requires secure aggregation, robust aggregation, and/or local DP to resist poisoning and gradient leakage.
- Secure Multi-Party Computation (SMPC) protects cross-silo gradients from server inference at the cost of higher communication overhead.
- Homomorphic Encryption (HE) provides encrypted inference and limited encrypted training for legally sensitive environments with higher latency tolerance.

Key Contributions

The survey highlights that effective PPML deployments rely not on isolated techniques but on hybrid stacks, such as:

- FL + Secure Aggregation + DP for collaborative statistical learning

- HE + DP for confidential inference with bounded leakage
- SMPC + DP for strict non-disclosure with formal guarantees

To guide practitioners, the project offers:

- A threat → defense → framework map that links each attack to concrete tools (e.g., Opacus, TF-Privacy, Flower, FLARE, TenSEAL, CrypTen).
- Indicative micro-benchmarks comparing accuracy and runtime trade-offs across non private, DP, FL, and SMPC-based pipelines.
- A PPML framework maturity matrix, rating popular libraries by usability, scalability, and compliance readiness.
- A reproducible DP-SGD notebook with ε-computation and accuracy curves.

Impact and Future Directions

A central practical element of the work is a deployment decision checklist and mini decision tree, helping engineers map their threat model and data-locality constraints to an appropriate PPML stack. Examples include:

- Cross-institutional data → start with FL + Secure Aggregation → add DP if membership risk is high.
- Centralized data with strong MI threat → choose DP-SGD.
- Legal non-disclosure requirements → prefer SMPC or HE.
- Low-risk environments → standard ML with structured privacy auditing.

These rules are demonstrated through micro-use-cases (e.g., hospital EHR consortia, keyboard prediction models, student performance datasets), giving practitioners intuition about trade-offs in utility, compute cost, and engineering complexity.

Impact and Future Directions

The paper concludes with a targeted research agenda focused on resolving persistent challenges: balancing privacy–utility–accuracy (especially for LLMs and PEFT + DP), scaling PPML under non-IID and limited-bandwidth setups, lowering cryptographic overhead, integrating poisoning robustness with DP guarantees, and building standardized privacy auditing and certification pipelines. Overall, the work positions PPML as a practical engineering discipline, providing a structured threat-driven map, reproducible artifacts, and decision tools to support compliant and trustworthy AI deployments in real-world systems.