

Research on a Quantitative Evaluation Model for Privacy Computing Technology Selection

Chengming Li *

School of Electronic and optical engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu, 210000, China

* Corresponding author Email: 18112874375@163.com

Abstract: Data collaboration in finance and healthcare has become unavoidable, though seldom for lack of competing proposals. In practice, privacy-preserving technology selection remains a half-informed exercise—shaped as much by vendor narratives as by empirical evidence—with the boundaries between Federated Learning (FL) and Secure Multi-Party Computation (MPC) frequently left vague in the literature. Rather than pursuing a theoretically pristine model, this work adopts an engineering stance, grounding its analysis in three pragmatic dimensions: privacy strength, model utility, and system overhead. By drawing on empirical benchmarks from top-tier venues instead of absolute performance predictions, the study comparatively examines FL and MPC across credit scoring and medical image segmentation tasks. The findings resist a universal optimum: MPC delivers robust security at a punitive cost, rendering it economically fragile for deep learning pipelines, whereas FL proves resilient under data heterogeneity yet remains exposed by its semi-honest assumptions. From these trade-offs, a visual decision tree is derived—intentionally coarse, but usable—to anchor technology selection amid conflicting business and compliance constraints.

Keywords: Privacy Computing; Technology Selection; Multi-Criteria Decision Making; Federated Learning; Secure Multi-Party Computation.

1. Introduction

With the enactment of the *Data Security Law* and the *Personal Information Protection Law*, the conflict between data silos and privacy compliance has become acute. Extracting value from encrypted data without compromising privacy is now a core challenge in financial risk control and smart healthcare. Federated Learning and Secure Multi-Party Computation are the two paths most people point to, but in real deployments the choice is rarely clean. Engineers end up hedging: FL is easier to ship, MPC is harder to get wrong on paper yet brutal in practice. The dilemma is less “which is better” than “which set of trade-offs is acceptable..”

The academic community has produced extensive optimizations for individual technologies. For instance, McMahan et al. proposed the FedAvg algorithm to address communication efficiency [1], while Demmler et al. developed the ABY framework to significantly reduce the overhead of secure computation [2]. However, most existing studies treat these problems in isolation, focusing either solely on accuracy or latency, with few attempts to weigh them on the same scale. Worse still, current surveys often remain at a qualitative level, leaving decision-makers without actionable quantitative tools.

This paper attempts to fill that gap. We must concede, however, that constructing a universally applicable quantitative model is extremely challenging, as privacy protection is inherently a multi-dimensional game. Therefore, instead of pursuing absolute predictive precision, this study focuses on revealing the relative strengths and weaknesses of different technical solutions within the “Privacy-Utility-Overhead” trilemma. By systematically reviewing authoritative empirical data from recent years, we conduct an in-depth comparison of FL, MPC, and their hybrid variants across two scenarios: joint credit scoring and medical image segmentation. The goal is not merely to rank algorithms, but to clarify the physical boundaries of technology selection—

specifically, determining when sacrificing efficiency for security is justified, and when excessive cryptographic protection becomes a barrier to business implementation.

2. Related Work

2.1. Performance Trade-offs and Heterogeneity in Federated Learning

As a distributed machine learning paradigm, FL achieves efficient model aggregation via the FedAvg algorithm.[1] In real-world deployments, however, data is often Non-Independently and Identically Distributed (Non-IID), a phenomenon particularly evident in medical imaging and financial risk control. Li et al. demonstrated that Non-IID data causes local model updates to deviate from the global optimum, triggering “Client Drift” and hindering convergence. To mitigate this, the FedProx algorithm introduces a proximal term into the loss function to constrain local deviations, proving significantly more stable on Non-IID datasets [3].

Furthermore, the privacy mechanism of native FL is heuristic rather than cryptographically complete. The Deep Leakage from Gradients (DLG) attack proposed by Zhu et al. proved that original training images can be reconstructed with high fidelity from gradient exchanges [4]. To enhance privacy, FL is often combined with Differential Privacy (DP), adding calibrated noise to gradients to provide ϵ -DP guarantees, inevitably trading off model accuracy [5]. None of this balances neatly. Privacy, utility, and overhead pull in different directions, and FL deployment ends up a set of uneasy trades. Mobile keyboard prediction deployments make the point plainly: communication overhead is not a side issue, it is a primary source of deployment delays and cost overruns[6]. We do not pretend this triad admits a clean optimum.

2.2. Efficiency Bottlenecks in Secure Multi-Party Computation

MPC offers stronger security semantics than FL. The ABY

framework by Demmler et al. combines Arithmetic, Boolean, and Yao circuits to provide efficient mixed-protocol support. The cryptographic overhead is not a small tax. It is punitive. Benchmarks show orders-of-magnitude slowdowns and relentless round trips, enough that anyone who has tried MPC in production will tell you: while the cryptographic foundations are rigorous, computational overhead remains prohibitive[2]. We treat these numbers not as warnings but as facts that shape what can realistically be built. Experiments on MNIST indicate that introducing Fully Homomorphic Encryption (FHE) may degrade training speed by over 20 times, while Secure Aggregation (SecAgg) protocols significantly increase communication latency [6]. Consequently, MPC is currently more applicable to scenarios with lower feature dimensions and simpler model structures; applying full-fledged MPC directly to high-dimensional image processing remains a formidable engineering challenge.

2.3. Technology Selection Methods in Information Systems

In the field of Multi-Criteria Decision Making (MCDM), the Analytic Hierarchy Process (AHP) and TOPSIS have been widely used for IT architecture and security technology selection [7]. However, few studies have applied MCDM to privacy computing, and even fewer have incorporated empirically derived quantitative inputs. Most selection advice remains at the qualitative level (e.g., "FL for big data, MPC for high security"). This paper contributes by integrating the MCDM model with empirical data extracted from literature to establish a quantifiable decision support system for privacy computing.

3. Construction of the Quantitative Evaluation Model

3.1. Problem Definition and Scenario Settings

This study sets up two representative high-value scenarios:

Joint Financial Credit Scoring: Multiple banks possess horizontally partitioned user data aiming to train a logistic regression model. This scenario involves medium-dimensional data and imposes high demands on real-time performance and auditability.

Medical Image Segmentation: Multiple hospitals possess CT/MRI data aiming to train a U-Net model. This scenario involves extremely high-dimensional data with severe Non-IID distribution, requiring high model accuracy but being relatively tolerant of latency.

Three typical technical schemes are compared:

Scheme A: Pure Federated Learning (FedAvg or FedProx), optionally with Differential Privacy.

Scheme B: Pure Secure Multi-Party Computation (based on ABY or SPDZ protocols).

Scheme C: Hybrid Architecture (FL framework + MPC/HE for secure aggregation).

3.2. Evaluation Dimensions and Quantifiable Metrics

To objectively compare technologies, this paper establishes an evaluation system comprising three core dimensions, with all metrics sourced from public literature.

Privacy Protection Strength (Cp)

Measures the mathematical security guarantee level. If Differential Privacy is applied, the ϵ value serves as the

quantitative indicator (smaller ϵ implies stronger privacy). Otherwise, grading is performed based on the type of attack resisted (e.g., resistance to malicious adversaries scores higher).

Model Utility Retention (Cu)

Measures the degradation of business effectiveness. Defined as the ratio of the privacy-preserving model's performance to that of the plaintext baseline (percentage). AUC retention is used for finance, and Dice coefficient retention for healthcare[3,8].

System Extra Overhead (Co)

Measures the additional costs incurred by privacy protection. Combines the end-to-end training time multiplier and per-round communication volume multiplier, followed by normalization.

3.3. Multi-Criteria Evaluation Model

Let the set of technical schemes be $S=\{A, B, C\}$, and the set of evaluation dimensions be $C=\{Cp, Cu, Co\}$. A weight vector $W=\{Wp, Wu, Wo\}$ is introduced, representing preferences for privacy, utility, and overhead. The comprehensive score function is defined as:

$$Score_i = w_p \cdot x_{ip} + w_u \cdot x_{iu} + w_o \cdot (1 - x_{io})$$

By adjusting weight configurations, different organizational preferences can be simulated. Furthermore, a 3D scatter plot will be drawn to identify the Pareto Optimal Set that cannot be simultaneously outperformed by other schemes.

4. Case Studies and Model Application

Statement on Data Sources and Modeling Limitations

Before proceeding with calculations, it is necessary to transparently state the data sources and processing methods used in this chapter. To ensure objectivity, all performance metrics are strictly extracted from empirical studies published in top-tier conferences and journals in recent years[1,2].

However, an unavoidable reality is that hardware configurations, network bandwidth, dataset scales, and even code implementations vary significantly across different literatures. Directly forcing these absolute values into the model would not only introduce huge errors but also mislead readers regarding technical feasibility. Therefore, this study adopts a relative comparison approach rather than absolute prediction. That would be dishonest given how scattered the underlying benchmarks are. Instead, we settle for relative trends and order-of-magnitude differences—coarse, but at least not pretending to a precision the data cannot support. This is less elegant, but more usable. During modeling, raw data are normalized to strip away the influence of hardware environments, thereby exposing the essential trade-offs within the "Privacy-Utility-Overhead" trilemma. In other words, the goal is not to tell you exactly *how much slower* MPC is than FL, but to determine *at what level of data sensitivity* such a loss of speed becomes worthwhile.

4.1. Case Setup and Parameter Input

4.1.1. Financial Joint Credit Scoring Scenario

Assume two banks aim to jointly train a logistic regression model for credit assessment. The data consists of low-dimensional features (approx. 100 dimensions) and is typically IID.

Baseline: Plaintext training yields an AUC of 0.82[1].

Scheme A (Pure FL): Based on FedAvg, AUC retention exceeds 99%, but end-to-end training time increases by about 50%[1]. This aligns with observations from large-scale mobile keyboard prediction deployments, where communication latency is a primary bottleneck [6].

Scheme B (Pure MPC): Based on ABY, offering strong security against malicious adversaries, but communication volume may surge to hundreds of times that of plaintext, resulting in extremely high training latency, which is often unacceptable in practice[2].

Scheme C (Hybrid): FL framework combined with Paillier Homomorphic Encryption, effectively preventing gradient leakage, with almost no AUC loss and significantly lower overhead than pure MPC.

4.1.2. Medical Image Segmentation Scenario

Assume three hospitals jointly train a U-Net model for brain tumor segmentation (BraTS dataset)[8]. The data consists of high-dimensional images with severe Non-IID distribution[3].

Baseline: Centralized training yields a Dice coefficient of 0.85[8].

Scheme A (Pure FL): Using FedProx, Dice retention reaches 97.6%, but due to the large model size, per-round communication pressure is immense[8,9].

Scheme B (Pure MPC): Direct ciphertext training of CNNs incurs prohibitively large computational overhead, potentially increasing training time by more than three orders of magnitude (over 1000x), making it practically infeasible under current hardware[2].

Scheme C (Hybrid): FedProx combined with Differential Privacy. To meet stringent medical data compliance, noise must be added to gradients, likely reducing Dice retention to around 91.8% and worsening training stability[5].

4.2. Data Summary

Table 1. Technical Metrics for the Financial Credit Scoring Scenario [1,2]

Technical Scheme	Privacy Strength (Cp)	Model Utility (AUC Retention) (Cu)	System Overhead (Co)
Scheme A (FedAvg)	0.4 (Semi-honest)	0.99	0.6 (High Comm.)
Scheme B (MPC/ABY)	1.0 (Malicious)	0.995	0.1 (Very Slow)
Scheme C (FL+HE)	0.8 (Anti-Gradient Leak)	0.99	0.4 (Med. Overhead)

Table 2. Technical Metrics for the Medical Image Segmentation Scenario [2,5,9]

Technical Scheme	Privacy Strength (Cp)	Model Utility (Dice Retention) (Cu)	System Overhead (Co)
Scheme A (FedProx)	0.3 (Vulnerable to Inf.)	0.976	0.5 (Large Transfer)
Scheme B (MPC/CNN)	1.0 (Full Ciphertext)	0.98	0.01 (Near Infeasible)
Scheme C (FL+DP)	0.9 (ϵ -DP)	0.918	0.45 (Slow Conv.)

4.3. Model Calculation and Results

4.3.1. Comprehensive Scores under Different Weight Configurations

Scenario 1: Financial Scenario (Efficiency-First vs. Security-First)

Configuration 1 (Efficiency-First): $W_p=0.2$, $W_u=0.4$, $W_o=0.4$

Score_A = 0.716, Score_B = 0.598, Score_C = 0.676 → Ranking: A > C > B

Configuration 2 (Security-First): $W_p=0.6$, $W_u=0.2$, $W_o=0.2$

Score_A = 0.398, Score_B = 0.819, Score_C = 0.758 → Ranking: B > C > A

Scenario 2: Medical Scenario (Accuracy-First vs. Compliance-First)

Configuration 3 (Accuracy-First): $W_p=0.1$, $W_u=0.7$, $W_o=0.2$

Score_A = 0.813, Score_B = 0.788, Score_C = 0.823 → Ranking: C > A > B

Configuration 4 (Compliance-First): $W_p=0.7$, $W_u=0.1$, $W_o=0.2$

Score_A = 0.408, Score_B = 0.800, Score_C = 0.812 → Ranking: C > B > A

4.3.2. Pareto Frontier Analysis

Plotting the 3D "Privacy-Utility-Overhead" space reveals that in the financial scenario, Schemes A and C constitute the Pareto Frontier. In the medical scenario, however, pure MPC (Scheme B) falls outside the frontier due to excessive overhead, marking it as a "dominated solution." This suggests that full ciphertext training of CNNs is currently economically inefficient [2].

5. Discussion and Decision Support

5.1. Real-World Implications of the Quantitative Results

Once regulation starts demanding defense against malicious insiders, the middle ground of FL plus HE is what survives, not because it is clean, but because the alternatives are worse [2,4]. Medical settings are even less forgiving: MPC drops out early, and you live with FL supplemented with ad hoc privacy measures. In medical scenarios, pure MPC essentially exits the competition early. Reality forces us to accept FL as the backbone, patched with DP or HE for compliance, despite the inevitable loss in accuracy. [5,8]

5.2. Construction of the Technology Selection Framework

Based on the above analysis, we derive the following visual decision tree for engineering practice.

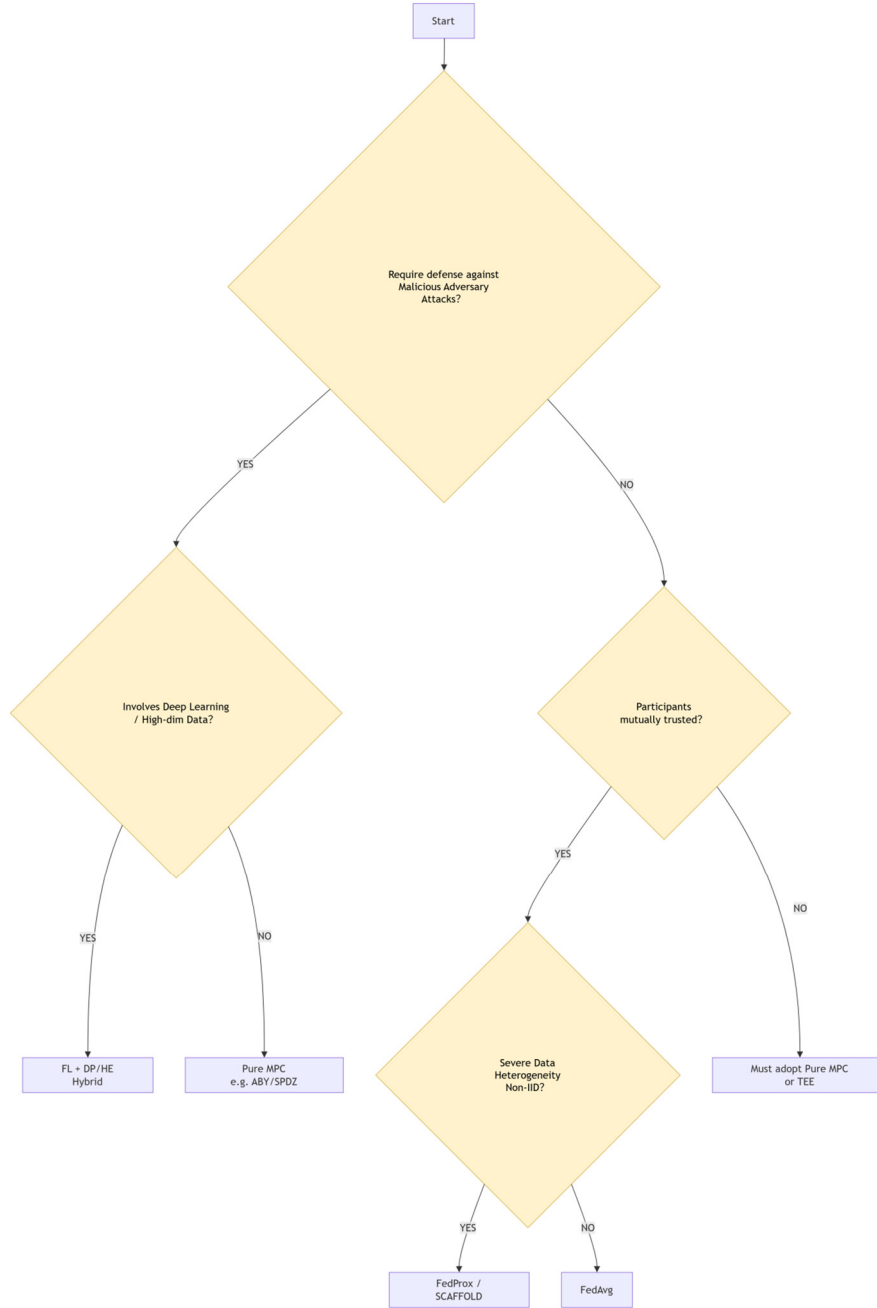


Fig 1. Technology Selection Decision Tree for Privacy Computing

6. Conclusion and Future Work

6.1. Research Summary

Addressing the lack of quantitative criteria for privacy computing technology selection, this paper conducted a comparative study based on multi-criteria decision-making. Rather than pursuing absolute performance predictions, the study focused on revealing the relative strengths and weaknesses of FL and MPC within the "Privacy-Utility-Overhead" trilemma. Through a systematic review of authoritative empirical data, instantiation was performed in financial and medical scenarios. The research clarified the Pareto optimal boundaries of different technical schemes and derived a visual technology selection decision tree. This framework aims to help practitioners rationally assess the costs and benefits of different choices, avoiding blind adherence to trends or over-engineering.

6.2. Limitations and Future Work

Although this paper strives for rigor, limitations remain:

Data Timeliness: Model inputs rely on public benchmarks; future work needs dynamic calibration mechanisms to adapt to hardware evolution.

Technology Extensibility: Future evaluations should incorporate emerging paradigms such as TEE and post-quantum cryptography.

Automated Tooling: Future work could develop a Web-based automated recommendation system that outputs optimal technical combinations based on business descriptions.

References

- [1] McMahan, B., Moore, E., Ramage, D., et al. (2017). Communication-efficient learning of deep networks from

- decentralized data. In *Artificial intelligence and statistics* (pp. 1273–1282). PMLR.
<https://doi.org/10.5555/3173033.3173186>
- [2] Demmler, D., Schneider, T., & Zohner, M. (2015). ABY-A framework for efficient mixed-protocol secure two-party computation. In *NDSS Workshop on Network and Distributed System Security (WAHC)*.
- [3] Li, T., Sahu, A. K., Zaheer, M., et al. (2020). Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2, 429–450.
- [4] Zhu, L., Liu, Z., & Han, S. (2019). Deep leakage from gradients. In *Advances in Neural Information Processing Systems* (pp. 14774–14784).
- [5] Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4), 211–407.
<https://doi.org/10.1561/04000000042>
- [6] Hard, A., Rao, K., Mathews, R., et al. (2018). Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*.
- [7] Saaty, T. L. (2008). Decision making with the analytic hierarchy process. *International Journal of Services Sciences*, 1(1), 83–98. <https://doi.org/10.1504/IJSSCI.2008.017249>.
- [8] Sheller, M. J., Edwards, B., Reina, G. A., et al. (2020). Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports*, 10(1), 12598. <https://doi.org/10.1038/s41598-020-69255-z>.
- [9] Nasr, M., Shokri, R., & Houmansadr, A. (2020). Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2020 IEEE Symposium on Security and Privacy (SP)* (pp. 739–753). IEEE.
<https://doi.org/10.1109/SP40000.2020.00023>