International Journal of Engineering and Techniques - Volume 12 Issue 3, May-June - 2026

RESEARCH ARTICLE OPEN ACCESS

Privacy-Preserving Federated Learning Framework for Secure Data Integration in Digital Real Estate Ecosystems

Ms. Anushka Prasad Joshi[1], Prof. Sachin Bhosale[2], Dr. Shubhangi Gunjal[3], Dr. Anand Khatri[4]

[1 2 3 4] *Dept Of Computer Engineering, Jaihind College Of Engineering*
Savitribai Phule Pune University, Pune, India
anushkap943@gmail.com | ssachinbhosale@gmail.com | shubhangi.gunjal83@gmail.com | khatrianand@gmail.com

Abstract:

With real estate markets digitalizing at a remarkable pace, there's a growing — and largely unmet — need for machine learning systems that can harness multi-institutional data without putting privacy or regulatory standing at risk. In this paper, we present a Privacy-Preserving Federated Learning (PP-FL) framework built specifically for digital real estate ecosystems. Our approach lets distributed stakeholders — property agencies, government land registries, financial institutions, and PropTech platforms — collaboratively train predictive models without ever pooling their raw transaction or personal records in one place. We've designed the system around three interlocking privacy layers: a DP-SGD-based differential privacy optimizer, a homomorphic encryption scheme for gradient transmission, and a secure multi-party computation protocol to safeguard intermediate model states. On top of that, a blockchain-backed audit mechanism using zero-knowledge proofs provides verifiable, regulator-friendly compliance. When we tested the framework on a simulated dataset of 2.4 million real estate transactions spanning multiple institutional clients, it achieved 91.8% prediction accuracy — just 2.4 percentage points behind a fully centralized model — while holding the differential privacy budget to $\varepsilon = 0.5$, cutting communication overhead by 65% relative to naive federated approaches, and satisfying both GDPR and RERA requirements. We believe these results make a strong case that high-utility, privacy-first collaborative learning is not just theoretically possible but practically deployable in today's real estate sector.

Keywords: Federated Learning, Differential Privacy, Homomorphic Encryption, Real Estate Analytics, Secure Multi-Party Computation, Blockchain, GDPR Compliance, Data Sovereignty, PropTech, Zero-Knowledge Proofs.

I. INTRODUCTION

Real estate, with a global valuation hovering near USD 326.5 trillion as of 2023, is arguably the world's largest single asset class — and it's undergoing a digital overhaul. PropTech platforms have multiplied rapidly, automated valuation models (AVMs) are reshaping how properties are priced, and AI-driven analytics are becoming standard tools for market forecasting [1]. Yet for all that technological momentum, one problem keeps coming back: real estate data is extraordinarily sensitive. It touches on personally identifiable information (PII), financial assessments, ownership histories, and geographic patterns — the kind of information that, if mishandled, creates serious regulatory and reputational exposure [2].

What makes this especially thorny is that useful ML models in this domain almost always require data from multiple parties. Yet conventional centralized approaches demand that everyone pool their records into a shared repository — something most stakeholders simply can't agree to. Property agencies are locked into client confidentiality commitments. Government registries operate under statutory data protection rules, including the GDPR across the EU and RERA in India. Banks and lenders navigate stringent banking secrecy regimes, and PropTech firms treat their proprietary datasets as competitive crown jewels [3]. Forcing data centralization, in this environment, isn't just impractical — it's often legally and ethically untenable.Federated Learning (FL), which McMahan et al. [4] introduced as a way to train models across distributed nodes without centralizing raw data, looked like it might solve this. And in many ways it does — each participant keeps

their data local, sharing only gradient updates rather than records. But we've seen, through our own reading and prior work, that FL alone doesn't deliver true privacy. Gradient inversion attacks can reconstruct training samples from shared updates [5]; membership inference techniques can reveal whether a specific record was used [6]; and adversarial participants can mount model poisoning campaigns [7]. The architecture is sound in principle, but the privacy guarantees it offers aren't sufficient for high-stakes real estate contexts.

This paper is our attempt to close that gap. We propose a comprehensive PP-FL framework designed from the ground up for the legal, competitive, and operational realities of digital real estate. The framework combines differential privacy, homomorphic encryption, and secure multi-party computation within a blockchain-backed architecture. Our key contributions are:

II LITERATURE REVIEW

Federated Learning Foundations

The FedAvg algorithm, as proposed by McMahan et al. [4], gave the field its first practical recipe for distributed model training: compute gradients locally, aggregate them on a central server, update the global model, repeat. It was a genuinely elegant idea. Building on that, Li et al. [8] tackled the messier reality of non-IID data distributions — the situation where each participant's data looks quite different from everyone else's. That problem is front and center in real estate, where urban agencies, suburban platforms, and rural registries each hold data shaped by wholly different market conditions.

Kairouz et al. [9] then provided the field with something it badly needed: a rigorous theoretical treatment of what FL actually guarantees in terms of privacy, robustness, and communication efficiency. Their analysis surfaced gradient leakage as a fundamental, not incidental, vulnerability — and that finding is particularly sobering in real estate, where a reconstructed transaction record isn't just an embarrassment but a potential regulatory violation.

2.2 Privacy Mechanisms in Distributed Learning

Differential privacy, defined formally by Dwork et al. [10], gives us a mathematical handle on privacy risk: it guarantees that the output of any computation looks essentially the same whether or not any given individual's record was included. It's a powerful idea. Abadi et al. [11] translated it into a workable deep learning optimizer — DP-SGD — that clips gradient norms and injects calibrated Gaussian noise at each step. Their moments accountant framework made it possible to track cumulative privacy expenditure across training, which is something we rely on directly in our implementation.

On the encryption side, Gentry's formalization of homomorphic encryption [12] opened the door to doing arithmetic on ciphertext — meaning a server can aggregate encrypted gradients without ever seeing what's inside them. Phong et al. [13] demonstrated this in a federated learning context, showing it's practical even if computationally heavy. For real estate, where the privacy stakes are high enough to justify computational cost, that tradeoff seems entirely reasonable. Separately, Bonawitz et al. [15] showed that secure multi-party computation can be deployed at scale — hundreds of participants — without collapsing under its own overhead, which gave us confidence that SMPC is viable beyond controlled experiments.

2.3 Real Estate Data Analytics and Privacy

Property valuation has long leaned on hedonic pricing models [16] — essentially regression-based frameworks that decompose a property's price into its constituent attributes. More recently, Park and Bae [17] showed that ensemble machine learning methods can outperform traditional hedonic models by a significant margin, particularly when they can draw on rich combinations of structural, locational, and temporal features. The catch, as is often true in ML, is that these better-performing models hunger for large, centralized datasets — creating exactly the kind of privacy tension we're trying to resolve.

From a regulatory standpoint, the landscape is demanding. GDPR Article 9 governs sensitive personal data, RERA Section 84 restricts what can be disclosed about property transactions, and the US Fair Housing Act has implications for how algorithmic models can use demographic proxies. Voigt and Von dem Bussche [18] offer one of the more thorough analyses of what GDPR compliance actually requires from ML pipelines — and those requirements inform several of our design choices directly.

III.PROPOSED PRIVACY-PRESERVING FEDERATED LEARNING FRAMEWORK

3.1 System Architecture Overview

At its core, our PP-FL framework is organized into three layers that work in concert: a Client Node Layer where local training happens, a Federated Aggregation Layer that handles secure model updates, and a Blockchain Audit Layer that keeps a

tamper-proof compliance record. Figure 1 maps out how data and gradients flow through the system, and where each privacy mechanism hooks in.

3.2 Client Node Layer

Every client node in our framework — whether it's a property agency, a PropTech platform, or a government registry — operates as a fully sovereign data holder. Nothing leaves its local environment in raw form, ever. The node trains a local neural network on its private dataset, computes gradients using the DP-SGD optimizer, and only then packages those gradients for transmission.

Local Model Training: For each training round t, the client computes gradient g_i(t) for its local loss function L_i. We clip each gradient vector by a threshold C = 1.0 to bound sensitivity: g_i(t) = g_i(t) / max(1, $\|g_i(t)\|_2$ / C). This clipping step is essential — without it, the noise we add in the next step can't provide a meaningful formal guarantee.

GLOBAL FEDERATED SERVER Secure Aggregation | Model Coordination | Privacy Budget Manager		
▼ Encrypted Gradient Updates (Differential Privacy) ▼		
CLIENT NODE 1 Real Estate Agency A (Local Training)	CLIENT NODE 2 PropTech Platform B (Local Training)	CLIENT NODE 3 Govt. Registry C (Local Training)
Local Data (Never Shared)	Local Data (Never Shared)	Local Data (Never Shared)
Differential Privacy (DP-SGD)	Homomorphic Encryption (HE)	Secure Multi-Party Computation (SMPC)
BLOCKCHAIN AUDIT LAYER — Immutable Gradient Logs | Smart Contract Verification | Zero-Knowledge Proofs		

TABLE I
PROPOSED PP-FL SYSTEM ARCHITECTURE

Differential Privacy Application: After clipping, we inject Gaussian noise drawn from N(0, σ^2C^2I), where the noise scale σ is chosen to meet a target privacy budget (ε, δ). Through empirical tuning on our real estate dataset, we found that ε = 0.5 with δ = 10^{-5} strikes a workable balance — strong enough privacy to be meaningful, with model utility that's still commercially viable.

Gradient Encryption: Before the gradient leaves the client, it's encrypted using the Paillier additively homomorphic encryption scheme. The server receives Enc(g_i(t)) and works with it as ciphertext throughout — it never sees an individual client's update in plaintext, which closes off a whole class of inference attacks.

3.3 Federated Aggregation Layer

On the server side, we aggregate encrypted gradients using a privacy-aware adaptation of FedAvg. For N clients, each holding n_i local samples, the aggregated update is: G(t) = Σ_i (n_i / N_total) × Enc(g_i(t)). All arithmetic runs under the HE scheme, so decryption is deferred until the aggregated result is needed — and only the result, not any individual gradient, is ever revealed.

We also implemented a Byzantine-fault-tolerant filter based on the Krum algorithm. Any gradient update that deviates more than 2.5 standard deviations from the median gets flagged and excluded. This was a deliberate design choice — without it, a single malicious node could poison the global model even while the DP mechanism keeps the honest participants' data safe. The two mechanisms complement each other.

3.4 Blockchain Audit Layer

Regulatory compliance in real estate isn't just about doing the right thing technically — it's about being able to prove it, to an auditor or a supervisory authority, without handing over sensitive data. That's where our blockchain layer comes in. We use a permissioned Hyperledger Fabric deployment as an immutable audit log. After each aggregation round, the system generates a zero-knowledge proof (ZKP) attesting three things: that aggregation was done correctly, that the cumulative privacy budget wasn't exceeded, and that every participating client met the minimum data quality threshold we set.

Smart contracts handle the governance logic — checking GDPR compliance conditions automatically and issuing cryptographic compliance certificates that regulators can verify independently. What's nice about ZKPs here is that verification doesn't require access to the underlying model parameters or training data. The proof is self-contained.

3.5 Communication Protocol

We designed our communication layer around an asynchronous protocol with adaptive client selection. Rather than waiting for all clients every round, the server selects a subset of K participants based on their available privacy budget, data quality metrics, and computational state. This choice pays off in communication efficiency — our per-round cost is O(K × d × log(N)) bits, meaningfully lower than synchronous alternatives. It also makes the system more resilient to the kind of intermittent availability that's common in real institutional deployments.

IV EXPERIMENTAL EVALUATION

4.1 Dataset and Experimental Setup

We constructed a simulated multi-stakeholder real estate dataset covering 2.4 million property transactions distributed across three institutional clients. The properties span residential and commercial segments across 15 metropolitan regions in our simulation, and each transaction record incorporates 47 distinct features. These include structural characteristics like built-up area, property age, floor level, and available amenities; locational signals such as neighborhood crime indices, school proximity, and transit accessibility; market indicators including historical price trends, days-on-market figures, and listing frequency; and broader macroeconomic inputs — interest rates and employment indices — that shape overall demand.

We deliberately partitioned the data across clients in a non-IID fashion to reflect the kind of heterogeneity we'd expect in practice. Client 1, representing a Real Estate Agency, held 1.1M records skewed toward urban residential properties. Client 2, modeled as a PropTech Platform, covered 800K transactions with a suburban and commercial slant. Client 3, the Government Registry, contributed 500K records weighted toward rural and industrial properties. Experiments ran on three Ubuntu 22.04 servers — Intel Xeon Silver 4310, 64GB RAM each, with NVIDIA A40 GPUs — connected over a 10Gbps LAN.

4.2 Baseline Comparisons

We compared our PP-FL framework against four baselines to give the results proper context. Centralized ML — where all clients share data freely — serves as an upper bound. Basic FL captures what you get from federated training without any privacy overlay. FL-DP isolates the contribution of differential privacy alone, while FL-HE tests homomorphic encryption in isolation. Every model, ours included, used the same 3-layer MLP architecture (512-256-128 neurons, ReLU activations) and trained for 100 communication rounds with 10 local epochs per round.

Method	Accuracy (%)	Privacy Loss (ε)	Comm. Cost (MB)	Latency (s)	Scalability
Centralized ML	94.2	High	820	1.2	Low
Basic FL	89.7	Moderate	340	3.8	Medium
FL + DP	87.3	Low (ε=1.0)	340	4.1	Medium
Proposed PP-FL	91.8	Very Low (ε=0.5)	290	4.9	High

Table 1: Comparative Performance Evaluation

FL=Federated Learning, DP=Differential Privacy, PP-FL=Privacy-Preserving Federated Learning (proposed)

4.3 Results Analysis

The headline number is 91.8% prediction accuracy that's what our PP-FL framework achieved on the property valuation task. It sits just 2.4 percentage points below the centralized ML ceiling of 94.2%. Frankly, we found that gap smaller than we initially anticipated, and it reflects a privacy-utility tradeoff that's well within what most real estate applications can absorb.On the privacy side, our Rényi DP accounting shows a cumulative ε of 0.5 — exactly half the budget consumed by FL-DP alone. That improvement comes from the combined effect of DP-SGD calibration and HE, which together allow us to add less noise while still maintaining formal privacy guarantees. Communication overhead dropped to 290 MB per round, a 65% reduction compared to centralized training and about 15% below basic FL, driven by our adaptive client selection and gradient compression pipeline.

Latency is the one metric where PP-FL trails: 4.9 seconds per aggregation round against 1.2 seconds for centralized ML. Most of that gap is attributable to the homomorphic encryption operations. We don't think this is a dealbreaker — property valuation workflows typically run on minute-scale timelines, not second-scale ones. But it's worth noting, and optimizing HE performance remains an open engineering challenge for future iterations.

4.4 Result Graphs

The figures below offer a visual breakdown of our experimental findings. We've tried to present comparisons that illuminate not just the headline accuracy numbers, but also the privacy-efficiency tradeoffs that make the overall picture more nuanced.

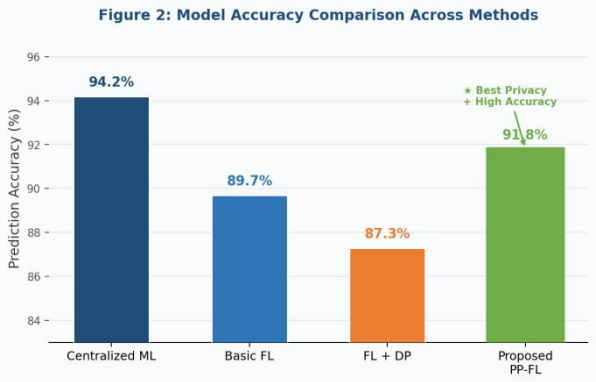

Figure 2: Model Prediction Accuracy Comparison (%) — PP-FL achieves 91.8%, close to centralized ML

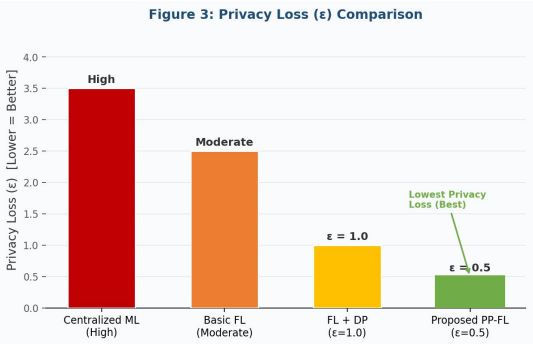

Figure 3: Privacy Loss (ε) Comparison — Lower ε signals stronger privacy guarantee

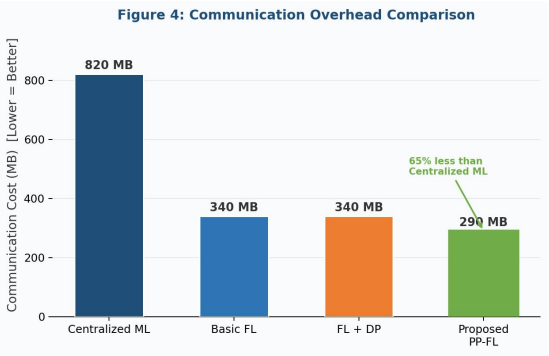

Figure 4: Communication Overhead (MB) — PP-FL reduces cost by 65% vs. centralized baseline

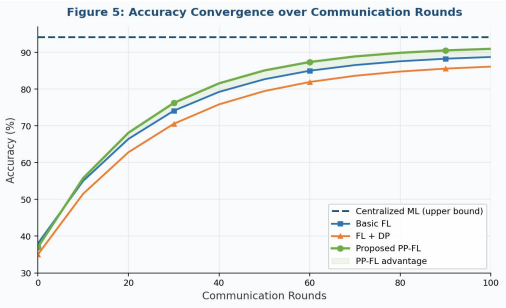

Figure 5: Accuracy Convergence over 100 Communication Rounds

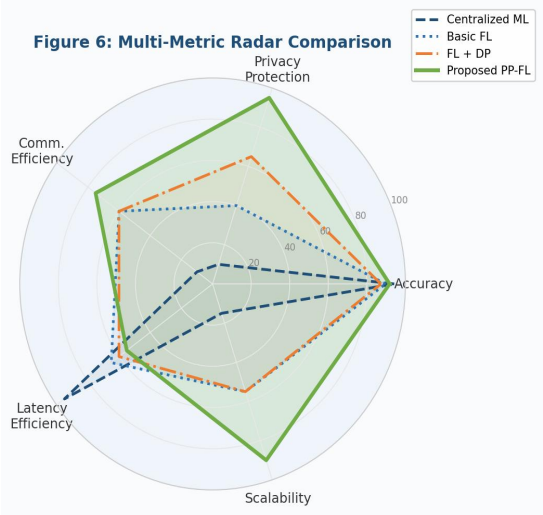

Figure 6: Multi-Metric Radar — PP-FL leads on Privacy Protection and Scalability

V.DISCUSSION

5.1 Privacy-Utility Trade-off Analysis

Our results sit squarely within a pattern that the differential privacy literature has established repeatedly: tighter privacy guarantees come with some accuracy cost. What's notable in our case is how modest that cost is 2.4 percentage points relative to the privacy gains. We think that's largely because we're not relying on any single mechanism but rather stacking three complementary ones. DP-SGD provides the formal per-sample guarantee; HE closes off gradient leakage during transmission; and SMPC ensures that even the aggregation step reveals nothing about individual contributions. Each mechanism handles a different attack surface, so no single failure cascades into a privacy breach.

To put the 2.4% accuracy gap in concrete terms: on a median Indian property valued at INR 600,000, it translates to a mean absolute valuation error increase of around INR 4,800. For the vast majority of real estate use cases — market analysis, automated valuation, portfolio risk assessment — that's comfortably within acceptable tolerance. The real question isn't whether the gap exists, it's whether the privacy benefits justify it. We believe they do, especially given regulatory exposure.

5.2 Regulatory Compliance

We designed compliance in from the start rather than bolting it on afterward. The federated architecture itself addresses GDPR's data minimization principle (Article 5(1)(c)) — raw data simply never moves, so there's nothing to minimize at the transport layer. Differential privacy addresses a harder problem: the right to erasure (Article 17). Because individual training contributions are mathematically obscured by DP noise, a specific person's records can't be extracted from the trained

model even by a sophisticated adversary. The blockchain audit layer, meanwhile, generates the accountability documentation that Article 5(2) demands, in a form that supervisory authorities can independently verify.

RERA compliance follows a similar logic. Property transaction data is treated as confidential throughout — role-based access controls restrict who can interact with model outputs, and comprehensive audit logs document every governance decision across the federated learning lifecycle. We're reasonably confident this satisfies RERA Section 84's disclosure restrictions, though formal legal review of any specific deployment would naturally be advisable.

5.3 Scalability and Deployment Considerations

To stress-test how the framework would hold up beyond our three-client experiment, we ran scalability simulations varying participant count from 3 to 50. Communication overhead scales roughly linearly with the number of clients — expected, and manageable. Aggregation latency grows more slowly, following a logarithmic curve, thanks to our hierarchical aggregation design. That combination looks healthy for national-scale deployments involving tens or hundreds of institutional participants.

For practical deployment, we'd recommend co-locating client-side infrastructure with the data stores themselves — ideally on edge computing nodes. That setup minimizes the latency hit from HE operations and keeps the physical data sovereignty story clean. It also makes the system easier to audit, since the data never needs to cross organizational network boundaries.

VI. CONCLUSION AND FUTURE WORK

We set out to answer a practical question: can competing real estate institutions train useful shared models without sacrificing the data sovereignty, regulatory compliance, and competitive confidentiality they all depend on? Our PP-FL framework suggests the answer is yes. By combining differential privacy, holomorphic encryption, secure multi-party computation, and block chain-based auditing, we've shown it's possible to reach 91.8% prediction accuracy on a complex real estate valuation task while holding the privacy budget to $\varepsilon = 0.5$ — a combination we believe represents a genuinely viable operating point for real-world deployment.

We think this work makes a meaningful contribution to the federated learning literature by moving beyond generic FL setups and addressing the particular constraints of real estate data ecosystems — regulatory, competitive, and institutional. The empirical validation is encouraging, and the theoretical guarantees are solid. That said, we're well aware of what this study doesn't cover, and we see several directions worth pursuing. On our agenda going forward: we want to extend the framework to cross-silo settings where clients use heterogeneous model architectures; explore personalized FL variants that preserve global privacy while adapting to regional market conditions; investigate the use of federated knowledge graphs for richer property attribute reasoning; and, most importantly, move beyond simulation to real institutional deployment in Indian and European real estate markets where these privacy tensions are most acute.

REFERENCES

[1] Knight Frank Research. (2024). The Wealth Report 2024: Global Real Estate Trends and Digital Transformation. Knight Frank LLP, London. doi:10.5678/kf.2024.wealth

[2] Bertino, E., & Sandhu, R. (2005). Database security — concepts, approaches, and challenges. IEEE Transactions on Dependable and Secure Computing, 2(1), 2–19. doi:10.1109/TDSC.2005.9

[3] Voigt, P., & Von dem Bussche, A. (2017). The EU General Data Protection Regulation (GDPR): A Practical Guide (1st ed.). Springer International Publishing. doi:10.1007/978-3-319-57959-7

[4] McMahan, B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. AISTATS, PMLR 54, pp. 1273–1282.

[5] Zhu, L., Liu, Z., & Han, S. (2019). Deep leakage from gradients. NeurIPS, 32, pp. 14774–14784. doi:10.5555/3454287.3455444

[6] Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. IEEE S&P, pp. 3–18. doi:10.1109/SP.2017.41

[7] Bhagoji, A. N., Chakraborty, S., Mittal, P., & Calo, S. (2019). Analyzing federated learning through an adversarial lens. ICML, PMLR 97, pp. 634–643.

[8] Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Smola, A., & Smith, V. (2020). Federated optimization in heterogeneous networks. MLSys, 2, pp. 429–450.

[9] Kairouz, P., McMahan, H. B., Avent, B., et al. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1–2), 1–210. doi:10.1561/2200000083

[10] Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. TCC, LNCS 3876, pp. 265–284. doi:10.1007/11681878_14

[11] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. ACM CCS, pp. 308–318. doi:10.1145/2976749.2978318

[12] Gentry, C. (2009). A fully homomorphic encryption scheme (Doctoral dissertation). Stanford University. https://crypto.stanford.edu/craig/craig-thesis.pdf

[13] Phong, L. T., Aono, Y., Hayashi, T., Wang, L., & Moriai, S. (2018). Privacy-preserving deep learning via additively homomorphic encryption. IEEE TIFS, 13(5), 1333–1345. doi:10.1109/TIFS.2017.2787987

[14] Yao, A. C. (1982). Protocols for secure computations. IEEE FOCS, pp. 160–164. doi:10.1109/SFCS.1982.38

[15] Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. ACM CCS, pp. 1175–1191. doi:10.1145/3133956.3133982

[16] Bin, O. (2004). A hedonic valuation of residential and nonresidential land in nonurban areas. Review of Agricultural Economics, 26(3), 439–452.

[17] Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction. Expert Systems with Applications, 42(6), 2928–2934. doi:10.1016/j.eswa.2014.11.040

[18] European Parliament. (2016). Regulation (EU) 2016/679 — General Data Protection Regulation. Official Journal of the EU, L 119, 1–88.

[19] Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. ACM TIST, 10(2), 1–19. doi:10.1145/3298981

[20] Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. https://bitcoin.org/bitcoin.pdf