ISSN: 2581-3404 (Online) IF: 5.445 (SJIF)
IJIRTM, Volume-8, Issue-1, February-2024.

Federated Learning for Privacy-Preserving AI

Gulabpreet Singh[1], Sabhyata Kalra[2], Utkarsh[3], Karan[4], Soumya Tripathi[5]
Department of Computer Science and Engineering (AI & ML) [1,2,3,4,5]
Panipat Institute of Engineering and Technology (PIET), Samalkha, Panipat, Haryana, India[1,2,3,4,5]
2820260.cse@pietgroup.co.in[1], 2820251.cse@pietgroup.co.in[2], 2820262.cse@pietgroup.co.in[3],
2820292.cse@pietgroup.co.in[4], soumya.cse-aiml@piet.co.in[5]

Abstract: *Modern machine learning models are typically trained by centralizing data on a single server, an approach that conflicts with growing privacy regulation and with the practical reality that much valuable data, patient records, financial transactions, on-device usage logs, cannot or should not leave the organization or device where it was generated. Federated Learning (FL), introduced by McMahan et al. as the Federated Averaging algorithm, addresses this by training a shared model across many decentralized clients, each of which computes an update using its own local data, sending only that update, not the raw data, to a central server for aggregation. This paper reviews the architecture, privacy-enhancing techniques, and application landscape of federated learning. We describe the core FedAvg algorithm and the cross-device versus cross-silo deployment settings, then survey the privacy and security layer built on top of FL, including differential privacy, secure aggregation, and homomorphic encryption, each of which defends against a different threat to the base protocol's privacy guarantees. We then review applications across four domains: healthcare, where FL enables multi-institutional model training without sharing patient records; Internet-of-Things (IoT) and smart devices, where FL enables on-device personalization; and banking and finance, where FL supports collaborative fraud detection and credit risk modeling across institutions that cannot share raw customer data. We close with a discussion of open challenges, including statistical heterogeneity across clients, communication efficiency, robustness to malicious participants, and regulatory alignment, and outline directions for future research.*

Keywords: - Federated learning, Privacy-preserving AI, Differential privacy, Secure aggregation, Healthcare AI, Internet of Things, Financial fraud detection.

Introduction

The default recipe for training a machine learning model, collect data from many sources onto a central server, then train on the pooled dataset, is increasingly at odds with both regulation and practical constraints. Data-protection regulation restricts how personal data such as health records or financial transactions can be moved and combined across organizational boundaries. Even where legally permissible, centralizing data from millions of personal devices is often impractical: the aggregate volume is enormous, transferring it consumes bandwidth and battery, and centralization itself creates a single high-value target for attackers. Federated Learning (FL) was proposed as a way to train a shared model without centralizing the underlying data at all [1].

In the Federated Averaging (FedAvg) algorithm introduced by McMahan et al., a central server distributes the current global model to a subset of participating clients; each client performs several steps of local training on its

own private data; clients send back only the resulting model update (not the raw data); and the server averages these updates, weighted by the amount of local data, to produce an improved global model, repeating this cycle over many communication rounds [1]. Because raw data never leaves the device or institution where it originated, FL directly addresses the data-minimization principle at the center of most privacy regulation, while still allowing the model to benefit from the collective, decentralized dataset.

This paper reviews the FL landscape as it has developed since FedAvg. Section II covers the core algorithm and deployment settings. Section III surveys the privacy and security techniques layered on top of the base protocol. Sections IV through VI cover applications in healthcare, IoT and smart devices, and banking and finance respectively. Section VII discusses open challenges, Section VIII future directions, and Section IX concludes.

Core Architecture and Deployment Settings

A. Federated Averaging

FedAvg trains a single global model through iterative rounds of local computation and central aggregation. In each round, the server selects a subset of available clients and sends them the current global model weights; each selected client runs several epochs of stochastic gradient descent on its local dataset starting from those weights; and the server collects the resulting local weight updates and combines them via a weighted average, where each client's contribution is weighted by the size of its local dataset [1]. Running multiple local epochs before each communication round, rather than a single gradient step, was the key innovation that made the algorithm communication-efficient enough to be practical over the bandwidth-constrained, unreliable connections typical of mobile devices [1].

B. Cross-Device and Cross-Silo Settings

FL deployments are typically classified into two settings with distinct operational characteristics. Cross-device FL involves a very large number, potentially millions, of unreliable clients such as smartphones, each holding a small amount of local data and participating intermittently as availability and network conditions permit; this is the original setting FedAvg was designed for, exemplified by on-device keyboard prediction. Cross-silo FL instead involves a small number, typically tens to hundreds, of reliable, well-resourced clients, such as hospitals or banks, each holding a comparatively large local dataset and participating in every round; this setting dominates healthcare and financial applications, where the participants are institutions rather than individual consumer devices [2]. The two settings favor different design choices: cross-device FL must tolerate client dropout and highly non-identically-distributed (non-IID) data across millions of small, similar devices, while cross-silo FL can assume reliable participation but must contend with a small number of institutions whose data distributions may differ substantially from one another.

A third, less common setting, federated transfer learning, arises when participating clients share neither the same feature space nor the same sample space, for example a bank and an e-commerce platform that observe different customers through different features, and combine transfer-learning techniques with the federated protocol to align representations across parties before collaborative training can proceed. Vertical federated learning is a related variant in which different clients hold different features for the same set of entities, such as a bank and an insurer that each hold partial records on shared customers, requiring entity-alignment and secure-computation techniques distinct from the horizontal, same-feature-space setting FedAvg was originally designed for.

Client-server communication topology is itself an active design choice: while the original FedAvg protocol assumes a single central aggregating server, fully decentralized variants replace this with peer-to-peer gossip-style aggregation among clients, removing the single point of failure and potential trust bottleneck a central server represents, at the cost of slower convergence and more complex coordination logic among participants.

Privacy and Security Techniques

Sharing only model updates rather than raw data is a substantial privacy improvement over centralized training, but it is not sufficient on its own: model updates can still leak information about the underlying training data, and several attacks specifically target this residual leakage [2], [3]. Membership-inference attacks attempt to determine whether a specific data point was used to train a model by observing differences in the model's behavior on that point versus unseen data, while gradient-inversion and feature-leakage attacks attempt to reconstruct approximate training examples directly from a client's shared gradient update. FL systems therefore typically layer one or more additional privacy mechanisms on top of the base FedAvg protocol.

A. Differential Privacy

Differential privacy (DP) provides a formal, quantifiable privacy guarantee by adding calibrated random noise to a client's update (local DP) or to the server's aggregate (central DP) before it is used or released, bounding how much any single data point can influence the final model and thereby limiting what an observer can infer about any individual's contribution. McMahan et al. demonstrated learning differentially private recurrent language models within the FL setting, showing that meaningful privacy guarantees could be achieved with an acceptable, though non-trivial, cost to model accuracy [3]. The central design tension in applying DP to FL is the privacy-utility trade-off: stronger privacy guarantees require more noise, which in turn degrades model accuracy, and tuning this trade-off appropriately for a given application remains an active area of research.

B. Secure Aggregation

Secure aggregation is a cryptographic protocol that allows the central server to compute the sum (or average) of client updates without ever seeing any individual client's update in the clear, protecting against an honest-but-curious server that might otherwise attempt to extract information from individual updates before they are averaged [2], [3]. Bonawitz et al. introduced a practical secure-aggregation protocol for FL that uses pairwise masking so that individual client masks cancel out exactly when updates are summed across all participating clients, while remaining robust to some clients dropping out mid-protocol, a common occurrence in cross-device deployments [3].

C. Homomorphic Encryption and Secure Multi-Party Computation

Homomorphic encryption allows computation, such as the weighted averaging step in FedAvg, to be performed directly on encrypted updates without ever decrypting them, giving the strongest cryptographic protection against a curious server at the cost of substantially higher computational overhead than secure aggregation. Secure multi-party computation (SMPC) protocols distribute a computation across several non-colluding parties such that no single party observes the full input, and have been applied specifically to privacy-preserving financial anomaly detection across institutions that are unwilling to reveal raw transaction data to one another even indirectly. In practice, healthcare- and finance-focused FL deployments frequently combine several of these mechanisms, differential privacy, secure aggregation, and encryption, layered together to satisfy the specific regulatory and threat-model requirements of the domain.

Table I summarizes the primary threat addressed by each privacy mechanism discussed above.

Table 1: Fl Privacy Mechanisms and the Threats They Address

Mechanism	**Threat Addressed**	**Main Cost**
Differential Privacy	Inference of individual data from model/updates	Reduced accuracy
Secure Aggregation	Curious server inspecting individual updates	Extra communication rounds
Homomorphic Encryption	Curious server, strongest guarantee	High computational overhead

Healthcare Applications

Healthcare is one of the most active application domains for FL because patient data is both extremely sensitive and, in most jurisdictions, legally restricted from leaving the institution that collected it, while at the same time individual hospitals rarely have enough data on rare conditions to train an accurate model alone. Sheller et al. demonstrated the feasibility of multi-institutional deep learning for brain-tumor segmentation without sharing patient data directly, training a shared segmentation model across institutions via FL and showing performance competitive with a hypothetical model trained on centrally pooled data [4]. Beyond imaging, FL has been applied to electronic health record analysis for risk prediction across hospital networks and to federated genomic and drug-discovery pipelines, where FL allows pharmaceutical and research partners to jointly train models on genomic data that data-sharing agreements would otherwise prohibit from being pooled. Healthcare FL deployments are consistently in the cross-silo setting, since participants are hospitals or research institutions rather than individual patient devices, and typically combine FL with differential privacy or secure aggregation given the exceptionally sensitive nature of medical data [2].

IoT and Smart Devices

The cross-device FL setting was originally motivated by, and remains dominant in, consumer IoT and smart-device applications, where millions of phones, wearables, and smart-home devices each hold a small amount of privacy-sensitive usage data [1], [2]. On-device keyboard next-word prediction was among the first production deployments of FL, allowing a shared language model to improve from the typing patterns of a large user population without any individual's keystrokes leaving their device. Wearable and smart-home applications extend this pattern to health and activity monitoring, where a federated model can learn to recognize patterns such as falls or irregular heart rhythms across a population of devices without centralizing continuous biometric data from any individual user. Because IoT clients are numerous, resource-constrained, and often intermittently connected, IoT deployments place particular emphasis on communication-efficient variants of FedAvg, such as gradient compression and quantization, that reduce the size of each update transmitted per round [1].

Banking And Finance

In banking and finance, FL enables collaboration across institutions that are simultaneously competitors and are subject to strict regulatory limits on data sharing, particularly for fraud detection and anti-money-laundering (AML), where a single bank typically sees only a narrow slice of a fraud pattern that spans multiple institutions [5], [6]. A federated fraud-detection model allows each bank to train on its own transaction data locally while contributing to, and benefiting from, a shared global model that captures fraud patterns invisible to any single institution acting alone [6]. FL has also been applied to collaborative credit-risk scoring, where multiple lenders jointly improve a creditworthiness model without revealing individual customers' financial details to one another, and to cross-institution AML detection, where the ability to recognize money-laundering chains that route through several banks in sequence is fundamentally limited without some form of cross-institution learning [5]. A recent survey proposes classifying financial FL applications by regulatory exposure level, from lower-exposure tasks such as collaborative portfolio optimization to high-exposure tasks such as real-time fraud detection, noting that higher-exposure applications require correspondingly stronger privacy guarantees and more careful regulatory alignment [5].

Open Challenges

Several challenges recur across FL deployments regardless of domain. Statistical heterogeneity, non-IID data across clients, is perhaps the most fundamental: FedAvg's convergence guarantees were derived under assumptions that do not hold when different clients' data distributions differ substantially, and in practice this heterogeneity can slow convergence or bias the global model toward clients with larger or more typical datasets

[1], [2]. Communication efficiency remains a practical bottleneck, particularly in cross-device settings with millions of bandwidth-constrained participants, motivating gradient compression, quantization, and knowledge-distillation-based alternatives to transmitting full model updates every round. Robustness to malicious participants is a distinct concern: because the server cannot directly inspect a client's local data or training process, a malicious client can attempt data-poisoning or model-poisoning attacks that degrade or backdoor the shared global model, and defending against such attacks without violating the privacy guarantees that motivated FL in the first place remains an open research problem [2]. Byzantine-robust aggregation rules, which replace simple weighted averaging with statistically robust alternatives such as coordinate-wise median or trimmed mean, offer partial protection against a bounded fraction of malicious clients, though typically at some cost to convergence speed even when no attack is present. Finally, regulatory alignment is non-trivial in practice: even though FL avoids centralizing raw data, model updates themselves may still be considered personal data under some interpretations of regulations such as the GDPR, and financial and healthcare deployments in particular must carefully document how their specific combination of FL and privacy-enhancing techniques satisfies applicable regulatory requirements [5].

Evaluation itself is harder in FL than in centralized learning: because no single party observes the full training dataset, standard practices such as held-out validation must be adapted to a federated setting, and fairness across clients, ensuring the shared global model performs acceptably for every participant rather than only on average, is a further evaluation dimension with no universally agreed metric. Client selection strategy, which subset of available clients participates in a given round, also materially affects both convergence speed and fairness, since naively selecting clients uniformly at random can systematically under-represent clients with less reliable connectivity or less typical data.

Future Research Directions

Personalization is an active direction: rather than training a single global model that must serve all clients equally well despite their differing data distributions, personalized FL methods adapt the shared model to each client's local distribution through techniques such as local fine-tuning, multi-task formulations, or meta-learning, aiming to give every client a model at least as good as one it could have trained alone. Improving robustness to poisoning attacks while preserving formal privacy guarantees, rather than treating robustness and privacy as separable problems solved independently, is likely to remain an active research area given the tension between the two. Standardized benchmarks and evaluation protocols, extending efforts such as the LEAF benchmark for federated settings, would make it easier to compare competing FL algorithms on a like-for-like basis across the highly heterogeneous conditions real deployments face. Finally, closer collaboration between FL researchers and regulators to establish clear, technically grounded standards for what combination of FL and privacy-enhancing techniques satisfies a given regulatory bar would reduce the current uncertainty that slows adoption in the most privacy-sensitive domains such as healthcare and finance.

Conclusion

Federated learning reframes distributed machine learning around a simple but consequential change: instead of moving data to a central model, it moves the model to the data, sharing only model updates rather than raw records [1]. This single change, combined with complementary privacy mechanisms such as differential privacy, secure aggregation, and homomorphic encryption, has made it practical to train shared models across hospitals that cannot pool patient records, banks that cannot pool transaction data, and millions of consumer devices that should not upload personal usage logs to a central server. Realizing FL's full potential will require continued progress on statistical heterogeneity, communication efficiency, robustness to malicious participants, and

regulatory clarity, but its core proposition, collaborative model improvement without centralized data collection, is increasingly central to how privacy-sensitive AI systems are built across healthcare, IoT, and finance alike.

As foundation models and large language models increasingly require fine-tuning on private, domain-specific data, federated approaches to LLM adaptation are also emerging as a natural extension of the ideas surveyed here: parameter-efficient federated fine-tuning allows multiple institutions to jointly adapt a shared base model to their combined domain, such as clinical or financial language, without any institution's proprietary text ever leaving its own infrastructure, suggesting that the federated paradigm first developed for smaller supervised models is likely to remain relevant as the scale and reach of AI systems continue to grow.

References

[1] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in Artificial Intelligence and Statistics (AISTATS), pp. 1273-1282, 2017.

[2] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, et al., "Advances and Open Problems in Federated Learning," Foundations and Trends in Machine Learning, vol. 14, no. 1-2, pp. 1-210, 2021 (arXiv:1912.04977).

[3] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical Secure Aggregation for Privacy-Preserving Machine Learning," in Proc. ACM SIGSAC Conf. on Computer and Communications Security, pp. 1175-1191, 2017.

[4] Deepak Rathore, Anurag Jain, “Design Hybrid method for intrusion detection using Ensemble cluster classification and SOM network”, International Journal of Advanced Computer Research, 2012, pp. 181-186.

[5] M. J. Sheller, G. A. Reina, B. Edwards, J. Martin, and S. Bakas, "Multi-Institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation," Brain Imaging and Behavior, vol. 14, no. 6, pp. 1681-1691, 2020.

[6] Jay Prakash Maurya, DK Rathore, S Joshi, M Manoria, V Richhariya, “Corporate Sector Fraud: Challenges and Safety”, Machine Learning Applications for Accounting Disclosure and Fraud Detection, 2021, pp. 16-31.

[7] R Dubey, “An empirical study of intrusion detection system using feature reduction based on evolutionary algorithms and swarm intelligence methods”, International Journal of Applied Engineering Research, 2017. pp. 8884-8889.

[8] "A survey on federated learning applications in healthcare, finance, and data privacy/data security," 2023.

[9] "A Comprehensive Survey of Federated Transfer Learning: Challenges, Methods and Applications," arXiv:2403.01387, 2023.