Vol. 16 (2026): Proceedings of the 2026 International Conference on AI and Technological Innovation Applications (AITIA 2026)
E-ISSN: 3105-0522 P-ISSN: 3105-0514
DOI: https://doi.org/10.70267/aitia.2026346352

Federated Learning as a Privacy-Enhancing Framework: Risks, Protection Mechanisms, and Applications

Zhiyue Zhang*

University of Dayton, Ohio, United States of America

**Corresponding author: Zhiyue Zhang.*

Abstract

Federated learning has become a major approach for training artificial intelligence systems when data is distributed across institutions, devices, or users. Its central appeal is that raw data can remain local while model updates are coordinated through a shared training process. This paper argues that federated learning should be evaluated as a privacy-enhancing framework rather than as a complete privacy guarantee. It first identifies key privacy risks, including model-update leakage, gradient inversion, membership inference, and risks created by malicious participants. It then classifies protection mechanisms into architectural choices, optimization design, differential privacy, secure aggregation, and deployment constraints. The paper proposes a comparative framework that asks what federated learning protects, what it still exposes, and which auxiliary mechanisms are needed in different application settings. Healthcare AI and mobile keyboard prediction are used as contrasting case studies: the former is typically cross-silo and institutionally governed, while the latter is cross-device and large-scale. The analysis concludes that federated learning is most effective when combined with explicit threat models, layered privacy protections, security controls, and governance arrangements.

Keywords

federated learning, privacy-preserving artificial intelligence, differential privacy, secure aggregation, healthcare AI, mobile keyboard prediction

1. Introduction

Artificial intelligence now depends on data from many parts of daily life. Search engines, mobile devices, hospitals, banks, and online platforms all generate records that can improve machine learning models. The same data, however, may contain sensitive information about users, patients, employees, or organizations. Traditional centralized machine learning collects data in one location before training. That design is convenient, but it also creates a large repository of raw information that can be misused, breached, or repurposed beyond the original context.

Federated learning offers a different training architecture. Instead of moving raw data to a central server, clients such as mobile phones, hospitals, or financial institutions train a shared model locally and send model

updates to a coordinator. Federated Averaging (FedAvg), proposed by McMahan et al., became the standard baseline because it reduced the communication burden of distributed training by allowing clients to perform multiple local updates before aggregation [1]. Subsequent surveys and taxonomies have shown that federated learning is now a broad research area encompassing challenges in optimization, privacy, security, systems, and governance [2, 3].

A simple description of federated learning can nevertheless be misleading. Keeping raw data local does not automatically guarantee privacy. Gradients and model updates may still reveal information about training examples; a malicious client may poison the training set; a curious server may infer properties of participants; and deployment constraints may force design choices that weaken either utility or privacy. Therefore, federated learning should not be treated as a finished privacy solution. It is better understood as a privacy-enhancing architecture whose effectiveness depends on threat models, aggregation protocols, differential privacy guarantees, and the deployment environment.

Existing reviews have already explained the basic algorithms and open problems of federated learning. The contribution of this paper is narrower and more comparative. It asks how federated learning changes the privacy problem, which risks remain after data are kept local, how differential privacy and secure aggregation complement each other, and why application contexts such as healthcare and mobile keyboard prediction require different assumptions. The paper is organized around a layered framework: architecture, optimization, privacy risks, protection mechanisms, deployment constraints, and application-specific governance.

2. Foundations of Federated Learning

2.1 Federated Learning Architectures

Federated learning is a distributed machine learning approach in which multiple clients train a shared model while keeping raw data decentralized. The clients may be smartphones, hospitals, banks, factories, or other data holders. A central server usually coordinates the process by selecting clients, sending the current model, receiving updates, and aggregating those updates into a new global model.

Two architectural distinctions are especially important. The first is cross-device versus cross-silo federated learning. Cross-device learning involves a very large number of personal devices, such as phones, where clients may be unreliable, intermittently connected, and individually small. Cross-silo learning involves a smaller number of organizations, such as hospitals or banks, where each participant may hold substantial data and may be governed by formal agreements. The second distinction is horizontal versus vertical federated learning. Horizontal federated learning assumes that clients share a similar feature space but hold different users or records; vertical federated learning assumes that clients hold different features about overlapping users [3]. These distinctions matter because they shape privacy risk, system design, and application feasibility.

2.2 Optimization Foundations: FedAvg and Non-IID Data

FedAvg is the most common starting point for explaining federated optimization. In each communication round, the server sends the current global model to a sampled group of clients. Each selected client trains locally for several steps and sends back an updated model. The server then computes a weighted average of those local models, giving each client a weight proportional to the amount of local training data it used. In plain terms, the next global model is the data-size-weighted average of the participating clients' locally updated models.

This differs from a simple federated stochastic gradient descent setting, where clients may send gradients after fewer local steps. FedAvg reduces communication by allowing more local computation before synchronization. That benefit is valuable because communication is often more expensive than local computation in federated settings [1]. The cost is that local updates may drift apart when the data are non-IID. One hospital may treat a different patient population from another hospital, and one phone user may type in a different language or style from another user. This heterogeneity creates challenges for convergence and fairness; theoretical analyses of FedAvg therefore emphasize the relationships among local steps, client sampling, learning rates, and non-IID distributions [4].

2.3 Privacy Assumptions: What FL Protects and What It Does Not Protect

The privacy benefit of federated learning begins with data minimization: raw training data need not be copied into a central database. This reduces the exposure associated with large centralized repositories and can enable collaboration when direct data sharing is legally or institutionally difficult.

However, federated learning does not protect everything. Model updates are derived from local data, and therefore may leak information about that data. The server may see metadata about participation, timing, device behavior, or institutional availability. The final global model may memorize unusual examples. A privacy analysis must therefore ask three questions: what information leaves the client, who can observe it, and how much influence a single user or client can have on the final model.

3. Privacy Risks and Protection Mechanisms

3.1 Privacy Risks: Update Leakage, Gradient Inversion, and Membership Inference

The first risk is model-update leakage. Even when raw data remain local, gradients or parameter updates may encode information about the examples used to compute them. Zhu et al. showed that training data can sometimes be reconstructed from shared gradients in the Deep Leakage from Gradients attack [5]. Geiping et al. further demonstrated that gradient inversion can be surprisingly effective under realistic assumptions [6]. These results are important because they show that the difference between sharing data and sharing updates is not always as large as it appears.

A second risk is membership inference. In membership inference attacks, an adversary tries to determine whether a particular record was included in a training set. Shokri et al. showed that machine learning models can leak membership information through their outputs [7]. In federated learning, this concern can apply to both the final model and intermediate updates, especially when clients are small groups or individual users.

A third risk comes from adversarial behavior inside the training process. A malicious server may attempt to isolate or manipulate client updates, and a malicious client may send poisoned updates or backdoors. These are partly security problems, but they also affect privacy because attacks on the training protocol can create new opportunities for inference.

3.2 Differential Privacy

Differential privacy provides a formal way to limit the effect of a single individual record or client on a computation. The classical formulation adds calibrated noise so that the output of an analysis changes by at most a bounded amount when an individual is added or removed [8, 9]. In federated learning, differential privacy can be applied by clipping updates and adding noise before aggregation. Client-level differential privacy treats an entire client as the protected unit, which is important when a phone or institution may hold many related examples [10].

The central trade-off is between privacy and utility. Stronger noise can reduce the influence of individual data, but it may also reduce model accuracy. This trade-off is especially visible in language modeling, where rare words or personal phrases may be useful for prediction but also sensitive. Work on differentially private recurrent language models shows how privacy accounting, clipping, and noise interact with model quality [11].

3.3 Secure Aggregation

Secure aggregation addresses a different part of the privacy problem. It is designed so that the server can compute an aggregate update without seeing each client update. Bonawitz et al. proposed a practical secure aggregation protocol for privacy-preserving machine learning that remains usable even when some clients drop out [12]. This is important in cross-device settings, where phones may disconnect or stop participating during a training round.

Secure aggregation and differential privacy are complementary. Secure aggregation protects individual updates from the server; differential privacy limits the identifiable influence of an individual on the result. Secure aggregation alone does not guarantee that the final aggregate or model is private. Differential privacy alone may still require trust in the party that clips and adds noise. A layered design can therefore use secure

aggregation to hide individual updates and differential privacy to reduce what can be inferred from the aggregate output.

3.4 Layered Protection Framework

The mechanisms above are best understood as layers. Federated architecture reduces raw-data movement. Secure aggregation reduces server visibility into individual updates. Differential privacy limits individual influence on the training result. Security controls defend against poisoning, backdoors, and malicious participants. Governance mechanisms define who may participate, what data are eligible, how consent is handled, and who is accountable if a model fails. No single layer is sufficient by itself. The strength of a federated system depends on whether the layers match the actual threat model.

4. Deployment Constraints and System-Level Challenges

Communication efficiency should be treated as a deployment constraint rather than as a privacy mechanism. Federated learning can require many rounds of communication between clients and the server. In cross-device systems, clients may be offline, battery-limited, or connected only under certain conditions. In cross-silo systems, communication may be more stable, but coordination, auditability, and institutional approval may be harder to achieve.

Large-scale federated systems, therefore, rely on client selection, scheduling, fault tolerance, compression, and aggregation protocols that can handle dropout. Google's production-oriented work on federated learning at scale emphasizes that practical deployment is as much a systems problem as an algorithmic one [13]. These constraints matter for privacy because a privacy-preserving method that is too slow, unstable, or costly may not be adopted. They also matter for fairness: if only powerful devices or large institutions can participate, the resulting model may underrepresent weaker clients.

5. Application Case Studies

5.1 Healthcare AI: Cross-Silo Federated Learning

Healthcare is one of the clearest areas where federated learning appears valuable. Hospitals, clinics, and laboratories hold data that could improve diagnostic models, medical imaging systems, risk prediction tools, and electronic health record models. At the same time, patient data are highly sensitive, regulated, and institutionally controlled. Directly pooling patient records across organizations can be legally difficult and ethically controversial.

Federated learning enables institutions to collaborate without transferring raw patient records. Rieke et al. describe federated learning as a promising direction for digital health because it can support multi-institutional model development while reducing the need for direct data sharing [14]. Concrete examples include medical image segmentation, multi-site diagnostic modeling, and electronic health record prediction. Sheller et al. showed the feasibility of multi-institutional deep learning for brain tumor segmentation without sharing patient data [15]. Dayan et al. used federated learning to predict clinical outcomes for COVID-19 patients across multiple institutions [16].

The healthcare case also shows the limits of federated learning. Hospitals may use different imaging machines, coding systems, label definitions, and clinical workflows. These differences create non-IID data and can cause a global model to work better for some institutions than for others. Clinical responsibility also remains unresolved: if a federated model makes a harmful recommendation, responsibility may be shared among developers, hospitals, data providers, and governance bodies. In healthcare, privacy protection is therefore inseparable from validation, fairness, auditability, and clinical accountability.

5.2 Mobile Keyboard Prediction: Cross-Device Federated Learning

Mobile keyboard prediction is a contrasting application because it is usually cross-device rather than cross-silo. Typed text can reveal names, locations, relationships, habits, and private messages. A central collection

of raw typed text would create serious privacy concerns. Federated learning allows a keyboard model to improve from user behavior while keeping raw text on the device.

In this setting, the system normally selects eligible devices only when they meet conditions such as being idle, charging, and connected to Wi-Fi. Each device trains locally on recent user interactions and sends an update rather than raw text. Work on mobile keyboard prediction demonstrates how federated learning can be used for next-word prediction and related language tasks at a large scale [17]. Differential privacy can reduce the risk of one user's phrases being memorized, while secure aggregation can reduce servers' visibility into individual updates [11, 12].

This case differs from healthcare in three ways. First, it involves many small and unreliable clients rather than a smaller set of institutions. Second, each client may hold highly personal but relatively small local data. Third, the system must operate under strict device constraints. As a result, cross-device keyboard learning depends heavily on client sampling, dropout handling, communication efficiency, differential privacy, and secure aggregation.

5.3 Comparison of the Two Cases

The two case studies show that federated learning cannot be evaluated as a single uniform design. Healthcare AI is usually closer to cross-silo federated learning: the number of participants is smaller, each institution may hold a large and clinically meaningful dataset, and institutional agreements, regulations, validation standards, and clinical accountability shape participation. Mobile keyboard prediction is usually closer to cross-device federated learning: the number of clients is much larger, each client contributes a smaller amount of highly personal data, and system design must handle intermittent availability, battery limits, Wi-Fi constraints, dropout, and device heterogeneity.

The privacy assumptions also differ. In healthcare, the central concern is not only whether raw medical records leave the hospital, but also whether the resulting model is clinically valid across institutions and whether responsibility is clear if the system fails. Secure aggregation and differential privacy may help, but they do not replace institutional governance, auditability, or clinical review. In mobile keyboard prediction, the primary concern is protecting individual users at scale. Client-level differential privacy and secure aggregation are especially important because typed text can reveal intimate personal information and because the server may otherwise observe updates from many individual devices.

These differences support the main argument of this paper. Federated learning is most effective when the architecture, privacy mechanisms, and deployment assumptions are aligned with the application. Healthcare requires strong governance and cross-institutional validation; mobile keyboard prediction requires scalable protocols, client-level protection, and robust handling of unreliable devices. In both cases, federated learning reduces raw-data centralization, but it becomes a credible privacy-enhancing framework only when combined with mechanisms and governance appropriate to the setting.

6. Challenges and Future Directions

Privacy challenges should be separated from security challenges. Privacy risks involve what can be inferred about training data, clients, or users from updates, aggregates, or model outputs. Future work should improve privacy accounting, client-level differential privacy, auditing of update leakage, and evaluation methods that test whether deployed systems actually protect sensitive information.

Security risks involve malicious behavior in the training process. Federated systems may include clients that send poisoned updates, attempt to insert backdoors, or manipulate aggregation. A malicious or curious server may also try to isolate clients or alter the protocol. Defenses may require robust aggregation, anomaly detection, client authentication, secure aggregation, and monitoring of model behavior.

System challenges involve communication cost, non-IID data, client dropout, device heterogeneity, and convergence stability. Future systems must reduce communication without excluding weaker clients or worsening model fairness. Personalization- and fairness-aware federated optimization are especially important when a single global model cannot serve every client group equally.

Governance challenges remain even when technical privacy protections are strong. Users and institutions need to know what data are involved, how updates are protected, who controls the model, and what happens when the system fails. Consent, transparency, accountability, and regulation are not optional additions to federated learning; they are part of what makes a privacy-enhancing framework trustworthy in practice.

7. Conclusion

Federated learning is an important approach to privacy-preserving artificial intelligence because it changes the assumption that useful models require centralized raw data. Keeping data on local devices or within local institutions reduces one of the most direct sources of privacy risk. This is especially valuable in areas such as healthcare and mobile text prediction, where data are both useful and sensitive.

At the same time, federated learning is not a complete privacy solution. Model updates, gradients, and final model outputs can still reveal information. Differential privacy can limit individual influence, secure aggregation can hide individual updates from the server, and system-level design can make deployment practical. These mechanisms work best as layers rather than substitutes for one another.

The strongest conclusion is therefore conditional. Federated learning is a significant step toward better data governance, but only when paired with clear threat models, privacy mechanisms, security defenses, deployment discipline, and institutional accountability. Its value lies not in making privacy automatic, but in creating a framework within which privacy can be designed, measured, and governed.

References

[1] McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 1273-1282.

[2] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1-2), 1-210.

[3] Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology, 10(2), Article 12.

[4] Li, X., Huang, K., Yang, W., Wang, S., & Zhang, Z. (2020). On the convergence of FedAvg on non-IID data. International Conference on Learning Representations.

[5] Zhu, L., Liu, Z., & Han, S. (2019). Deep leakage from gradients. Advances in Neural Information Processing Systems Workshop.

[6] Geiping, J., Bauermeister, H., Dröge, H., & Moeller, M. (2020). Inverting gradients: How easy is it to break privacy in federated learning? Advances in Neural Information Processing Systems, 33, 16937-16947.

[7] Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. IEEE Symposium on Security and Privacy, 3-18.

[8] Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. Theory of Cryptography Conference, 265-284.

[9] Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3-4), 211-407.

[10] Geyer, R. C., Klein, T., & Nabi, M. (2017). Differentially private federated learning: A client level perspective. arXiv preprint arXiv:1712.07557.

[11] McMahan, H. B., Ramage, D., Talwar, K., & Zhang, L. (2018). Learning differentially private recurrent language models. International Conference on Learning Representations.

[12] 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. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 1175-1191.

[13] Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konečný, J., Mazzocchi, S., McMahan, H. B., Van Overveldt, T., Petrou, D., Ramage, D., & Roselander, J. (2019). Towards federated learning at scale: System design. Proceedings of Machine Learning and Systems, 1, 374-388.

[14] Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M., & Cardoso, M. J. (2020). The future of digital health with federated learning. npj Digital Medicine, 3, Article 119.

[15] Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D., Colen, R. R., & Bakas, S. (2020). Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. Scientific Reports, 10, Article 12598.

[16] Dayan, I., Roth, H. R., Zhong, A., Harouni, A., Gentili, A., Abidin, A. Z., Liu, A., Beardsworth Costa, A., Wood, B. J., Tsai, C. S., et al. (2021). Federated learning for predicting clinical outcomes in patients with COVID-19. Nature Medicine, 27, 1735-1743.

[17] Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., Eichner, H., Kiddon, C., & Ramage, D. (2018). Federated learning for mobile keyboard prediction. arXiv preprint arXiv:1811.03604.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Open Access