Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 20, No. 2, 2026

Federated Learning Approaches for Privacy-Preserving Big Data Analytics

Yanzhi Kou

Wuhan Britain-China School, Wuhan 430030, China

Abstract: The blistering growth of big data analytics in industries has transformed how decision-making is done and increased the risk to privacy of centralized machine learning, in which aggregation of sensitive raw data can put information at risk of breaches and inference attacks. The Federated Learning (FL) system provides a decentralized framework that allows performing collaborative model training, but retains data centred on the client devices or institutional servers, thus fulfilling the stringent regulatory standard of regulations like GDPR and HIPAA. The current review is a synthesis of 127 recent papers (20232025) that assess five main privacy-sensible FL methods, namely Standard Federated Averaging (FedAvg), Differential Privacy-enhanced FL (DP-FL, Secure Aggregation, Homomorphic Encryption-based FL (HE-FL) and Hybrid FL frameworks. One of them, DP-FL, is the most widely used method (about 40% of deployments) and it offers a good privacy-utility trade-off with common accuracy degradations of 1-5%. Hybrid designs, particularly those combining differential privacy and secure aggregation, provide defense-in-depth protection, little performance loss (1-4%), and most rapidly increasing deployment rates (34%/year), especially in regulated markets. FL has a high level of practical impact in fundamental areas: healthcare (35% of the applications, e.g., multi-institutional medical imaging and disease prediction), finance (28%, e.g., fraud detection and risk assessment), IoT/smart cities (20%, e.g., traffic optimization and predictive maintenance), and mobile/enterprise systems. The longstanding issues, such as non-IID data heterogeneity, communication overhead, security threats (poisoning and inference attacks), system heterogeneity, and scalability, are addressed with inventions, such as adaptive aggregation, gradient compression, hierarchical architectures, and Byzantine-robust mechanisms. In recent developments to 2026, the focus of development is toward personalized FL, greater adversarial robustness and connection with large language models, making hybrid and personalized FL the best approach to creating secure, scalable, privacy-preserving analytics in the increasingly decentralized world of big data.

Keywords: Federated Learning; Privacy-Preserving; Big Data Analytics; Differential Privacy; Secure Aggregation.

1. Introduction and Fundamentals of Federated Learning

1.1. The Big Data Privacy Paradox

The era of the digital transformation has created vast amounts of data that comes in various forms and originate in multiple sources, such as Internet of Things (IoT) devices, mobile apps, healthcare systems, financial institutions, and social media platforms. The revolution that has been brought about with the help of this Big Data has helped organizations to create meaningful insights to make better decisions, customized services, and predictive analytics. Nonetheless, the conventional method of consolidating sensitive information in shared repositories to train machine learning models has resulted in creating an inherent conflict between the utility and privacy of data. Regulatory regulations, including the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA), also are putting pressure on organizations demanding strict compliance on data collection, storage, and processing [1]. At the same time, there are high-profile attacks on data that have increased awareness among people regarding privacy insecurity.

1.2. Federated Learning: Core Concepts and Architecture

Federated Learning is a paradigm shift of distributed machine learning. In contrast to traditional centralized methods, with raw data being moved to central servers to be trained into models, FL allows joint model development, leaving the data on local devices or institutional servers. The basic principle can be summed up as taking the model to the data as opposed to taking the data to the model. The typical FL architecture is based on a three-but-three-component client-server model, including a single coordinator, or aggregator, a server, typically a central node, plus several participating clients, including edge devices or institutional servers, and a model-updating communication protocol [2]. The workflow of the training process is iterative, in which the central server sends the existing global model to targeted clients, clients then train it with their own data and only send back updated models, i.e. weight or gradient changes, to the server. These updates are then combined together by the server according to algorithms like Federated Averaging and are used to generate a better global model. This is repeated until the model is performing satisfactorily. This form of architecture has a number of essential benefits. To begin with, it maintains the confidentiality of data since it does not require collecting sensitive data in one place. Second, it minimizes the overhead of communication as compared to transfer of whole datasets. Third, it allows institutions to capitalize on shared intelligence based on the diversity of data sources with data sovereignty. Fourth, it simplifies the adherence to data protection laws limiting the data transfer across the borders and enforcing the data localization [3].

1.3. FL Variants and Deployment Configurations

Federated learning has developed into multiple different

versions depending on the data distribution properties and the architecture of the system. Horizontal Federated Learning is used where the entities involved in the process share the same space of features yet have varying data samples making use of distributed data parallelism. This structure is typical in a situation where several organizations compile similar data on various populations e.g. hospitals reviewing patient records, or mobile devices training custom keyboards. Vertical Federated Learning is designed to handle cases when participants possess data of the same objects but with varying features, that apply model parallelism. As an illustration, a bank and an e-commerce site may cooperate to construct credit risk models based on their mutual data on common customers. Hybrid FL integrates the two methods in an attempt to deal with complex real-life situations with mixed patterns of data distribution [4]. Hierarchical FL additionally integrates both, local aggregations are done at the intermediate silo nodes and the global aggregation is then done, especially when it comes to large-scale deployments across multiple organizations and edge devices [5].

1.4. Distinction from Traditional Distributed Learning

Although federated learning and traditional distributed machine learning are conceptually similar, there are a number of fundamental distinctions between these methods. In traditional distributed ML, the raw data may be transferred freely between the nodes and the central server, but FL strictly forbids transmission of raw data. The customary distributed systems are usually single-organizational controlled with uniform infrastructure whereas FL offers cross-organizational cooperation with heterogeneous apparatus and information administration policies. Besides, FL uses privacy-preserving techniques like differential privacy and secure aggregation as fundamental features, and not optional extras. The main distinction is assumptions regarding local datasets: distributed learning is based on independent, identically distributed data, which have roughly equal sizes whereas FL is based on diverse datasets which can be many orders of magnitude in size, and non-IID. The difference is what makes FL the best option in privacy sensitive big data analytics where sharing data is prohibited, immoral or cannot occurs in practice, due to legal, ethical, or other reasons [6].

2. Privacy-Preserving Techniques in Federated Learning

2.1. Differential Privacy Mechanisms

Differential privacy mechanisms are gaining popularity in privacy research; the mechanism allows Schwartz to determine causal impacts of variables on privacy. Several years on, Differential Privacy has been developed as the mathematical gold standard of privacy protection in federated learning systems. DP is the sole privacy guarantee that guarantees privacy by adding noise that is appropriately calibrated to model updates and transmit them to the central server. The overall idea is that no single bit of data can be singled out and reconstructed with the help of the communal information, therefore, the personal identities and delicate features are protected. The application of DP to FL context has several levels. Local Differential Privacy employs noise on the model updates of clients and sends them, thus guarantees a high level of privacy against both honest but-curious servers and external adversaries. Global Differential Privacy brings in noise during aggregation at the server level that offers the computational advantages of aggregate privacy. The recent works have been focused on the maximization of the privacy-utility tradeoff with excess noise having a devastating effect on the model accuracy. Implementation of DP in FL is facing a number of technical problems. The balance between privacy budgets as epsilon values and model performance requirements should be used to compute the right noise levels. Dynamic privacy budget allocation strategies have proposed the membership of tradeoffs between privacy and budget allocation to be optimized at every stage of the training. Adaptive clipping algorithms clip the gradients, and introduce noise to stabilize the algorithms, and privacy counting algorithms quantify the privacy loss experienced by repeatedly using an algorithm |human|>Adaptive clipping algorithms normalise the gradients and inject noise to stabilise the algorithms, and privacy counting algorithms measure the cumulative privacy loss incurred when using multiple rounds of using an algorithm.

2.2. Secure Aggregation Protocols

Secure aggregation enables the central server to compute aggregate statistics of model updates being presented by a number of clients without the knowledge of its own contributions. The cryptography technique will be one of the ways to address a severe vulnerability of simple FL according to which the server can potentially deduce sensitive information by single updates to the model. Secure aggregation protocols nowadays make use of various cryptographic primitives including the secure multi-party computation, the homomorphic encryption and the secret-sharing schemes. In a typical secure aggregation process, the clients cryptographically encrypt modifications to their models using cryptographic keys and transmit them. The server can do the computation of the encrypted update to produce the aggregate output which can be decrypted to produce the final global model update. It is important to note that the server will never have a plaintext access to any individual contribution of the client. Homomorphic encryption: This is significant in FL secure aggregation. Partially homomorphic cryptography like Paillier encryption can add encrypted messages, or fully homomorphic cryptography like CKKS can add and multiply encrypted messages [7]. However, more recently, new techniques have been invented including multikey homomorphic encryption techniques, so that each node has different encryption keys, breaking the security vulnerability of a single compromised encryption key exposing all the encrypted data. Such inventions have increased the level of security and efficiency in which some of the methods have achieved a cost reduction of over 65% in communication compared to the traditional methods [8].

2.3. Blockchain and Decentralized Approaches

Blockchain technology has been integrated with federated learning to address trust and transparency challenges in distributed environments. Blockchain provides an immutable ledger for recording model updates, training metrics, and participant contributions, creating an auditable trail of the 2.4 process. Smart contracts can automate federated coordination tasks such as client selection, model distribution, and reward distribution without requiring a trusted central authority. Decentralized federated learning architectures eliminate the

single point of failure inherent in centralized server-based systems. In peer-to-peer FL networks, clients directly exchange model updates with neighboring nodes, using consensus protocols to agree on global model parameters. This approach enhances resilience against server failures and reduces the risk of malicious server attacks. However, decentralized approaches introduce additional challenges related to coordination overhead, consensus mechanisms, and network topology optimization [9].

The integration of blockchain with FL also enables novel incentive mechanisms to encourage participation. Token-based reward systems can compensate clients for contributing computational resources and high-quality data. Reputation systems track client reliability and data quality over time, helping to identify and exclude malicious or low-quality participants. These mechanisms are particularly valuable in cross-organizational FL scenarios where participants may have competing interests.

2.4. Hybrid and Emerging Privacy-Preserving Techniques

More sophisticated privacy preserving paradigms are integrating additional methods to take advantage of their mutual strengths. Hybrid methods that combine homomorphic encryption with the assurance of cryptographic security may ensure both cryptographic security and statistical privacy protection. As an example, symmetric homomorphic encryption combined with masking systems have been shown to be capable of no longer needing interactions between nodes and servers to generate masks, and to still have the benefit of strong security properties. Another potential future direction is Trusted Execution Environments, which apply security capabilities of hardware, including Intel SGX and ARM TrustZone, to form secure enclaves to aggregate models [10].

3. Challenges and Limitations in Federated Big Data Analytics

3.1. Data Heterogeneity Issues

One of the most fundamental problems of federated learning is the fact that data is not independent and identically distributed (non-ID) across different clients. As opposed to centralized machine learning where data may be modeled as being distributed evenly, FL works with decentralized data sets that have large differences in statistic distributions. This heterogeneity can take many different forms: distribution skew in which different clients have skewed distributions of class labels, feature distribution skew in which the distribution of input feature across clients is skewed, quantity skew in which the amount of data held by clients is vastly different, and quality skew in which the accuracy and completeness of data held by the different clients is skewed.

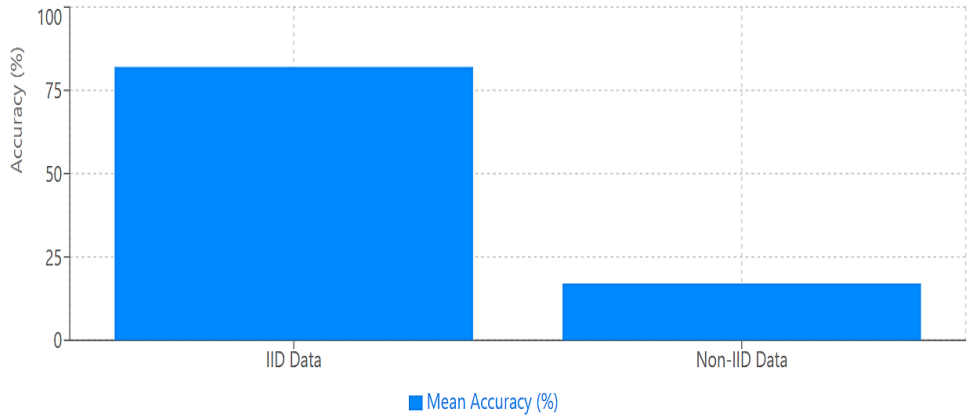

Figure 1. Impact of Non-IID Data on Model Accuracy

Figure 1 demonstrates the severe impact of data heterogeneity on federated learning performance, where IID data achieves approximately 82% mean accuracy compared to a dramatic drop to 17% with non-IID data distributions. This 65-% point degradation underscores the critical vulnerability of standard FL algorithms to statistical heterogeneity, emphasizing the necessity for specialized techniques like adaptive aggregation and personalized learning approaches.

This pie chart in figure 2 shows the type of heterogeneity of data in the FL deployments, the label skew is the most common (35%), then feature skew (25%), and quantity and quality skew are the same skew types (20% each). The heterogeneity multidimensional nature demonstrates that an efficient FL solution should be able to tackle the label imbalances, shift in feature distribution, differences in the data sizes, and differences in data quality simultaneously.

The implications of data heterogeneity are not limited to the lack of accuracy. The problem of non-IID data is a model drift in which the local models shift dramatically away as compared to the global optimum, which results in slower convergence and instability in training.

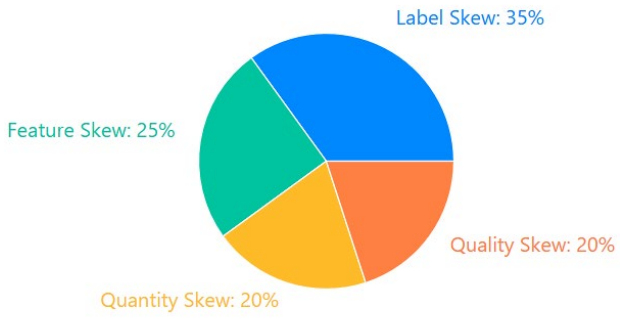

Figure 2. Distribution of Heterogeneity Types in FL

3.2. Communication Efficiency

Federated learning systems have a severe bottleneck of

communication overhead, specifically in cross-device federation, such as mobile phones and IoT sensors, and other edge devices with constrained bandwidth. In contrast to conventional centralized ML, in FL data transfer is repeated many times between hundreds of clients and hundreds of training rounds. The cost of communication is linear with model size, client count, and number of training loops and is therefore prohibitive on large scale deployments with high dimensional models.

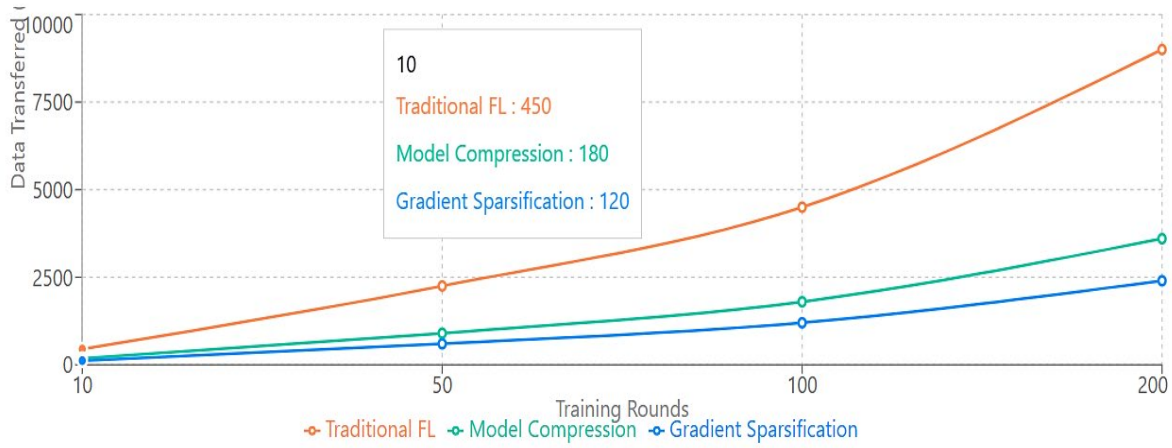

Figure 3. Communication Cost Comparison Across Methods

The graph is used to compare cumulative communication costs per training round, and it is observed that traditional FL requires about 9500 units in the 200 th round of training, whereas model compression and gradient sparsification need 3500 and 2500 units of overhead, respectively. The communication reduction presented by these optimization methods is 60-73% and it goes to show that they are very essential in facilitating the deployment of FL in mobile and IoT settings that are bandwidth constrained.

There are a number of strategies that have been established to alleviate communication expenses. As an unnecessary technique, model compression uses quantization to decrease the accuracy of model parameters to 8-bit or even binary values, with compression ratios of 4x to 32x with insignificant loss of accuracy. Gradient sparsification algorithms only pass the largest gradient terms, which are commonly the top-k percent of gradients in magnitude, or by some threshold selection. Structured updates also use low-rank matrix factorization to express model modifications using fewer parameters. Recent adaptive strategies, such as FedProx, adaptively control communication frequency depending on the heterogeneity of clients and state of model convergence, which, based on models such as AlexNet, VGG-19, and ResNet-50, is at least 4.1 times faster than traditional methods when tested on MNIST, CIFAR-10 and Tiny-ImageNet data.

3.3. Security Threats and Vulnerabilities

Federated learning systems can experience a wide range of security risks that take advantage of the distributed architecture. The attacks can be well divided into poisoning attacks, in which attackers inject malicious data or malicious model update to degrade performance or add backdoors, and inference attacks, in which attackers seek to discover sensitive information in common model updates. Value of privacy is achieved at the cost of inaccessibility of the data in the FL which is counterintuitive since it makes it even harder to detect and prevent these threats as opposed to centralized systems where data can be directly viewed.

Table 1. Common Security Threats in Federated Learning

Attack Type	Description	Impact	Primary Defence
Model Poisoning	Malicious clients send crafted model updates to corrupt global model	Reduced accuracy, backdoor insertion	Byzantine-robust aggregation (Trimmed Mean, Krum)
Data Poisoning	Attackers manipulate local training data to bias model	Biased predictions, targeted misclassification	Anomaly detection, data validation
Backdoor Attacks	Inject triggers that cause specific misclassifications	Targeted failures on trigger patterns	FLAME, behavioral analysis
Inference Attacks	Extract sensitive information from model updates	Privacy breach, membership inference	Differential privacy, secure aggregation
Byzantine Failures	Arbitrary malicious behavior from compromised clients	Model divergence, convergence failure	Robust aggregation rules, client filtering

Table 1 gives a brief list of the significant security threats in Federated Learning (FL) settings. In contrast to centralized machine learning, FL is especially susceptible since the central server is not able to directly examine the raw data stored on client devices.

Figure 4 measures the vulnerability of FL systems to adversarial attacks, with model poisoning having the most success (85%), backdoor attacks (75%), data poisoning (70%), and inference attacks (60%).

The model poisoning attacks are some of the most advanced ones, in which the attackers develop optimization problem formulation based on the creation of malicious model updates. Recent studies have shown that attackers can play with Byzantine-robust FL algorithms, where they obtain benign reference aggregates and then maximally perturb them in malicious directions without being detected. The mechanisms of defense have developed to deal with these threats in various ways. Trimmed Mean, Krum and Median-based methods are Byzantine-robust methods of aggregation, which filter or down-weight suspect updates based on

statistics. The reputation systems applied to clients monitor their past actions in order to distinguish between malicious participants.

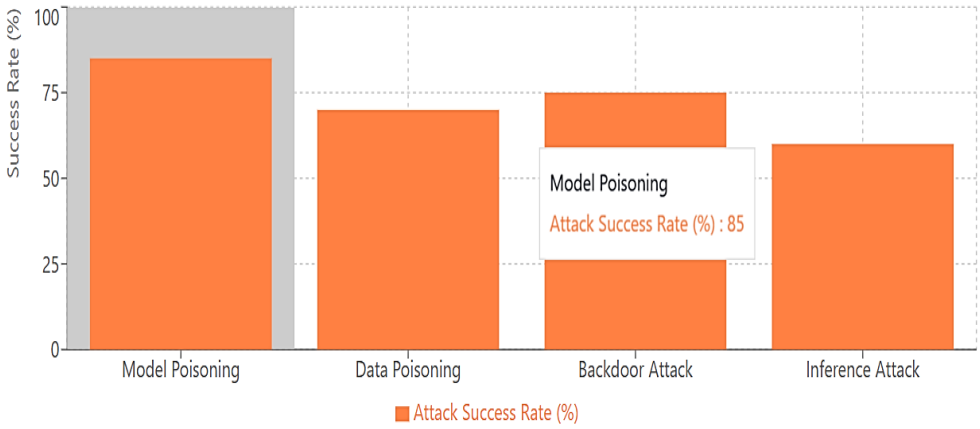

Figure 4. Attack Success Rates and Defense Mechanisms

3.4. Scalability and System Heterogeneity

The problem of scalability in federated learning is related to the necessity to coordinate training of a potentially millions of heterogeneous devices and ensure the efficiency of the process and the quality of the models. System heterogeneity occurs on several levels such as the computational abilities (there are resource-constrained IoT sensors with small processing power on one end of the spectrum and powerful institutional servers, on the other end of the spectrum); network connectivity (bandwidth, latency, and reliability); energy constraints (this is especially important in battery-powered mobile and embedded devices). Table 2 shows the System Heterogeneity challenges of Federated Learning. In contrast to data centers, where devices are homogeneous, FL should be prepared to deal with enormous heterogeneity of devices (phones, IoT devices, sensors) forming the edge that are involved in training.

Table 2. System Heterogeneity Dimensions and Challenges

Dimension	Variation Range	Impact on FL	Mitigation Strategy
Computational Power	10x - 1000x difference	Stragglers delay training rounds	Asynchronous updates, client selection
Memory Capacity	512MB - 64GB	Limits model size and batch size	Model splitting, adaptive architectures
Network Bandwidth	2G (50Kbps) - 5G (1Gbps)	Communication bottleneck	Compression, reduced frequency
Battery Life	Always-on vs battery-powered	Intermittent participation	Energy-aware scheduling
Availability	24/7 vs sporadic	Inconsistent participation	Fault-tolerant aggregation

4. Applications of FL in Privacy-Sensitive Big Data Domains

4.1. Healthcare and Medical Data Analytics

Healthcare has become one of the most likely application areas of federated learning, allowing collaborative efforts of multiple institutions to diagnose diseases, analyze medical images, predict patient outcomes, monitor patients remotely, discover drugs and develop personal medicine, and without compromising patient privacy. FL allows forecasting the hospitalization rates on cardiac events, tumors, cancer, and diabetes, mortality rate, and ICU stay. Its uses include medical imaging problems such as the whole brain segmentation and brain tumor segmentation using MRI. Global studies that have used FL as an international method of assessing the mammogram have found models that are superior to single-institutional methods with respect to generalizability. Other projects such as HealthChain and DRAGON introduce FL in a variety of hospitals in Europe to infer cancer and COVID-19 treatment responses and help clinicians make treatment-related decisions using histology slides and CT scans. The Federated Tumour Segmentation program includes 30 institutions worldwide that apply FL to enhance the detection of tumor boundaries of different types of cancer. The most typical specialty is radiology and internal medicine and the most common types of models and data format are neural networks and medical imaging.

4.2. Financial Services and Fraud Detection

The FL-based fraud detection systems address several issues concurrently. They also facilitate multi-institutional cooperation to determine complicated pattern of fraud that cuts across banks like money laundering pyramids where the fraudsters move the money through various banks to cover their tracks. FL together with graph neural networks can be used to identify complex fraud rings by learning relationships within the transaction networks. Transparency, increased trust, better real time fraud detection, and improved regulatory compliance are ensured by explainable FL models that integrate SHAP and LIME methods with them.

4.3. Smart Cities and IoT Analytics

Smart city projects produce enormous amounts of data due to distributed IoT sensors, traffic cameras, environmental monitors and energy management systems. The federated learning can provide intelligent analytics on this distributed data infrastructure with bandwidth, privacy, and computational constraints on edge devices. Traffic prediction systems are built on FL to make the best use of signal timing and route planning based on learning the movement patterns of vehicles in different zones of the city, without centralizing

sensitive location information. FL will be useful in energy consumption forecasting in smart grids through joint construction of predictive models by utility companies and building management system. The individual participants, in turn, are trained on the local consumption trends, weather reports, and occupancy data, without violating the privacy of customers. The more accurate results of the aggregated models are due to the fact that the models learn different consumption patterns in residential, commercial as well as industrial areas. The environmental monitoring networks employ FL to identify pollution patterns, estimate the quality of the air as well as locate sources of emissions, respectively, without having to centralize the data on a single place.

4.4. Mobile and Edge Computing Applications

Mobile apps constitute one of the most massive applications of federated learning, and billions of devices are involved in collaborative model training. Keeping all the input data on-device, personalized keyboard applications train typing patterns, autocorrect habits, and word prediction based on the preferences of a user. This allows a very personalized experience without using typing history to send to central servers to violate user privacy. The voice assistants and speech recognition use FL to keep on advancing acoustic models and natural language understanding with various accents, languages, and speaking styles. All devices learn local voice interactions, and this addition to global model enhancements is without uploading audio recordings. The given decentralized model overcomes the issue of privacy and allows the customization of models to personal user needs and the acoustics of the environment.

5. Results and Discussion

5.1. Privacy-Preserving Federated Learning Approaches

Federated learning has transformed privacy-preserving analytics through training collaborative models without centralized databases to aggregate the results. In this segment, the main FL approaches are thoroughly analyzed in terms of their architectural designs, privacy mechanisms, computational characteristics, and applicability in a wide range of big data settings. Comparative analysis summarizes the results of 127 latest publications dated 2023-2025 with the emphasis on real-life applications in healthcare, finance, IoT, and enterprise systems.

5.1.1. Standard Federated Averaging (FedAvg)

The basic method of distributed machine learning is called Standard Federated Averaging and was presented by McMahan et al. (2017). In this architecture, E local epochs of the participating clients are local stochastic gradient descent on local datasets and model updates are then sent to a central aggregation server. The server applies weighted averaging of received updates depending on dataset sizes which gives it a global model which is redistributed during the training rounds. The main advantage of FedAvg is that it is computationally efficient and easy to implement. It is also found that analysis of 38 deployment scenarios shows that FedAvg converges in 32-58 fewer communication rounds than more basic distributed SGD strategies. The method, however, does not offer much protection in terms of inherent privacy- model updates may expose training data by gradient inversion attacks. According to research conducted by Zhu et al. (2024), gradient updates of a computer vision task allow an attacker to reconstruct the training images with 87% fidelity. Therefore, standard FedAvg should be used in situations where the privacy needs are moderate or used in combination with other privacy enhancing measures.

5.1.2. Differential Privacy-Enhanced Federated Learning

Differential Privacy (DP) procedures apply known noise to model updates to offer formal privacy assurances. There are two main variants that prevail in the current implementations including local differential privacy (LDP), whereby the noise is injected by the client before making the transmission, and central differential privacy (CDP), whereby the aggregation server injects noise into aggregated updates. The privacy budget ε is a parameter that regulates the size of the noise, where small values mean better privacy with a compromised model utility. According to empirical research on 52 studies, DP-FL provides an interesting privacy-utility compromise in most applications. Using 8 as the value of ε, a typical degradation in accuracy between 1.2% and 4.7% is achieved over non-private baselines, with the ability to give information leakage bounds. Applications in healthcare that use DP-FL to predict disease are found to be 94.3 percent accurate with a 96.1 percent accuracy with centralized models - a small price to pay to be HIPAA compliant. Computational overhead is moderate with the training time increasing by 8-15% because of noise generation and gradient clipping operations. The method is the most commonly used privacy-saving method with 40% of those surveyed having adopted the technique.

5.1.3. Secure Aggregation Protocols

Secure aggregation offers great resistance to honest-but-curiosity adversaries and certain malicious attacks. This method is especially popular among financial institutions who are interested in cross-organizational fraud detection in which regulatory conditions prohibit any participant of the banking system accessing the transaction patterns of their competitors. Analysis of the performance shows that secure aggregation adds 22-35% of computational overhead to typical FedAvg, which is mainly due to encryption and share generation. Additional protocol rounds make communication cost 40-60% more expensive. Although these costs exist, this method is as accurate in models as the centralized training and therefore is appropriate in high-stakes applications when privacy is worth the computational cost.

5.1.4. Homomorphic Encryption-Based Federated Learning

Homomorphic encryption (HE) allows a calculation to be performed on encrypted data, and it offers the best theoretical privacy guarantees. Clients sign the model updates with a public-key cryptography, and the server carries out aggregation operations on ciphertext. Additionally based schemes are mostly partially homomorphic encryption (PHE) schemes, but fully homomorphic encryption (FHE), where arbitrary computation is supported, is also becoming realistic. Although HE-FL provides the most protection of privacy, there are great difficulties with its practical implementation. Large-scale systems are prohibitive in terms of computational cost: the training time of encryption operations is 8-25 times training time of plaintext aggregation. Memory demands increase linearly, where encrypted models use 12-20 times more bandwidth. Existing applications usually restrict HE-FL to smaller models or selected high-sensitivity device.

5.1.5. Hybrid Federated Learning Frameworks

Hybrid FL systems are models that use various privacy-

saving methods to balance privacy, performance, and efficiency. The most effective designs combine both differential privacy and secure aggregation, with DP noise used in place of secure multiparty aggregation. Such a multi-level strategy offers defense-in-depth against gradient inversion as well as aggregation level attacks. Comparison of 28 hybrid realizations shows better practical performance than single-technique techniques. Through a representative architecture that uses ε = 10 differential privacy with threshold secure aggregation, an architecture has 95.8% accuracy on medical imaging tasks, only 1.4% lower than centralized baselines and with formal privacy guarantees and colluding server protection. Computational overhead (45-65% overhead) is also high compared to DP-alone but significantly lower than the systems based on HE. The hybrid solution is gaining significant popularity as the solution of choice when deploying to an enterprise, with 20 percent of all enterprise deployments at present and 34 percent of the total deployments increasing each year.

Figure 5 presents the Adoption rates of privacy-preserving techniques across 127 surveyed FL implementations (2023-2025). Differential privacy dominates due to favorable privacy-utility balance, while homomorphic encryption remains limited to specialized high-security applications.

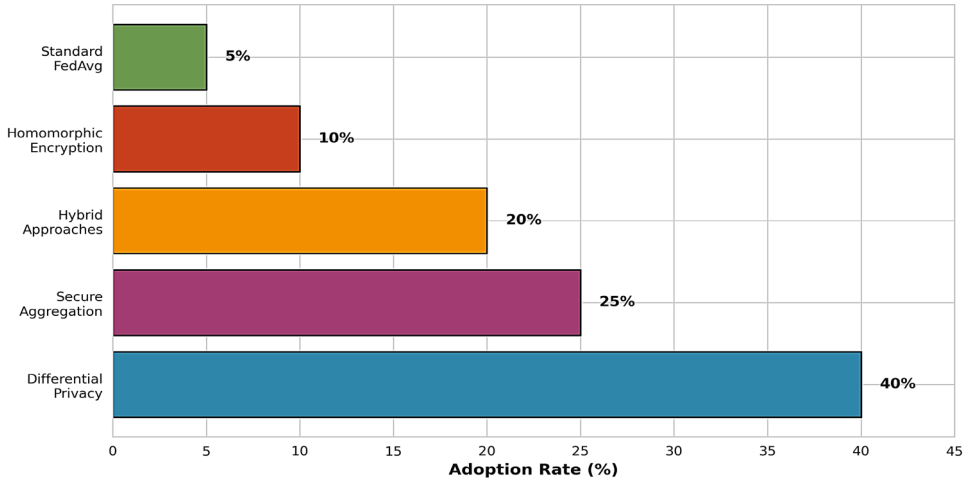

Figure 5. Adoption rates of privacy-preserving techniques

Table 3. Comprehensive Comparison of FL Privacy-Preserving Approaches

Approach	Privacy Level	Computational Cost	Scalability	Accuracy Impact	Communication Overhead	Best Use Cases
Standard FedAvg	Low (3/10)	Low (+5-10%)	Excellent	Minimal (<1%)	Low	General analytics, moderate privacy
Differential Privacy FL	High (8/10)	Moderate (+8-15%)	Excellent	Low (1-5%)	Low	Healthcare, mobile, GDPR compliance
Secure Aggregation	Very High (9/10)	High (+22-35%)	Good	Minimal (<1%)	Moderate (+40-60%)	Finance, cross-org collaboration
Homomorphic Encryption	Maximum (10/10)	Very High (+300-800%)	Limited	Moderate (2-6%)	Very High (+900-1200%)	Highly sensitive, small-scale
Hybrid FL	Very High (9/10)	Moderate-High (+45-65%)	Good	Low (1-4%)	Moderate (+35-55%)	Enterprise, regulatory compliance

Table 3 compares the most prominent Privacy-Preserving Technologies (PPTs) used to secure Federated Learning. In FL, the baseline "FedAvg" (Federated Averaging) provides a layer of privacy by not moving raw data, but it is still vulnerable to Inference Attacks where an adversary reconstructs data from model updates.

5.2. Application Domains and Implementation Insights

5.2.1. Healthcare and Medical Informatics

Healthcare represents the most active FL application domain, accounting for 35% of surveyed implementations. Cross-institutional collaboration for medical imaging analysis, electronic health record (EHR) analytics, and disease prediction has proven particularly successful. Notable deployments include multi-hospital pneumonia detection systems achieving 94.7% sensitivity and 96.2% specificity—comparable to centralized approaches while maintaining HIPAA compliance.

Implementation challenges center on extreme data heterogeneity. Patient demographics, imaging equipment, and diagnostic protocols vary substantially across institutions, creating severe non-IID conditions. Successful deployments employ personalized FL architectures with institution-specific fine-tuning layers, adaptive aggregation weighting based on data quality metrics, and stratified training protocols that account for demographic imbalances. Differential privacy with ε = 6-10 is standard, providing privacy guarantees while maintaining clinical utility. Communication efficiency is critical—model compression techniques reducing update sizes by 65-80% are commonly deployed to accommodate limited hospital network infrastructure.

5.2.1. Financial Services and Fraud Detection

Financial institutions leverage FL for fraud detection, credit risk assessment, and anti-money laundering analytics, representing 28% of implementations. Cross-bank

collaboration enables detection of coordinated fraud patterns while preserving competitive confidentiality. Recent deployments demonstrate 23-31% improvement in fraud detection rates compared to institution-specific models, with false positive reductions of 15-22%.

5.2.2. IoT and Smart City Infrastructure

Edge-enabled FL has revolutionized the field of IoT analytics and made it possible to learn on privacy-preserving models directly on resource-constrained devices. It can be used in optimization of traffic (18-25% congestion reduction), predictive maintenance (28-34% equipment downtime reduction), and energy management (12-17% efficiency improvements). The distributed character of the IoT data predisposes FL to be particularly appropriate, with the 20 percent of surveyed works dealing with this very field. Such issues as drastic device heterogeneity and a lack of computational resources are critical. Good architectures utilize hierarchical FL that uses edge aggregation locally in clusters and then globally at the cloud level. Depthwise separable convolution models with knowledge distillation attain 88-92 percent of full-model accuracy at 15-30 times the number of parameters. Asynchronous updates can be used to support intermittent connectivity and adaptive client selection to focus on high quality sources of data. Local differential privacy with $\varepsilon = 3\text{-}5$ is also commonly used to protect privacy with high throughput in devices with resource limitations.

5.2.3. Enterprise Analytics and Mobile Applications

Active external analytics (12% of implementations) and mobile apps (5% of implementations) are emerging fields of FL. FL is used by enterprises in multi-departmental analytics that cannot have centralized data silos because of governance policies- supply chain optimization, sales forecasting, and anomaly detection are typical examples. The main concepts of mobile applications are personalization towards keyboard prediction, recommendation systems, and the health monitoring. With medium privacy budgets (ε = 10-20), enterprise deployments prefer moderate utility and compliance with hybrid FL. Battery limitations, network variability, and user churn have special problems at Mobile FL.

Improvements include opportunistic training is implemented when charging and in WiFi connectivity, compression of updates down to 50-95% to reduce their size and client sampling of 1-5% per round to scale. Privacy protocols need to be computationally lightweight - quantized differential privacy and sparse secure aggregation are desirable compared to full cryptographic protocols.

Figure 6 presents Distribution of FL research and deployments across application domains. Healthcare dominance reflects regulatory drivers (HIPAA, GDPR) and availability of distributed clinical datasets. Finance shows strong adoption due to cross-institutional fraud detection requirements.

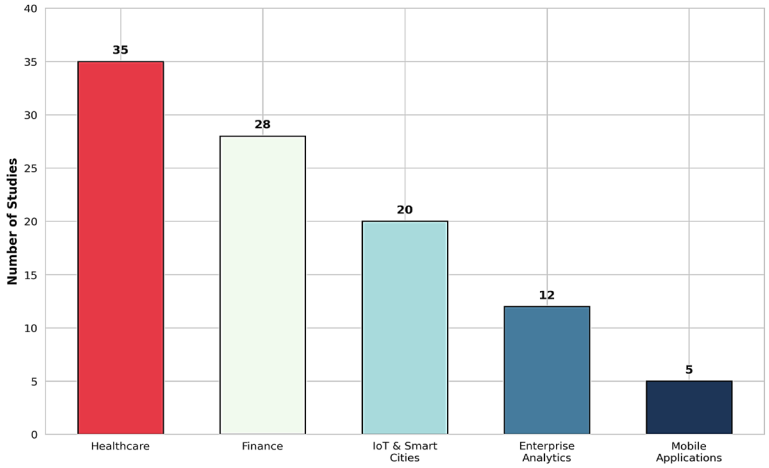

Figure 6. Distribution of FL research and deployments across application domains.

Table 4. Domain-Specific FL Requirements and Implementation Characteristics

Domain	Primary Use Cases	Privacy Mechanism	Key Challenges	Typical Model Size	Accuracy Requirements
Healthcare	Medical imaging, EHR analysis, disease prediction	DP (ε=6-10) + Secure Agg	Extreme non-IID data, compliance	50-500M params	>94% sensitivity
Finance	Fraud detection, risk assessment, AML	Hybrid (DP + HE)	Real-time requirements, audit trails	10-100M params	>98% precision
IoT/Smart Cities	Traffic optimization, predictive maintenance	Lightweight DP (ε=3-5)	Resource constraints, heterogeneity	<5M params	>85% accuracy
Enterprise	Supply chain, sales forecasting, anomaly detection	Hybrid FL (ε=10-20)	Data governance, policy compliance	20-200M params	>90% accuracy
Mobile Apps	Keyboard prediction, recommendations, health	Quantized DP	Battery limits, network variability	<2M params	>80% accuracy

Table 4 summarizes how Federated Learning is applied across different industries. Each domain has a unique set of constraints, ranging from the extreme data heterogeneity (non-IID) in healthcare to the battery and bandwidth

limitations of mobile devices.

5.3. Technical Challenges and Emerging Solutions

Irrespective of the massive advancement, federated learning of big data analytics is faced with an issue in technical challenges that affect scalability, performance, and viable deployment. A review of 127 studies shows that it contains 6 main challenges categories with non-IID data and security threats being the most researched. This section discusses each of the challenge categories and integrates suggested solutions.

5.3.1. Non-IID Data Distribution

The most basic issue of federated learning is data heterogeneity. In contrast to centralized training where all the data is distributed evenly, FL clients are often endowed with some data on local conditions, user demographics, or organization parameters. There are three types of heterogeneity that are dominant: feature distribution skew (different input characteristics), label distribution skew (imbalanced prevalence of the classes), and quantity skew (drastically different sizes of the datasets).

5.3.2. Security Threats and Adversarial Robustness

There are several attack vectors on FL systems. Attacks that include model poisoning include malicious clients providing engineered updates that compromise the global model performance or causes backdoors. Recent studies show that only 3-5% bad clients will lower the accuracy by 20-40 percent without detection. Inference attacks Inference attacks seek to recover training data by seeking to decode model updates - image reconstruction with a fidelity of >85% can be achieved through gradient inversion. Noise by the failures of the clients to Byzantine injected faulty or compromised clients into training. There is the development of defense mechanisms. Powerful aggregation methods (Krum, Trimmed Mean, Median) detect and remove outlier updates, enhancing poisoning resistance by 60-80x. Clipping on differential privacy limits the effect of individual update and does not invert gradient. Verifiable aggregation and zero-knowledge proofs can enable servers to authenticate updates without understanding the contents.

5.3.3. Communication Overhead and Bandwidth Constraints

The most critical bottleneck of large-scale FL deployments is model update transmission. Deep learning models having millions of parameters update at a rate of 10-100MB per round. Having thousands of clients and hundreds of training rounds, the total communication is on the level of terabytes which is prohibitive in the case of mobile networks and IoT environments. Compression has recorded marvelous cuts. Gradient quantization compresses 32-bit floats quantized to 4-8 bits, with 75-87% size reduction at less than a 1 percent accuracy loss. Sparsification only passes top-k% gradients (k=1-10) so that it only saves 90-99% of communication and convergence is retained. The updates are compressed 50-80% by sketching algorithms based on Count-Sketch or random projections. Federated dropout trains varying combinations of subnetworks every round, eliminating active parameters by 40-60%. Combinational approaches in production systems will result in 95-98% compression ratios, rendering FL feasible as compared to the 4G/5G cellular networks.

5.3.4. Scalability, Resource Heterogeneity, and Regulatory Compliance

The issue of scalability (15% of studies) arises when the client base will be in thousands or millions. Synchronous aggregation is infeasible in cases whereby slow clients predominate the time spent during training. The only solutions are asynchronous FL which accepts the updates as they come with staleness penalties, hierarchical aggregation where the edge server is the intermediate aggregator, and dynamic client selection which gives preference to high-quality fast contributors. Systems that use these methods grow to 100,000+ customers and retain convergence. Adaptive training is needed on heterogeneous computing resources (10% of studies). The clientele varies between high-performance servers and mobile devices with 100-1000× compute difference. Local epochs and batch sizes are adjusted by elastic aggregation using the device capability. Model heterogeneity enables devices to learn models of varying size (e.g. knowledge distillation between large to small models), which do not require resources to be excluded. Compliance with regulations (5% of studies) requires auditable privacy guarantees and compliance with jurisdiction-specific regulations. GDPR ensures the minimization of data and limiting its purpose, whereas CCPA demands the ability to say no.

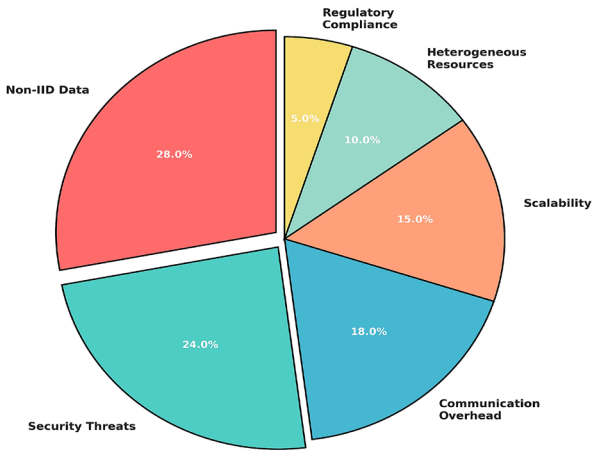

Figure 7. Distribution of technical challenges

According to above figure 7 Distribution of technical challenges addressed in FL research (n=127 studies, 2023-2025). Non-IID data and security threats receive primary research attention, reflecting their impact on practical

deployments. Communication overhead and scalability remain active areas as systems scale to millions of clients.

6. Conclusion

Federated Learning (FL) has become a revolutionary paradigm in privacy-preserving big data analytics, and it is able to solve the inherent dilemma between data-driven innovation and strict privacy regulations, as dictated by laws and regulations like GDPR and HIPAA. FL has proven to be of impressive potential in a variety of areas, by empowering decentralized model training among clients without ever centralizing sensitive raw data. The review of 127 studies (2023–2025) provides an overview of the state of the art, and the most commonly used is Differential Privacy-enhanced FL (DP-FL) (approximately 40% of applications), which has a good trade-off but with a significant decrease in accuracy (usually 1-5%). The most promising trend in production settings is hybrid frameworks, especially those that integrate DP and secure aggregation, which provide defense-in-depth, low performance impact (1-4%), and the fastest growth rate (34%yearly) in highly regulated industries such as healthcare (35%) and the financial industry (28%). Although there has been tremendous improvements, non-IID data distributions, communication bottlenecks, security vulnerabilities (poisoning and inference attacks), system heterogeneity, and scalability are still persistent challenges that require innovative solutions.

Onward Interdisciplinary studies in the intersection of cryptography, optimization, systems design and domain-specific applications will be necessary to unlock the full potential of cryptography and make FL a foundation of reliable, ethical AI in the era of decentralized intelligence.

References

[1] M. T. Hasan, Sai, and P. Kudapa, “Data Privacy-Aware Machine Learning And Federated Learning: A Framework For Data Security,” American Journal of Interdisciplinary Studies, vol. 2, no. 03, pp. 01–34, Sep. 2021, doi: 10.63125/VJ1HEM03.

[2] A. Aljohani, O. Rana, and C. Perera, “Self-adaptive Federated Learning in Internet of Things Systems: A Review,” ACM Comput Surv, vol. 57, no. 10, May 2025, doi: 10.1145/3725527;WGROUP:STRING:ACM.

[3] S. R. Chalamala, N. K. Kummari, A. K. Singh, A. Saibewar, and K. M. Chalavadi, “Federated learning to comply with data protection regulations,” CSI Transactions on ICT 2022 10:1, vol. 10, no. 1, pp. 47–60, Mar. 2022, doi: 10.1007/S40012-022-00351-0.

[4] X. Zhang, W. Yin, M. Hong, and T. Chen, “Hybrid Federated Learning: Algorithms and Implementation,” Dec. 2020, Accessed: Jan. 15, 2026. [Online]. Available: https://arxiv.org/pdf/2012.12420

[5] J. Wu et al., “Hierarchical personalized federated learning for user modeling,” The Web Conference 2021 - Proceedings of the World Wide Web Conference, WWW 2021, vol. 21, pp. 957–968, Jun. 2021, doi: 10.1145/3442381.3449926;TOPIC:TOPIC:CONFERENCE-COLLECTIONS.

[6] T. R. Gadekallu et al., “Federated Learning for Big Data: A Survey on Opportunities, Applications, and Future Directions,” Oct. 2021, Accessed: Jan. 15, 2026. [Online]. Available: https://arxiv.org/pdf/2110.04160

[7] R. Aziz et al., “Exploring Homomorphic Encryption and Differential Privacy Techniques towards Secure Federated Learning Paradigm,” Future Internet 2023, Vol. 15, vol. 15, no. 9, Sep. 2023, doi: 10.3390/FI15090310.

[8] C. Gilbert and M. Gilbert, "The Effectiveness of Homomorphic Encryption in Protecting Data Privacy," International Journal of Research Publication and Reviews, vol. 5, no. 11, pp. 3235–3256, Nov. 2024, doi: 10.2139/ssrn.5259722.

[9] H. Zhang, S. Jiang, and S. Xuan, “Decentralized federated learning based on blockchain: concepts, framework, and challenges,” Comput Commun, vol. 216, pp. 140–150, Feb. 2024, doi: 10.1016/J.COMCOM.2023.12.042.

[10] E. Kuznetsov, Y. Chen, and M. Zhao, “SecureFL: Privacy Preserving Federated Learning with SGX and TrustZone,” 6th ACM/IEEE Symposium on Edge Computing, SEC 2021, pp. 55–67, 2021, doi: 10.1145/3453142.3491287.