Federated Learning: Privacy-Preserving Collaborative Machine Learning

SurveyForge

Abstract— Federated Learning (FL) offers a transformative approach to distributed machine learning by enabling collaborative model training across decentralized data sources while preserving privacy. This survey systematically explores FL's core architectural designs, privacy-preserving methodologies, and practical applications across sectors like healthcare and finance. By leveraging techniques such as differential privacy, homomorphic encryption, and secure multiparty computation, FL enhances data confidentiality, though it must balance these innovations with computational efficiency and model utility. Additionally, the survey highlights prevailing challenges, including communication overhead, data heterogeneity, and adversarial threats in client-server interactions. Advanced aggregation protocols and decentralized architectures, like blockchain integration, are proposed to address these constraints. The paper also emphasizes future research directions in personalized FL, federated unlearning, and cross-modal data integration to ensure robust, secure deployments. Ultimately, FL's ability to align with regulatory compliance, tackle data inequities, and adapt to dynamic multi-modal environments underscores its pivotal role in the evolution of privacy-preserving AI technologies.

Index Terms—Privacy-preserving machine learning, Decentralized model training, Federated data integration

1 Introduction

Federated learning (FL) represents a transformative paradigm in distributed machine learning by facilitating collaborative model training across data residing on decentralized devices or silos, all while maintaining strict privacy constraints. Unlike traditional centralized learning, which necessitates aggregating data in a central repository, FL ensures that sensitive data remains local to the clients. This approach fundamentally adheres to the principles of data minimization and privacy preservation, which have become paramount amidst growing societal concerns regarding data misuse and the enforcement of stringent privacy regulations such as the General Data Protection Regulation (GDPR) [1], [2].

The core principle of FL revolves around its decentralized architecture, allowing multiple participants, such as mobile devices, edge nodes, or institutions, to collaboratively train shared machine learning models. These participants exchange only model updates (e.g., gradients or parameters) rather than raw data, significantly reducing the risk of sensitive information leakage [1], [3]. This shift aligns with real-world needs in domains such as healthcare, finance, and IoT-based applications, where privacy risks and datasets fragmented across organizational or geographic boundaries often inhibit joint data analysis. For instance, FL has enabled use cases like secure medical diagnosis tools, federated recommendation systems, and fraud detection models, proving its applicability across a variety of demanding scenarios [4], [5].

The field of federated learning has rapidly evolved since its formal introduction by Google in 2016 [6]. From its origins in training predictive text models on mobile devices, FL has expanded into a robust discipline tackling challenges such as heterogeneity among client systems, communication constraints, inefficient aggregation protocols, and vulnerabilities to adversarial attacks. Heterogeneity, for instance, manifests in the form of statistical imbalances in non-independently and identically distributed (non-IID) data across clients, leading to degraded model performance compared to centralized learning frameworks [7], [8]. Addressing this challenge requires methods that enhance training robustness while preserving system scalability, such as client selection strategies, dynamic weighting techniques, and personalized federated models [9], [10].

Privacy concerns have also spurred significant research in integrating secure computation techniques, including differential privacy (DP), secure multiparty computation (SMPC), and homomorphic encryption (HE). These approaches aim to mitigate risks associated with model updates revealing sensitive patterns or reconstructing raw data [11], [12]. While differential privacy achieves rigorous mathematical guarantees of privacy through noise injection, cryptographic methods ensure secure aggregation during model updates. However, the additional computational expense and resulting trade-offs in model utility and training efficiency continue to pose open questions [13], [14].

The historical development and societal urgency driving FL adoption reflect a broader shift in machine learning paradigms from centralized to distributed intelligence. Emerging trends include decentralized FL frameworks leveraging blockchain for tamper-proof aggregation and coordination, especially to address the vulnerability of centralized systems to single points of failure [15], [16]. Furthermore, recent explorations into hierarchical and incentive-compatible FL architectures aim to make large-scale deployments feasible by addressing scalability and ensuring fair resource utilization across participating clients [17], [18].

This paper synthesizes the foundational aspects of FL as a privacy-preserving collaborative learning paradigm, its evolution over time, and the practical and technical challenges yet to be resolved. Subsequent sections delve deeper

into architectural foundations, privacy threats, system optimization, and domain-specific deployments, presenting a nuanced view of the state of the art and identifying open research questions. By addressing these challenges and advancing its methodologies, FL holds the potential to revolutionize how machine learning balances privacy preservation and model utility in a decentralized world.

2 Core Architectures and Functional Foundations of Federated Learning

2.1 Federated Learning System Components

Federated learning (FL) systems rely on a well-orchestrated interplay between fundamental components, each playing a crucial role in ensuring efficient, privacy-preserving, and decentralized machine learning processes. These system components include clients or edge devices, a central server, mechanisms to manage data distribution and heterogeneity, and communication interfaces and protocols. This subsection examines these components, evaluates their interactions, addresses inherent trade-offs, and highlights emerging challenges and opportunities.

Clients, often represented as edge devices, are the cornerstone of federated learning systems, as they locally store data and perform initial model training. These include smartphones, IoT devices, or institutional data stores, which operate under constraints such as limited computational resources, battery efficiency, intermittent connectivity, and privacy rules. Clients train local models using their native datasets, compute model updates (e.g., gradients or weights), and communicate these updates to the central server. However, this decentralized setting introduces challenges associated with hardware discrepancies, non-uniform energy availability, and data heterogeneity among clients. For instance, non-independent and identically distributed (non-IID) data across clients—a common phenomenon—can significantly degrade model convergence and fairness, as highlighted in [7]. Efforts to address these issues focus on adaptive optimization techniques, such as FedProx or Scaffold, which impose constraints to stabilize learning under varying local conditions [8]. These approaches seek to balance the trade-off between maintaining utility and ensuring equitable participation across clients.

The central server in traditional FL architectures acts as a model aggregator, coordinating training, managing participant selection, and centralizing the synthesized model for subsequent refinement. This server aggregates received client model updates via methodologies like Federated Averaging (FedAvg), which computes a weighted average of updates based on client data volumes. While effective, reliance on the central server introduces a single-point-of-failure risk and potential security vulnerabilities if compromised [17]. Decentralized or hybrid approaches, such as blockchain-based FL frameworks, propose eliminating or distributing the server's role to enhance reliability and resilience [15]. By utilizing mechanisms like consensus protocols, these designs achieve fault tolerance and mitigate malicious server risks but at the cost of additional computational complexity and latency.

Data distribution and its inherent heterogeneity across clients pose significant challenges to federated learning system performance. Non-IID conditions exacerbate gradient divergence, slowing convergence or biasing models toward overrepresented client distributions [8]. Horizontal federated learning (HFL) systems primarily operate on clients sharing similar feature spaces but varying samples, while vertical federated learning (VFL) applies to scenarios where clients contain complementary features of the same entities [19]. Although these paradigms address specific use cases, neither can fully resolve multi-party training challenges posed by industry-specific siloed data, indicating a need for sophisticated fusion techniques [14].

Communication interfaces are integral in enabling secure and efficient interactions between clients and the server. FL communication protocols focus on minimizing bandwidth usage, safeguarding transmissions from adversaries, and addressing issues of reliability in cases of partial client dropout. Advances in gradient sparsification, quantization, and encryption frameworks like Secure Multiparty Computation (SMPC) ensure confidentiality while optimizing network overhead [20]. Despite progress, the implementation of these strategies often incurs trade-offs between computational burden and communication efficiency [11]. Emerging innovations in asynchronous communication frameworks and federated dropout-tolerant aggregation highlight future directions to support dynamic and resource-constrained FL environments [6].

The modularity of FL architectures enables flexible adaptation for diverse scenarios, from edge-device-driven IoT networks to domain-specific health informatics [21]. Nevertheless, the integration of robust components and interaction frameworks continues to present technical and practical challenges. Striking the balance between privacy guarantees, system scalability, and fairness remains an open research area where architectural refinement and design are critical for successful deployments.

Overall, the evolution of federated learning system components reveals the field's enormous potential alongside its nuanced constraints. Designing architectures that proactively address client heterogeneity, server dependencies, and secure communications while minimizing computational or communication trade-offs represents a paramount objective for the advancement of federated machine learning systems. Incorporating decentralized and hybrid paradigms, rigorously evaluated under large-scale experiments, offers a promising trajectory for future research in this domain.

2.2 Architectural Paradigms in Federated Learning

Federated learning (FL) architectures are tailored to address the dynamic complexities of decentralized systems, prioritizing scalability, efficiency, and privacy preservation. This subsection explores three core architectural paradigms—centralized, decentralized, and hybrid—illustrating how each design aligns with specific system requirements while addressing associated trade-offs and limitations. These paradigms provide a natural progression from the architectural components outlined in the previous subsection and lay the groundwork for understanding the operational workflows detailed in the following subsection.

Centralized architectures have emerged as the foundational approach in FL, with a central server orchestrating the collaborative training process. The server aggregates model updates from client devices and distributes the refined global model, ensuring a cohesive system workflow. This structure offers simplicity, broad utility, and ease of deployment, enabling algorithms like FedAvg to achieve substantial convergence efficiency in diverse practical scenarios [22], [23]. However, centralized systems are hindered by vulnerabilities such as single points of failure, network bottlenecks, and susceptibility to adversarial attacks on the central entity [24]. Moreover, as client participation scales, these systems face significant challenges, including increased communication overhead and latency, which undermine reliability and performance [25]. These weaknesses align with the broader concerns of centralized dependency discussed earlier, reinforcing the need to explore alternative paradigms.

Decentralized architectures eliminate central server reliance, fostering a peer-to-peer communication structure where clients collaboratively exchange updates to achieve model consensus. This approach enhances fault tolerance, systemic robustness, and privacy, as sensitive data no longer flows through a single aggregation point. Techniques like blockchain-based FL introduce secure and immutable mechanisms for consensus and aggregation, addressing the vulnerabilities associated with centralization [26], [27]. Despite these benefits, decentralized architectures face challenges in synchronization, computation, and scalability. The absence of a trusted coordinator complicates update consistency, while cryptographic methods (e.g., blockchain) can impose substantial energy and computational costs, limiting practicality at scale. Addressing these trade-offs demands the development of efficient peer-to-peer protocols and lightweight consensus algorithms [17].

Hybrid architectures integrate the strengths of centralized and decentralized paradigms to balance efficiency, scalability, and robustness. A prevalent configuration in this category is hierarchical FL, where intermediate nodes, such as edge servers or regional hubs, facilitate localized aggregation before final updates are sent to a central server. This design reduces communication burdens, alleviates latency, and improves scalability, particularly in systems involving geographically distributed clients [21], [28]. Furthermore, hybrid architectures enable domain-specific optimizations by tailoring aggregation processes to groups of similar clients, making them particularly effective in sectors like healthcare and IoT applications [21], [29]. Nevertheless, managing the increased complexity of hybrid systems requires careful handling of edge-to-cloud interactions, resource allocation, and system coordination.

Looking forward, addressing challenges such as scalability, trust, and heterogeneity will be critical for advancing FL architectures. Decentralized systems, for instance, may achieve broader adoption if lightweight, resource-efficient consensus protocols or blockchain alternatives can be standardized [20]. Similarly, hybrid systems capable of dynamically transitioning between centralized and decentralized modes based on contextual demands and resource availability could optimize efficiency and reliability. Furthermore, combining federated transfer learning with multi-level hybrid designs offers exciting possibilities for fostering collaboration among heterogeneous clients at scale [30].

In summary, the architectural paradigms of FL—centralized, decentralized, and hybrid—each present unique trade-offs between simplicity, robustness, and adaptability. Centralized architectures prioritize ease of deployment, while decentralized designs emphasize resilience and privacy, and hybrid approaches address scalability and domain-specific needs. Tackling key challenges such as operational complexity, scalability, and resource management while leveraging the synergies between these paradigms is pivotal to advancing FL systems. This structural flexibility paves the way for integrating more efficient workflows and applications, setting the stage for the operational dynamics discussed in the next section.

2.3 Processing Workflows for Federated Model Updates

The operational workflows of federated learning (FL), starting from local computation to global model aggregation, underline the core mechanism that allows distributed learning while preserving data privacy. This subsection provides a critical analysis of these stages, examines trade-offs in their design, and highlights challenges and emerging trends.

The FL workflow begins with **local model training** at each client device, where the data remains decentralized. This process involves performing multiple epochs of gradient-based optimization (e.g., Stochastic Gradient Descent) on the client's local data, aiming to minimize a local loss function. However, heterogeneity in data distributions across clients—non-independent and identically distributed (non-IID) data—complicates local training dynamics and can hinder the global model's convergence. Methods such as FedProx, which introduces a proximal term to limit client updates from straying far from the global model, have been proposed to address these issues [8]. Local computation efficiency is further challenged by resource constraints on edge devices, necessitating optimization techniques like adaptive mini-batching, mixed precision arithmetic, and computational offloading to higher-capacity devices [17].

Once local models are trained, clients send their updated parameters or gradients to a central server for **global aggregation**. Federated Averaging (FedAvg), one of the most widely adopted aggregation algorithms, computes the weighted average of updates based on client data size, ensuring scalability [31]. While effective, FedAvg assumes well-behaved clients, which may not hold under adversarial settings; malicious participants can submit poisoned updates, thereby corrupting the global model. Techniques like robust aggregation algorithms (e.g., Krum, Multi-Krum) aim to mitigate such vulnerabilities by filtering out anomalous updates based on similarity criteria [32]. Furthermore, privacy risks associated with transmitting updates, such as gradient inversion attacks, necessitate secure aggregation protocols like Secure Multi-Party Computation and Homomorphic Encryption, which add computational and communication overheads [33], [34].

The aggregation can be performed either synchronously or asynchronously. **Synchronous FL systems** require all

client updates within a round, introducing delays due to slower devices or clients with intermittent connectivity. This bottleneck can exacerbate communication inefficiencies in large-scale systems with diverse client participation [17]. To address this, **asynchronous methods** allow partial updates to be aggregated as they arrive, yielding faster convergence in heterogeneous networks. However, asynchronous approaches often struggle with stale updates that could degrade model accuracy. Techniques to re-weight updates using time-decay functions or sequential correction mechanisms have been proposed to alleviate these challenges [35].

The iterative nature of FL demands multiple global aggregation rounds, introducing communication efficiency as a pivotal bottleneck. Strategies such as update compression, gradient sparsification, and quantization aim to reduce the volume of shared updates without compromising model accuracy significantly [31], [36]. Recent hybrid approaches, merging communication-efficient methods with differential privacy mechanisms, strike a balance between reducing bandwidth usage and ensuring robust privacy guarantees [11], [37].

Emerging trends in FL workflows highlight the potential of **hierarchical aggregation**, wherein intermediate servers aggregate updates regionally prior to finalizing at the central server. This multi-tier approach reduces communication costs in geographically distributed FL systems, as proposed in healthcare and IoT applications [4], [21]. Blockchain-assisted decentralized FL frameworks further enhance security and reduce reliance on a single point of failure in the aggregation process, though they introduce latency and scalability concerns [16], [38].

Despite these advancements, significant open challenges persist. Addressing system heterogeneity effectively, minimizing communication costs for resource-constrained clients, and enhancing robustness against data and model poisoning remain critical areas for future research. The development of hybrid aggregation models that dynamically adapt based on client availability and computational resources offers a promising direction to ensure scalability without compromising on privacy and security guarantees. As federated systems evolve, advancements in decentralized and hierarchical workflows coupled with robust privacy-preserving techniques are expected to drive the next generation of federated learning systems.

2.4 Variants of Federated Learning

Federated learning (FL) has emerged as a highly versatile framework, driving privacy-preserving collaborative machine learning across diverse application domains while addressing critical challenges in decentralization and data handling. To align with varying data distributions, system architectures, and application requirements, several specialized variants of FL have evolved. This subsection categorizes and analyzes these variants comprehensively, exploring their foundational frameworks, inherent trade-offs, and use cases. By situating these paradigms within the broader FL workflows of local training and global aggregation, as discussed in the previous section, and their direct connection to application scenarios detailed in the following section, this discussion highlights the dynamic adaptability of FL.

Horizontal Federated Learning (HFL), the most conventional FL paradigm, applies to settings where participating clients share similar feature spaces but differ in data instances. This setup is prevalent in cross-device networks, such as smartphones or IoT systems, where the goal is to collaboratively train a model without centralizing client data. Centralized aggregation strategies, such as the FedAvg algorithm, form the backbone of HFL's scalability and adoption [31]. Nonetheless, HFL must contend with challenges posed by non-IID data distributions among clients, which can lead to suboptimal global model performance. Efforts to address these issues include the incorporation of regularized aggregation techniques [39] and locally adaptive mechanisms [40], which have significantly enhanced performance and convergence, especially in edge-device environments with constrained resources.

Vertical Federated Learning (VFL), in contrast, is tailored for scenarios where datasets across clients differ in feature spaces while sharing overlapping data instances. This variant is particularly effective in cross-silo environments, such as inter-organizational collaborations in domains like finance or healthcare. Leveraging cryptographic protocols, including secure multi-party computation (MPC) and homomorphic encryption, VFL facilitates joint model training while ensuring the privacy of client-specific data. Advanced approaches such as two-phase MPC frameworks [41] enhance both security and scalability by employing hierarchical or committee-based aggregation. This capability to combine disjoint yet complementary feature sets has cemented VFL's relevance in use cases requiring robust joint models, such as fraud detection or disease diagnosis. However, VFL's reliance on secure protocols often introduces significant computational and communication overheads that remain a pressing challenge.

Federated Transfer Learning (FTL) extends the boundaries of FL by addressing cases where both feature spaces and data instances differ across clients. FTL is particularly beneficial in domain adaptation tasks, allowing collaborations between institutions or systems with disjoint yet mutually beneficial objectives. Combining FL and transfer learning paradigms, this variant enables flexible knowledge sharing while maintaining data privacy. Techniques such as privacy-preserving feature alignment [42] and fine-tuned adapter integration with foundation models [43] have enabled FTL to excel in personalization and other resource-constrained domains. However, ensuring effective and meaningful knowledge transfer across heterogeneous datasets still represents a significant research opportunity.

Hierarchical Federated Learning (HFL) introduces a multi-tiered aggregation approach, integrating edge-to-cloud hierarchies to enhance scalability across geographically distributed systems. Clients within a particular region can first aggregate locally, with these intermediate results undergoing further aggregation at global servers. This multi-layer structure not only reduces latency but also improves system efficiency, making it well-suited for large-scale settings like smart cities or healthcare networks [17]. While hierarchical architectures alleviate communication bottlenecks in expansive networks, they also pose challenges in maintaining consistent privacy guarantees and mitigating discrepancies during multi-level aggregation.

Each FL variant embodies specific trade-offs, reflecting the diversity of application-specific requirements, system constraints, and environmental heterogeneity. **Horizontal FL** is ideal for simpler, homogeneous scenarios, while **Vertical FL** and **Federated Transfer Learning** extend FL's capabilities to more complex, fragmented, or domain-adaptive contexts. **Hierarchical FL**, with its focus on scalability, is a natural choice for large distributed systems but brings heightened coordination complexity. Emerging hybrid paradigms further enrich this landscape, bridging functionalities from multiple variants—for example, blending Federated Transfer Learning with Hierarchical FL to enable personalized yet scalable workflows across diverse environments.

In summary, the evolution of FL paradigms underscores their adaptability to a range of real-world complexities. Addressing interdisciplinary challenges will require leveraging insights from these individual variants while exploring synergies across them. A promising future direction involves the development of unified frameworks that seamlessly integrate techniques such as distributed optimization, secure model fusion, and multi-modal aggregation. These advancements will play a pivotal role in scaling federated systems to meet dynamic and often conflicting requirements, as discussed in subsequent sections focusing on applied domains and deployment strategies.

3 Privacy Threats and Vulnerabilities

3.1 Risks from Model Updates

Federated learning (FL) systems, while designed to preserve privacy by keeping raw data decentralized, are not immune to privacy risks. A central concern lies in the sharing of gradients or model updates during the collaborative learning process. Such shared updates, despite originating from local computations, can inadvertently leak sensitive information about the underlying training data when exploited by adversaries. This subsection examines privacy risks arising from model updates, focusing on gradient inversion attacks, model reconstruction vulnerabilities, and overfitting-induced data leakage.

Gradient inversion attacks represent one of the most direct threats to FL's privacy-preserving intentions. These attacks leverage the gradients transmitted by clients to reconstruct private training inputs. As shown in [37] and [12], gradients encode rich information about the original data, particularly in deep learning settings where high-dimensional gradients capture detailed patterns. Mathematically, given a gradient $\nabla\mathcal{L}$, an adversary minimizes the difference between the reconstructed gradient $\nabla\mathcal{L}_{\text{reconstructed}}$ computed on a synthetic input x' and the received gradient by solving an optimization problem, such as $\min_{x'} ||\nabla\mathcal{L}(x';\theta) - \nabla\mathcal{L}||^2$, where θ represents the model parameters. This optimization often results in the recovery of data characteristics, enabling direct privacy breaches. While these attacks are computationally intensive, recent advancements in reconstruction algorithms have improved their efficiency, making them more viable in real-world scenarios [44].

Model reconstruction vulnerabilities add another layer of risk whereby attackers use aggregated model updates to reverse-engineer the global model or distill sensitive properties of participants' datasets. Particularly in vertical federated learning, where features are partitioned across entities, these models can inadvertently reveal correlations between feature splits, as discussed in [19]. By systematically analyzing periodic updates, adversaries can identify unique patterns or infer private feature values. This is especially concerning in domains like healthcare or finance, where inferred properties can lead to severe privacy violations [29].

Overfitting and data memorization risks further exacerbate privacy threats. Overfitted models tend to memorize specific client data points, particularly in personalized FL settings or in cases of highly non-IID (non-Independent and Identically Distributed) data distributions. As highlighted in [9], memorized information may inadvertently surface during model inference, enabling membership inference attacks. These attacks determine whether specific data points were present in a model's training dataset by analyzing its responses to crafted queries. Moreover, attackers exploiting overfitting gain an asymmetric advantage in exploiting vulnerabilities in small or poorly generalized datasets [2].

Defensive mechanisms to mitigate these risks include cryptographic techniques such as secure multiparty computation (SMPC) and homomorphic encryption (HE), which obscure gradients and prevent direct access to raw updates [45]. Additionally, differential privacy (DP)-enabled FL frameworks inject calibrated random noise into model updates to reduce the likelihood of data reconstruction attacks. These techniques safeguard training inputs against gradient inversion without excessively compromising model utility, as shown by [11]. However, achieving an optimal trade-off between privacy and utility remains a prominent challenge in adopting DP-based solutions, particularly in decentralized settings with ultra-low resource availability.

Emerging hybrid solutions that combine multiple privacy-preserving techniques are gaining traction. For instance, "crypto-aided" differential privacy integrates encryption techniques with noise injection to bolster defense mechanisms while mitigating individual limitations [34]. On the other hand, adaptive gradient obfuscation, which dynamically adjusts noise levels based on model sensitivity, represents another promising line of defense, as outlined in [46].

Despite progress, numerous open questions persist. Future research must address balancing communication efficiency and rigorous privacy guarantees, particularly in adversarial contexts or under evolving threat models. Moreover, optimizing FL systems to scale while preserving privacy across varying domains remains an unmet challenge, especially for resource-constrained applications like IoT or edge computing [21]. As these challenges are tackled, efforts that align advances in privacy-preserving techniques with practical deployment will shape the future trajectory of secure federated learning systems.

3.2 Adversarial Threats from Participants

Federated learning (FL) relies on collaboration among participants who may not trust each other, aiming to train machine learning models without sharing raw data. However, the presence of adversarial participants poses profound

risks to both the integrity and security of FL systems. This subsection delves into various adversarial threats, including data poisoning, model poisoning, and Byzantine attacks, which target either the global model's performance or the overall functionality of the training process.

Adversarial actors can launch **data poisoning attacks** by deliberately corrupting their local datasets to degrade the global model's performance. Examples of such attacks include label-flipping, where adversaries mislabel data points, and feature poisoning, which involves altering feature values to bias the model toward erroneous outputs [32], [44]. The effectiveness of these attacks is exacerbated in FL due to the distributed and non-IID nature of client data, which complicates the identification of anomalous or malicious inputs. For instance, label-flipping exploits the imbalances in non-IID datasets, allowing poisoned data to disproportionately influence the global model [7], [8]. Although contemporary defenses, such as robust aggregation techniques and anomaly detection methods, have shown promise, they often fall short in scaling to highly diverse or large-scale FL systems, especially when increased client participation introduces greater communication and computational overhead [1], [2].

In **model poisoning attacks**, adversaries manipulate their locally trained model updates before submitting them for aggregation, directly influencing the global model's behavior. This includes targeted attacks, such as embedding backdoors into the global model. Backdoors are designed to remain dormant until specific triggers activate them, causing malicious behavior in the model. Such sophisticated attacks exploit FL's reliance on the aggregation of client updates, allowing even a single well-strategized malicious participant to introduce subtle biases without immediate detection [32], [44]. For example, under the "little is enough" principle, attackers carefully craft infrequent but impactful updates to bypass standard anomaly detection mechanisms, embedding vulnerabilities into the global model over time [39]. Prevailing defenses, including differential privacy and secure aggregation frameworks, address some aspects of model poisoning but often at the expense of reduced model accuracy or increased computational complexity, limiting their practicality in resource-constrained environments [2], [34].

Another prominent adversarial risk is posed by **Byzantine failures**, where malicious or faulty participants send arbitrary updates to disrupt the FL training process. Unlike data or model poisoning, Byzantine threats do not aim to achieve a specific outcome but instead focus on degrading the overall performance or stability of the system. These adversaries can act randomly or coordinate their efforts, making detection and mitigation significantly harder. Byzantine-resilient aggregation algorithms, such as Multi-KRUM, attempt to filter out anomalous updates by identifying statistical outliers, yet they face challenges in highly heterogeneous systems or scenarios with strong collusive adversaries [44], [47]. Additional defenses, such as redundancy mechanisms and gradient masking strategies, have shown partial success. However, these methods often rely on stringent assumptions about adversary behavior and require synchronized contributions from participants, which may not hold in real-world FL settings [6], [48].

The continuously evolving strategies of adversarial participants highlight the increasing sophistication of threats in FL. For instance, adversaries can exploit cross-silo systems, where a smaller number of clients makes the system particularly sensitive to the actions of a single malicious participant [4]. Moreover, coordinated attacks involving collusion among multiple adversarial clients exacerbate the risks, circumventing many anomaly detection techniques that typically assume independent, uncoordinated actions from adversaries [1]. These challenges underline the urgency of designing dynamic, adaptive defense mechanisms capable of responding to diverse and evolving adversarial strategies.

Going forward, hybrid defense mechanisms that combine cryptographic techniques with statistical robustness promise significant improvements. For example, methods like crypto-aided differential privacy can simultaneously protect privacy while mitigating adversarial manipulations [11], [34]. Similarly, the development of adaptive anomaly detection frameworks that leverage machine learning to dynamically identify adversarial behaviors presents a promising avenue to enhance robustness without excessive constraints. Importantly, cross-disciplinary collaborations integrating FL with advancements in security, privacy, and fairness will be crucial to designing resilient federated ecosystems [2], [37].

This nuanced understanding of adversarial threats reflects the delicate balance required to achieve privacy, robustness, and scalability in federated learning systems. Addressing these vulnerabilities will be critical to ensuring the long-term effectiveness and trustworthiness of FL deployments across diverse application domains.

3.3 Membership and Inference Attacks

Membership inference and other inference attacks represent significant threats to the privacy-preserving guarantees of federated learning (FL). These attacks aim to uncover specific information about the training data, such as whether a particular data point was used in training (membership inference) or deducing private attributes of the data (property and feature inference). This subsection explores these attack mechanisms, analyzes their impact on FL systems, evaluates defenses, and discusses potential research directions in mitigating such threats.

Membership inference attacks exploit the overfitting tendencies and unintended memorization of data by machine learning models to determine whether a specific example was part of the training set. In the federated learning paradigm, the iterative sharing of model updates exacerbates this threat, as the updates encode information about the gradients and loss landscapes derived from client datasets. These shared updates can leak subtle details about the data distributions, enabling adversaries to perform targeted inference [32], [37], [44]. Formally, membership inference is modeled as a binary classification problem, where the attacker evaluates the likelihood $P(x \in \mathcal{D}_{\text{train}}|\theta)$, with x being the data point of interest and θ the model parameters or shared updates. Attackers often rely on shadow models that simulate the victim model's behavior to infer membership [32], [44].

In addition to membership inference, property inference attacks aim to identify statistical properties or attributes of client data. For example, an adversary might deduce whether specific demographic traits dominate a client's dataset or infer sensitive labels [29], [34]. Unlike membership inference, property inference does not require the attacker to isolate individual data points; instead, it leverages discrepancies in model gradients or parameters to recover aggregate information. This is particularly concerning in cross-silo FL, where client datasets may correspond to identifiable institutions like hospitals or financial entities, potentially revealing sensitive operational trends [4], [29].

Feature inference attacks represent another emerging threat, particularly in vertical federated learning (VFL). In such setups, input features are partitioned among parties, requiring collaboration to train the model. The feature inference threat arises as adversaries deduce a participant's private feature set through intermediate computations shared during the training process. This vulnerability is amplified when feature aggregation mechanisms, such as secure multiparty computation, are improperly implemented [8], [49]. While cryptographic techniques like homomorphic encryption mitigate some risks by masking features, high computational overhead limits widespread adoption [33].

Attack efficacy depends on multiple factors, including the model's complexity, the training data's heterogeneity, and the extent of adversarial knowledge. For instance, federated models trained on non-IID datasets have been shown to be particularly vulnerable to inference attacks due to uneven contributions to global updates [8], [50]. In addition, attackers with auxiliary knowledge, such as partial access to client data or side-channel information (e.g., timing or update size), can further enhance inference success rates [32], [44].

Emerging defenses against inference attacks primarily focus on differential privacy (DP) and secure aggregation mechanisms. By injecting noise into model updates, DP effectively reduces the signal-to-noise ratio accessible to adversaries, making precise inference more challenging [11], [37]. However, selecting appropriate privacy budgets under DP remains a challenging trade-off between model accuracy and privacy protection [11], [51]. Secure multiparty computation and homomorphic encryption provide robust guarantees against feature leakage but are computationally expensive and may introduce scalability issues in large-scale FL systems [33], [51]. Other mitigation strategies include adversarial training, which strengthens the model's resilience against inference attacks by simulating adversarial scenarios during training, and adaptive aggregation schemes that mask sensitive updates through obfuscation [44], [52].

Despite recent advancements, inference attacks on FL remain an active area of research, with attackers continuously evolving to exploit subtler vulnerabilities. Future research should focus on hybrid defense frameworks combining cryptographic, statistical, and adversarial resistance techniques to balance privacy, efficiency, and scalability. Additionally, robust evaluation metrics and standardized benchmarks are needed to assess the effectiveness of these defenses across diverse attack scenarios [36], [44]. Ultimately, mitigating membership and inference attacks is foundational for ensuring the broader acceptance of federated learning in privacy-sensitive domains.

3.4 Communication and Aggregation Vulnerabilities

Communication and aggregation in federated learning (FL) are foundational to its functionality, enabling the collaborative training of models while preserving data decentralization. However, these critical processes are fraught with privacy vulnerabilities that adversaries can exploit during both the communication of updates and the aggregation of model parameters. The sensitivity of these stages, coupled with the inherently distributed nature of FL, introduces significant challenges in ensuring robust privacy preservation.

A primary risk in communication arises from adversaries engaging in communication snooping. During the transmission of model updates or gradients, inadequately secured communication channels can expose sensitive information. For instance, gradient inversion attacks have demonstrated how private training data can be reconstructed from transmitted updates [53]. Even when encryption protocols are employed, timing-based side-channel attacks may still infer sensitive details by analyzing metadata such as the size, frequency, or timing of updates [54]. These risks are especially pronounced in FL setups involving resource-constrained devices, where the use of lightweight cryptographic protocols often leaves systems vulnerable to sophisticated adversarial techniques [55].

The aggregation process, integral to combining updates into a global model, also presents critical vulnerabilities. Secure aggregation schemes are intended to mask individual client contributions, yet improperly designed or implemented protocols may inadvertently leak aggregated statistics that adversaries can exploit to infer private attributes. For example, gaps in secure aggregation designs can lead to reconstruction attacks if model parameters are not appropriately verified [56]. Moreover, a malicious central server could abuse its privileged position to infer sensitive details under the guise of aggregation, posing additional risks to client privacy [54].

Collusion among clients adds another layer of complexity to aggregation vulnerabilities. In scenarios where subsets of clients conspire with each other or with the central server, adversaries could correlate shared updates across multiple rounds to infer sensitive details. This becomes particularly problematic in settings where differential privacy mechanisms or fragmented updates—introduced to enhance privacy—further compromise defenses by unintentionally amplifying the risk of collusion-based inferences [57]. Such scenarios underscore the pressing need for robust aggregation protocols that balance privacy preservation with the scalability demands of FL systems.

Efforts to mitigate these vulnerabilities have focused on advancing secure aggregation and communication protocols, though these solutions often involve trade-offs. For instance, secure aggregation techniques like LightSecAgg enhance resilience to participant dropout and improve computational efficiency but remain susceptible to side-channel risks [55]. Alternatively, homomorphic encryption allows computation on encrypted data, providing robust guarantees, albeit at the cost of significant computational and communication demands, which may be prohibitive

in resource-constrained environments [41]. Differential privacy introduces calibrated noise to safeguard privacy but can degrade model utility, especially when applied to non-IID or unbalanced datasets [58].

Emerging hybrid approaches aim to overcome such limitations by combining multiple defense mechanisms. For example, dynamic update masking and selective parameter encoding techniques offer promising directions for ensuring efficient and secure communication while preserving metadata privacy [59]. Similarly, decentralized architectures eschew reliance on a central server, mitigating central server vulnerabilities but introducing challenges in coordinating across clients [60]. Verification frameworks like EIFFeL further enhance security by ensuring the integrity of updates while maintaining privacy guarantees, indicating the importance of holistic, multi-faceted approaches [61].

Future advancements must focus on reducing communication overhead and enhancing privacy through optimizations like sparsification, gradient compression, and resource-efficient synchronization protocols [62]. Adaptive secure aggregation schemes and scalable, privacy-preserving mechanisms will be critical to navigating the increasingly sophisticated landscape of adversarial strategies. By addressing these challenges, FL systems can achieve a balance between privacy, efficiency, and scalability, strengthening their viability for widespread adoption across diverse and privacy-sensitive domains.

3.5 Emerging Threats in Privacy Breaches

Emerging privacy threats in federated learning (FL) represent a critical and sophisticated evolution of attack vectors, posing significant challenges to privacy-preserving mechanisms. These threats exploit both underexplored vulnerabilities and advanced adversarial capabilities, often bypassing traditional defenses. This subsection delves into these advanced threats, evaluates their mechanisms, and explores the implications for future mitigation strategies.

One concerning avenue for emerging attacks involves *backdoor injections*, where adversaries maliciously embed triggers into the global model during training. These triggers, designed to activate specific undesirable behaviors, can remain dormant during standard testing, thus evading detection. For instance, targeted backdoor strategies leverage the iterative aggregation process in FL to subtly introduce corruptions that compromise the overall model's functionality, particularly in personalized tasks [32]. Recent studies demonstrate that backdoor attacks can achieve high success rates even under strict aggregation protocols, particularly when data heterogeneity among clients allows adversaries to camouflage malicious updates amidst legitimate updates [32]. Notably, such attacks underscore the limitations of commonly deployed defenses like robust averaging, as these methods often fail to discriminate subtle adversarial contributions while maintaining utility for the global model.

Another emerging threat involves *cross-silo data correlation*. In this scenario, adversaries exploit shared model updates from multiple institutions or organizations to deduce sensitive patterns. For instance, in vertical federated learning (VFL), where features are distributed across institutions for common samples, the correlation between intermediary computed results from distinct parties can leak sensitive user properties or relationships without direct access to raw data [63]. This privacy leakage stems from statistical similarities in intermediate results, amplified when multiple parties collude to reconstruct more granular data patterns. The risk becomes acute in healthcare-focused FL systems, where these correlations can expose proprietary patient data while operating under the premise of privacy [4].

Timing-based side-channel attacks further reveal the multifaceted vulnerability landscape in FL. These attacks infer private characteristics by analyzing ancillary information, such as the size of model updates, communication frequencies, and timing patterns, rather than the content itself [64]. Such attacks exploit inconsistencies in communication across participants, especially in decentralized setups with resource-constrained clients. For example, subtle variations in communication latency or update compression ratios can inadvertently disclose client-specific data characteristics, including sensitive information about their underlying datasets. While cryptographic approaches like homomorphic encryption and secure multiparty computation offer protections for the data content, they are often insufficient against these timing leakages, which remain an open challenge in deployment at scale [33].

The development of adversarially-driven *gradient inversion attacks* exemplifies another frontier in FL privacy concerns. These attacks aim to reconstruct private client data by exploiting shared gradients during the optimization process. Advanced techniques utilizing generative models or optimization-based reconstructions further enhance attackers' ability to decode high-resolution raw data from gradients, revealing user-sensitive patterns even when data transformations, such as differential privacy, are applied [11]. This type of attack is exacerbated under non-IID data conditions, where gradient variations across clients provide additional context for adversarial inference [8].

As adversarial tactics evolve, hybrid threats leveraging combinations of the above attack vectors have begun to surface. For example, adversaries may combine timing-based side-channel analysis with backdoor injection strategies to enhance stealth and impact, creating broader attack spaces difficult to counter with current single-layer defenses [32]. Moreover, as FL frameworks adopt transfer learning or hierarchical learning to scale across verticals involving IoT and edge devices, these hybrid threats are likely to proliferate due to the increased complexity in communication networks and aggregation protocols [21].

Addressing these threats requires innovative and rigorously tested countermeasures. Techniques such as adaptive differential privacy and resilient cryptographic enhancements show potential, yet their deployment must consider the system-level trade-offs between robustness, computational overhead, and communication efficiency. Additionally, exploring adversarial-resistant aggregation methods and leveraging decentralized architectures may mitigate collusion-oriented and correlation-based attacks [65]. Future research should also prioritize the integration of cross-layer defenses, combining cryptographic solutions with anomaly detection algorithms and maintaining scalability to ensure practical adoption across FL systems [66].

In summary, emerging threats in federated learn-

ing encapsulate highly sophisticated attack methodologies that exploit systemic weaknesses at various levels of FL frameworks. While advancing defense mechanisms shows promise, these threats underscore the need for continuous innovation in safeguarding privacy in federated machine learning ecosystems.

4 Privacy-Preserving Techniques in Federated Learning

4.1 Differential Privacy Mechanisms in Federated Learning

Differential Privacy (DP) has emerged as a cornerstone in the quest for mathematically provable privacy guarantees in federated learning (FL), offering robust protections against information leakage from shared outputs, such as gradients or model parameters. By introducing calibrated random noise into the training process or the communication of updates, DP ensures that the inclusion or exclusion of any single data point has a limited and formally quantifiable impact on the model's output. In federated learning, this property aligns with the goal of protecting sensitive client information while enabling collaborative model training across distributed datasets.

At its core, differential privacy ensures that for any two adjacent datasets $\mathcal{D}$ and $\mathcal{D}'$ differing in a single data point, the probability of any particular output remains nearly indistinguishable: $P(M(\mathcal{D}) \in \mathcal{O}) \leq e^{\epsilon} \cdot P(M(\mathcal{D}') \in \mathcal{O}) + \delta$, where ϵ is the privacy budget (quantifying privacy loss) and δ captures the probability of larger deviations. In FL, the integration of DP entails the introduction of noise either at the client-side (local DP) or centrally at the server during aggregation (central DP). These methods embody distinct trade-offs between privacy guarantees, communication overhead, and model utility.

Local Differential Privacy (LDP) applies noise directly to individual client updates before transmitting them to the server. LDP has the advantage of decentralized privacy, ensuring that no raw or exact gradients leave the client device. However, excessive noise can degrade model utility, particularly in high-dimensional settings common in deep learning. Studies incorporating LDP mechanisms, such as locally private gradient perturbations, have shown that the use of large privacy budgets ϵ minimizes utility loss but weakens the privacy guarantees [13], [37]. The scalability of LDP remains another concern, as maintaining acceptable accuracy under strict privacy budgets across numerous clients proves challenging.

Conversely, **Central Differential Privacy (CDP)** introduces calibrated noise during aggregation at the server, masking the contributions of individual clients. This centralized approach benefits from reduced noise levels compared to LDP, as aggregation naturally averages individual noise contributions, thus maintaining better model utility. However, CDP's reliance on a trusted server represents a potential vulnerability if the server is compromised or fails to enforce stringent aggregation protocols [37], [67]. Cryptographic techniques such as Secure Multiparty Computation (SMPC) or Homomorphic Encryption (HE) are often employed in conjunction with CDP to mitigate these risks [47].

Emerging approaches explore **adaptive DP mechanisms** within FL to balance privacy and utility dynamically. For instance, noise levels can be adjusted based on the sensitivity of the data or the specific training epoch, with higher noise early in training to obfuscate sensitive updates and lower noise as convergence progresses [11], [29]. These strategies aim to mitigate the trade-offs inherent in static noise injection methods by optimizing privacy contributions over the course of learning, particularly in non-IID settings where heterogeneity exacerbates sensitivity [7].

Practical deployment of DP in FL systems also requires consideration of resource constraints. Techniques such as gradient clipping prior to noise addition are prominent, as they ensure that individual updates adhere to bounded sensitivity, allowing differential noise calibration without compromising computational efficiency. However, the interplay between clipping thresholds and the magnitude of added noise often induces additional degradation in learning performance, especially in federated systems with significant client heterogeneity [21], [35].

While differential privacy has achieved significant strides in mitigating inference attacks and membership inference risks in FL [37], it is inherently vulnerable to specific challenges. For example, DP mechanisms are less effective in defending against property inference attacks, where adversaries seek aggregate characteristics of client datasets rather than individual records [13], [44]. Additionally, the lack of standardized metrics for evaluating the privacy-utility trade-offs under DP implementations in diverse FL settings highlights the need for further research [11].

Looking ahead, hybrid frameworks combining DP with complementary privacy-preserving techniques such as SMPC and HE hold significant promise. These approaches can address CDP's reliance on trusted aggregators while enhancing utility and scalability beyond what LDP alone can achieve [45]. Moreover, context-aware privacy strategies tailored to domain-specific FL applications—such as healthcare and IoT—offer exciting directions for fine-grained privacy management [16], [67]. To further the adoption of DP mechanisms in practical FL systems, future efforts should focus on adaptable noise-calibration algorithms, realistic benchmarking tools, and formal guarantees demonstrating resilience against emerging attack vectors.

In summary, differential privacy represents a pivotal element of federated learning's privacy-preserving arsenal, though its real-world efficacy hinges on navigating the intricate trade-offs between privacy, utility, and scalability. By synergizing DP with advanced cryptographic safeguards and adaptive techniques, the next generation of FL systems can offer robust, provable privacy guarantees without compromising collaborative machine learning's transformative potential.

4.2 Cryptographic Techniques for Privacy in Federated Learning

Cryptographic techniques play a pivotal role in safeguarding privacy within federated learning (FL) environments, ensuring data confidentiality during the collaborative model training process. These methods complement differential privacy and anonymization techniques by ad-

dressing privacy concerns during computation and communication, particularly in cases where adversaries may intercept intermediate computations or communication channels. This subsection critically explores core cryptographic approaches, including homomorphic encryption (HE), secure multiparty computation (SMPC), and hybrid cryptographic techniques, analyzing their strengths, limitations, and practical implications in federated learning.

Homomorphic encryption (HE), a cornerstone in cryptographic privacy preservation, enables computations directly on encrypted data, ensuring that local model updates are encrypted before transmission. This allows the central aggregator to perform operations (e.g., summations) without accessing plaintext updates, thereby aligning with the overarching objective of protecting participant contributions. For instance, HE methods integrated into frameworks like FedML-HE [33] achieve strong privacy guarantees by rendering individual contributions entirely opaque. However, while HE eliminates risks of direct information leakage, its computational overhead poses a significant barrier to scalability. Techniques such as selective encryption of sensitive model parameters help mitigate these overheads while maintaining privacy protection [33]. Nevertheless, large-scale deployments for complex models still face high latency and memory demands, especially in resource-constrained FL environments [68]. Future research could focus on compact encryption schemes that better balance privacy, scalability, and computational efficiency.

Secure multiparty computation (SMPC) provides another robust cryptographic approach, enabling multiple parties to jointly compute a function over their inputs while keeping those inputs private throughout the process. SMPC-based aggregation techniques, such as additive secret sharing, distribute encrypted model updates across multiple non-colluding servers, ensuring privacy as long as at least one server remains honest. This capability makes SMPC particularly effective in cross-silo FL settings, where fewer parties and higher trust among participants often prevail [69]. Furthermore, lightweight SMPC protocols have made strides in reducing communication overhead, making them more attractive for real-time applications [47]. However, when applied to the large cross-device FL networks characteristic of consumer settings, SMPC encounters challenges such as high communication costs and reliance on reliable network infrastructure. Designing asynchronous SMPC protocols or integrating SMPC with communication-efficient techniques represents a promising path forward in addressing these challenges [11].

Hybrid cryptographic approaches combine the strengths of HE and SMPC to address their individual limitations. By using HE for encrypting sensitive model parameters and SMPC for securely aggregating the remaining updates, these frameworks optimize computational and communication efficiency while maintaining robust privacy guarantees [44]. Moreover, recent innovations explore integrating differential privacy into cryptographic schemes, adding noise to encrypted model updates to bolster protection against reconstruction attacks [11]. While hybrid methods offer a layered approach to privacy, they introduce increased implementation complexity, requiring careful calibration to avoid added inefficiencies or security vulnerabilities.

Despite their transformative potential, cryptographic techniques in FL face significant challenges in balancing scalability, efficiency, and robustness. One critical issue is their compatibility with gradient compression, a technique essential for reducing bandwidth costs in large-scale deployments. Similarly, enhancing fault tolerance to handle cryptographic failures, such as server compromises in SMPC scenarios, remains a pressing concern. Future research directions also include leveraging decentralized cryptographic methods—such as blockchain-based aggregation—for greater transparency and resilience in federated learning systems [26].

In summary, cryptographic techniques form an indispensable layer in the privacy architecture of federated learning, complementing differential privacy and anonymization strategies by providing robust safeguards during computation and communication. Homomorphic encryption, secure multiparty computation, and hybrid approaches present nuanced trade-offs between security, scalability, and efficiency. As federated learning expands to encompass more diverse deployment environments, ongoing innovation in lightweight cryptographic protocols, fault-tolerant mechanisms, and hybrid frameworks will be essential in building scalable and reliable FL systems capable of handling real-world complexities.

4.3 Anonymization and Data Obfuscation Strategies

Anonymization and data obfuscation strategies in federated learning aim to prevent direct or indirect identification of individual client contributions, focusing on minimizing risks of privacy breaches without compromising the global model's utility. Unlike cryptographic solutions such as secure multiparty computation (SMPC) and homomorphic encryption (HE), which directly ensure confidentiality during computation, anonymization and obfuscation techniques reformulate data or metadata itself to obscure identifiable elements, offering a complementary layer of protection. These strategies are particularly critical in scenarios where gradients or model updates, typically shared during training, can leak sensitive information via reconstruction or inversion attacks [37], [44].

One prominent approach is anonymity-enhancing aggregation protocols, where client contributions are aggregated in a way that decouples individual updates from their origins. Secure aggregation protocols exemplify this, ensuring that the server receives only the aggregated model updates without access to individual client updates. By leveraging cryptographic masking techniques, secure aggregation protocols preserve the privacy of individual updates even if the server colludes with other parties [34]. Protocols like this inherently reduce explicit identification risks, yet they face overhead and scalability challenges as resource demands increase under large-scale federated learning systems [23].

Synthetic data generation is an emerging strategy that utilizes artificially created datasets to obfuscate sensitive data properties while maintaining fidelity for model training. A major advantage of these techniques lies in their flexibility: sensitive attributes or statistical features are replaced with synthetic counterparts that mimic data distributions. While effective against direct reconstruction attacks,

synthetic datasets may fail to preserve high-dimensional relationships critical to model accuracy, underscoring a trade-off between utility and privacy [4], [11]. Additionally, synthetic data methods risk introducing bias into the global model, particularly when sensitive features are inadequately captured or over-generalized [10].

Dynamic anonymity mechanisms present adaptive frameworks that modulate anonymity levels over time. For example, methods can adjust obfuscation intensity based on the sensitivity of individual updates, the aggregation stage, or the type of attacks most likely to occur during specific iterations of federated model training. Implementations such as dynamic noise injection allow on-the-fly adjustments, wherein noise proportional to data sensitivity is injected into gradients. While dynamic mechanisms offer enhanced flexibility compared to static solutions, their efficacy relies heavily on accurate sensitivity estimation, which may itself expose vulnerabilities if adversaries exploit these dynamic adjustments [31], [51].

Another critical concept is hiding metadata about client participation or model updates. Techniques that randomize or delay contributions complicate the ability of adversaries to associate updates with specific clients, mitigating risks like membership inference attacks. However, randomized approaches introduce latency, potentially slowing the convergence of shared models in cross-device settings where efficiency is paramount [17], [70].

Evaluating these anonymization and obfuscation methods requires a dual-focus on privacy benchmarks and model utility. Studies demonstrate that combining anonymization techniques with approaches like differential privacy can achieve stronger privacy guarantees at the cost of increased performance overhead [11]. Emerging trends also explore hybrid anonymization frameworks, integrating synthetic data with secure aggregation to address diverse privacy threats synergistically [71]. Nonetheless, remaining challenges include preserving fairness across heterogeneous client populations and mitigating trade-offs between computational scalability and privacy robustness.

In conclusion, anonymization and data obfuscation strategies enrich the privacy toolbox for federated learning by mitigating direct and indirect identification risks. While advancements in aggregation protocols, synthetic data techniques, and dynamic obfuscation offer substantial promise, future research must focus on harmonizing scalability, fairness, and utility amidst increasing system complexity. Observing practical implementations, such as in healthcare and IoT systems, could yield critical insights for bridging theoretical guarantees with real-world constraints [4], [21].

4.4 Emerging Hybrid Privacy Frameworks

Hybrid privacy frameworks in federated learning (FL) represent a strategic fusion of cryptographic, statistical, and distributed learning techniques designed to address privacy risks comprehensively while simultaneously ensuring computational efficiency and model utility. These frameworks seek to mitigate the limitations of singular approaches—such as cryptographic or differentially private techniques used in isolation—and tackle emerging threats in diverse and dynamic real-world deployments.

A significant innovation within hybrid privacy frameworks is the integration of cryptographic mechanisms with differential privacy to secure data during model aggregation. Differential privacy traditionally involves the addition of random noise to client updates or aggregated results to obscure sensitive information, but this practice often compromises model utility, particularly in systems characterized by high data heterogeneity [39]. Cryptographic methods, such as homomorphic encryption (HE) or secure multiparty computation (SMPC), ensure that computations can be performed on encrypted data without revealing its content, thereby safeguarding privacy without directly degrading the model's performance. Hybrid frameworks synergize these methods by applying fine-grained noise injection at pivotal stages of aggregation, striking a crucial balance between strong privacy guarantees and reduced utility losses [41], [55].

Another promising direction in hybrid privacy approaches is secure representation learning, where clients exchange highly abstract latent features rather than raw gradients or sensitive model parameters. Techniques like federated knowledge distillation leverage centrally available proxy datasets to distill aligned outputs from locally trained models into a cohesive global model. This reduces the risk of gradient inversion or reconstruction attacks by minimizing the exchange of sensitive information [72]. When combined with privacy-enhancing methods such as differential privacy, these frameworks show great promise in handling heterogeneous data distributions while also improving communication efficiency and reducing computational demands.

Vertical federated learning (VFL), a setting where feature spaces rather than datasets are partitioned among clients, adds an additional layer of complexity for maintaining privacy. To address this, hybrid frameworks customized for VFL often integrate SMPC with data obfuscation techniques. For example, encrypted intermediate computations are shared among parties, allowing them to collaborate while keeping individual feature distributions private [41]. Frameworks such as Federated Learning with Feature Anchors (FedFA) further extend this concept by aligning client-specific features and classifiers while navigating inconsistencies introduced by heterogeneous data. This combined approach of global feature alignment and secure computation demonstrates the potential to enhance learning accuracy without compromising privacy [73].

Despite these advancements, hybrid privacy frameworks face important challenges that must be addressed to ensure their scalability and efficiency. Cryptographic techniques like HE often require exponential computational resources as the number of participants or the complexity of models increases, which can hinder their implementation in large-scale FL systems [17]. Additionally, blending privacy mechanisms necessitates careful trade-offs between privacy strength, computational cost, and scalability. For instance, while differential privacy offers robust guarantees, higher noise levels can impair learning accuracy, especially in personalization tasks or systems requiring real-time responsiveness [23], [40].

Recent innovations point to potential pathways for addressing these obstacles. Lightweight cryptographic designs

such as LightSecAgg, which employs encoded aggregate masks to optimize secure aggregation, have demonstrated advancements in scalability and resilience against user dropouts, all while reducing computational overhead [55]. Dynamic privacy adaptation represents another promising trend, wherein hybrid frameworks adjust privacy parameters during training to reflect the sensitivity of data and the operational constraints of the system [30]. Moreover, the integration of these hybrid methodologies with federated foundation models opens exciting prospects for enabling privacy-preserving collaborations across domains using extensively pre-trained global architectures [74].

In summary, hybrid privacy frameworks in FL embody a paradigm shift aimed at uniting diverse privacy-preserving techniques to create robust ecosystems optimized for real-world applications. By combining the complementary strengths of cryptography, differential privacy, secure learning frameworks, and adaptive methodologies, these frameworks successfully address multifaceted privacy threats that arise in federated learning workflows. However, widespread adoption hinges on the resolution of key challenges related to computational efficiency, scalability, and resource optimization. As FL continues to expand across domains such as healthcare, finance, and the Internet of Things, hybrid privacy frameworks are poised to play a pivotal role in reconciling privacy with performance, paving the way for practical and secure federated ecosystems.

4.5 Privacy-aware Federated Personalization

Privacy-preserving federated personalization represents an essential frontier in federated learning (FL), targeting the tension between the desire for personalized model optimization and the need to protect clients' sensitive data. Federated personalization aims to tailor global models to individual client preferences or environments while maintaining privacy guarantees, especially in settings characterized by significant data heterogeneity.

The core challenge in federated personalization lies in reconciling the global knowledge captured through federation with the local idiosyncrasies of client data. Traditional FL approaches, such as Federated Averaging, often fail in highly heterogeneous environments due to the objective mismatch between global and client-specific tasks. As a result, privacy-aware strategies for personalization have emerged, enabling users to benefit from both collaborative training and individual model customization.

One prominent privacy-oriented approach is local model fine-tuning, wherein a global model serves as an initialization point for further adjustments using local data. This technique minimizes communication overhead while enabling model customization. However, local fine-tuning alone may lead to overfitting on sparse, non-IID client data, jeopardizing generalization. To address this, meta-learning algorithms such as Model-Agnostic Meta-Learning (MAML) have been adapted for federated contexts [9], [10]. These approaches allow the global model to learn a representation that is rapidly adaptable to each client, improving both personalization and stability. Though effective, meta-learning often incurs high computational costs, particularly for edge devices with limited resources.

Shared representation learning constitutes another important method by decoupling local task-specific parameters (e.g., classifiers) from feature extractors, which are shared across clients. This paradigm reduces communication requirements by centralizing updates related only to shared components. Methods such as those proposed in [75], [76] illustrate the potential of this approach, showing that abstracting global representations while isolating fine-grained local adaptations enables significant improvements in both privacy protection and utility. Nevertheless, shared representation strategies entail trade-offs between the model's alignment with local distributions and the complexity of disentangling shared versus client-specific parameters. Ensuring that classifiers derived from shared features generalize effectively across diverse client preferences remains an open challenge.

Meanwhile, privacy-preserving transfer learning and hybrid personalization frameworks, incorporating vertical FL concepts, are gaining traction. By leveraging pre-trained global models tailored for specific tasks, transfer learning enhances local utility while upholding privacy constraints, as demonstrated in [30]. Adaptive obfuscation mechanisms also enhance privacy by dynamically masking gradients or model updates based on the sensitivity of personalization requirements. Methods such as entropy-based obfuscation in [77] highlight the potential of these approaches to balance privacy and model performance.

Group privacy techniques have also emerged for enhancing personalization among clients with shared data distributions or preferences. These methods aggregate individual client contributions into group-level updates to obfuscate individual identities. The techniques detailed in [39] exemplify this approach, clustering clients based on data similarity to mitigate the adverse effects of statistical heterogeneity. However, determining optimal grouping strategies in resource-constrained or highly dynamic environments remains an unresolved issue.

Despite significant progress, privacy-aware personalization faces numerous challenges. For instance, ensuring robust formal privacy guarantees, such as differential privacy, while enabling effective personalization is non-trivial. Techniques like cryptographic frameworks or client-specific noise injection, as explored in [11], significantly enhance privacy but can weaken model effectiveness, particularly in resource-limited deployments. Moreover, personalization's computational and communication efficiencies must be carefully balanced against the constraints of edge devices, which vary considerably in terms of hardware and connectivity [78].

In the future, combining multiple privacy-preserving techniques into integrated frameworks may further advance federated personalization. Hybrid systems that merge cryptographic protocols with meta-learning, or privacy-aware representation learning with adaptive obfuscation, hold promise for achieving robust personalization while safeguarding the sensitive data of clients. Expanding benchmarks to include real-world evaluations of heterogeneous personalization systems, as recommended in [10], will be crucial for grounding these developments in practical applications. As FL continues to blend personalization with privacy, interdisciplinary collaboration integrating expertise

in statistical learning, cryptography, and system optimization will be pivotal for scaling its adoption across domains.

4.6 Privacy Metrics and Evaluation in Federated Learning

Effective evaluation of privacy-preserving techniques in federated learning (FL) is critical to understanding their robustness, utility trade-offs, and formal guarantees. Privacy metrics provide quantitative insights into the resilience of FL systems against potential attacks, while evaluation frameworks simulate adversarial behaviors to measure privacy leakage under practical conditions. This subsection explores the landscape of privacy metrics and methodologies, emphasizing their significance, limitations, and future implications.

Differential privacy (DP) serves as a cornerstone approach in evaluating privacy, providing well-defined mathematical guarantees against information leakage. Privacy loss is quantified through privacy budgets (e.g., ε-DP), balancing the trade-off between privacy and model accuracy. Noise injection techniques, such as Gaussian or Laplace noise added to model updates, form the foundation for achieving DP in FL systems [12], [79]. However, practical deployment of DP mechanisms involves nuanced challenges: excessive noise injection can degrade model performance, while insufficient noise compromises privacy. Adaptive differential privacy, where noise levels dynamically vary during training, has emerged as a promising strategy to mitigate this trade-off, particularly in scenarios with heterogeneous client data [80].

Evaluation of gradient inversion attack defenses further strengthens the privacy guarantees of FL systems. Gradient inversion attacks represent a significant threat, as adversaries can exploit shared gradients to reconstruct private training data [81], [82]. Effective defenses must withstand rigorous evaluations under white-box attack scenarios. Reconstruction risk, quantified through methods such as mutual information (MI) analysis, offers strong theoretical grounding. MI-based metrics enable direct quantification of relationships between aggregated updates and private data [83]. However, reconciling empirical findings with theoretical MI bounds remains an ongoing challenge, with recent innovations suggesting promising directions in aligning practical results with theoretical guarantees.

Membership inference attacks constitute another benchmark for examining privacy protections. These attacks aim to determine whether specific data points were present in a training dataset, posing unique challenges in evaluating privacy. Simulation-based evaluation frameworks factor in diverse adversarial settings and FL configurations—such as data heterogeneity, local batch sizes, or aggregation methods—enabling precise assessment of defenses such as gradient perturbation and secure aggregation [84], [85]. Techniques like DP inherently offer protection against such attacks, but empirical evidence emphasizes the need for hybrid approaches integrating secure multiparty computation (SMPC) or homomorphic encryption for enhanced resilience [80].

Adversarial simulation frameworks, including generative adversarial network (GAN)-based analyses, serve as practical complements to theoretical evaluations. These simulations test FL defenses against sophisticated threat models, including attribute inference and property leakage [86], [87]. Insights derived from such simulations have catalyzed advancements in feature-level obfuscation and black-box knowledge transfer frameworks, which strategically reduce the attack surface by limiting the granularity of shared updates [88].

Despite this progress, existing privacy metrics and evaluation frameworks reveal lingering gaps that constrain their reliability. For example, many metrics fail to account for adaptive adversaries who modify their strategies iteratively throughout training [89], [90]. Moreover, balancing stringent privacy guarantees with utility requirements remains a persistent challenge, especially in real-world deployments like healthcare or finance, where predictive performance often takes precedence [91], [92].

Future research must embrace holistic evaluation paradigms that integrate adaptive adversary modeling, hybrid defense combinations, and compliance with evolving regulatory frameworks. Novel techniques like federated unlearning, which seeks to eliminate client-specific contributions from trained models, demand the development of new privacy evaluation metrics to assess retraining costs and residual information leakage [93]. Furthermore, adaptive privacy metrics capable of accommodating various client populations, data modalities, and deployment scenarios will be critical for fostering scalable and reliable FL systems.

In conclusion, privacy metrics and evaluation frameworks are indispensable for advancing the state of privacy-preserving federated learning. By bridging the gap between theoretical guarantees and empirical validations and addressing emerging attack vectors, the field is well-positioned to establish robust, practical metrics that enhance security, preserve utility, and ensure scalability across diverse, real-world applications.

5 System Optimization for Efficiency and Scalability

5.1 Communication Efficiency in Federated Learning

Optimizing communication efficiency is paramount to the scalability and performance of federated learning (FL) systems, which are inherently constrained by the high frequency of data exchange between clients and servers. Unlike centralized machine learning, where computations are conducted on localized datasets, FL involves iterative updates of model parameters, often over constrained network environments, introducing challenges related to bandwidth usage, latency, and reliability. This subsection explores contemporary strategies for minimizing communication overhead in FL without compromising model quality, focusing on gradient reduction techniques, quantization, structured updates, and asynchronous protocols.

Gradient Sparsification and Compression: A significant proportion of communication inefficiency in FL arises from the size of model gradient updates. Gradient sparsification, which involves transmitting only the most significant gradient components and dropping the rest, alleviates the bandwidth burden. Techniques such as Top-k sparsification

have demonstrated substantial reductions in communication costs while maintaining acceptable convergence rates [21], [94]. Compression methods further extend these benefits, where techniques such as quantization and entropy coding represent gradients using fewer bits. For example, stochastic quantization approximates gradients to predefined levels, significantly reducing transmission size [35]. However, these approaches introduce trade-offs between compression ratios and gradient fidelity, potentially slowing convergence or degrading model accuracy.

Model Weight Quantization: Quantizing model weights into lower-precision formats, such as 8-bit fixed-point representations, has emerged as another critical strategy for communication efficiency. Experiments show that low-precision representations achieve competitive performance compared to full-precision FL systems while drastically reducing communication costs [35]. However, quantization is particularly sensitive to the distribution of data across clients, requiring adaptive schemes that dynamically adjust precision based on data variability to maintain the balance between accuracy and resource savings.

Structured and Layer-wise Updates: Structured sparsity techniques impose constraints directly on the model architecture, such as by transmitting updates to specific layers or components of the neural network. Layer-wise updates, for instance, prioritize transmitting only higher-level representations (e.g., gradients of the last few layers) while freezing lower-layer parameters. These approaches are particularly valuable for deep neural networks with millions of parameters but require careful tuning to avoid penalizing model generalization [2], [9]. While structured updates significantly alleviate communication overhead, the resulting reduction in gradient diversity transmitted to the server may exacerbate the effects of statistical heterogeneity, particularly in non-IID data scenarios.

Asynchronous Communication Protocols: Traditional FL models rely on synchronous updates, where all clients complete training and transmit gradients in lockstep. While straightforward, this approach is network-constrained and suffers from latency spikes when slow or resource-limited clients delay the aggregation process. Asynchronous protocols address this issue by enabling clients to upload updates independently, even as others continue their local calculations. For instance, staleness-aware aggregation methods accommodate delays by weighting contributions based on their timeliness [50]. However, asynchronous systems introduce potential challenges in achieving model fairness, as the contributions of slower or less-reliable clients may be unintentionally diminished.

Hybrid Approaches and Dynamic Optimization: Emerging research highlights the integration of multiple approaches to optimize communication efficiency in FL. For instance, combining gradient sparsification with adaptive quantization schemes can amplify the benefits of both techniques [21]. Similarly, structured updates can be selectively applied to resource-constrained clients while more capable devices transmit full gradients, introducing hybrid participation models [2]. Dynamic techniques that adapt communication strategies in real time, based on factors such as client resource availability, local data quality, and network conditions, also represent a promising future direction.

Despite significant progress, several challenges remain. Balancing communication efficiency with robustness to non-IID data is an open problem, requiring techniques that preserve the diversity of gradient updates without overwhelming bandwidth constraints. Additionally, while current research often focuses on reducing transmission volume, few works explicitly evaluate the energy efficiency of proposed methods, a critical factor for FL systems deployed on IoT and mobile devices [67]. Future advancements in communication-efficient FL will also increasingly depend on interdisciplinary innovations, leveraging advances in network design and distributed optimization algorithms.

In conclusion, reducing communication overhead is essential for the scalability of FL systems, particularly in dynamic, resource-constrained environments. Techniques such as sparsification, quantization, and asynchronous protocols present viable solutions, though their integration and adaptation to heterogeneous and large-scale deployments remain areas of active research. Ensuring that communication optimization approaches are evaluated holistically, considering their impact on convergence, utility, and energy consumption, will be critical for the continued progress of federated learning technologies.

5.2 Scalability Optimization for Large-Scale Federated Systems

Scalability in federated learning (FL) is critical to harnessing its broader potential to include millions of distributed devices while ensuring efficient communication, computation, and reliability. Building on strategies that optimize communication efficiency, scalability efforts specifically address the challenges of large-scale deployments, including handling heterogeneous client populations, minimizing training overhead, and maintaining robust model convergence in dynamic network environments. This subsection examines key strategies for enhancing scalability, focusing on hierarchical frameworks, cluster-based partitioning, and adaptive client participation techniques, while assessing their respective strengths and limitations.

Hierarchical Federated Learning Frameworks: Hierarchical frameworks provide a multi-layered solution to scalability by reducing the bottleneck at the central server. In this architecture, client devices communicate with intermediate nodes, such as edge servers, to perform local aggregations before sending compressed updates to the central server. By offloading a portion of the computational and communication burden to edge layers, hierarchical FL mitigates bandwidth demands and latency issues, making it particularly well-suited for mobile and IoT deployments. For instance, hierarchical systems have demonstrated significant improvements in resource efficiency and system throughput in such contexts [17]. However, the inclusion of multiple layers in the communication hierarchy introduces synchronization challenges. Global aggregation must reconcile updates across these levels to preserve convergence and accuracy. Techniques like combining federated averaging with localized aggregation have shown potential in addressing such concerns [35].

Cluster-Based Training: Another effective strategy to enhance scalability is cluster-based training, which organizes clients into groups based on criteria such as data

similarity, geographic proximity, or computational capacity. By reducing the heterogeneity of data distributions within each cluster, this approach facilitates faster convergence and allows for localized training optimizations. Techniques like weighted hierarchical clustering or k-means algorithms are often employed for grouping clients, followed by intra-cluster training and less frequent inter-cluster synchronization [66]. This minimizes the strain on central servers by limiting cross-cluster communication. Despite these benefits, defining robust clustering metrics remains a significant challenge, especially when dealing with non-IID data distributions across clients—a common scenario in FL. Studies focusing on federated learning in diverse silos highlight the importance of developing advanced federated optimization techniques to address such variations [7].

Dynamic Client Selection Protocols: To manage scalability in expansive FL systems, dynamic client selection protocols prioritize active and reliable participants for each training round, reducing communication costs and computational load. These protocols evaluate clients based on factors like network stability, data utility, and device resource availability to optimize participation. Adaptive sampling frameworks have demonstrated success in selectively including clients that contribute more effectively to the global model objectives, thereby enhancing convergence and system efficiency [95]. However, over-reliance on high-performing clients raises concerns about fairness, particularly for under-represented datasets that may be excluded due to stringent selection criteria. Additionally, integrating participant evaluation mechanisms into the FL workflow introduces additional computational overhead, which may counteract scalability improvements.

Elastic Federated Systems: Elastic federated learning systems further extend scalability by dynamically adjusting resource allocation and aggregation intervals based on the system's operational load. For example, asynchronous update mechanisms allow the processing of client updates as they arrive, rather than waiting for all clients to complete local training, thus accommodating devices with intermittent connectivity [96]. These systems also enhance communication efficiency by employing techniques such as gradient sparsification and selective parameter updates, as evidenced by contemporary frameworks that integrate sparsification and quantization for communication optimization [25]. However, asynchronous updates can lead to the issue of update staleness, requiring advanced aggregation techniques to mitigate its impact on model performance and fairness.

While these strategies represent significant progress toward enabling scalable FL systems, substantial challenges remain in real-world implementation. For example, addressing the heightened variability of non-IID data distributions across a larger client population is essential for achieving sustainable model convergence. Similarly, managing device heterogeneity presents difficulties in synchronizing contributions from resource-constrained devices alongside more capable ones. Future work can focus on hybrid approaches that blend hierarchical frameworks with adaptive clustering techniques, optimizing both scalability and accuracy. Additionally, decentralized strategies, such as blockchain-enabled federated aggregation, offer promising avenues for enhancing system robustness while maintaining scalability [26].

In summary, scalability is a linchpin for the practical realization of federated learning across extensive, diverse client ecosystems. By refining hierarchical communication architectures, leveraging intelligent clustering, and adopting dynamic participation protocols, researchers can address many of the current barriers to large-scale FL deployment. Continued innovation in these areas will pave the way for federated systems that operate seamlessly across millions of devices, balancing performance, privacy, and efficiency in various application contexts.

5.3 Strategies for Personalization in Federated Systems

Personalization in federated learning (FL) is an essential avenue for addressing the diverse needs of individual clients arising from heterogeneous data distributions, resource constraints, and unique application requirements. This subsection examines strategies that balance global model generalization with client-specific adaptation, ensuring improved user-centric performance while maintaining the integrity of the global federated system.

A prominent approach to personalization in FL is **client-specific fine-tuning**, where each client adapts the globally aggregated model to its local dataset. This method often involves freezing certain layers of the global model (e.g., feature extraction layers) while fine-tuning others (e.g., classifier layers). Such techniques have proven effective in leveraging globally learned representations while enhancing local performance. For instance, Federated Reconstruction frameworks have demonstrated that partial local updates, rather than full model refinements, yield efficient personalization while meeting communication and privacy constraints [97]. However, this approach assumes clients have sufficient computational and data resources to perform fine-tuning, which can be challenging in resource-constrained or highly fragmented environments.

An increasingly sophisticated strategy for achieving personalization is **meta-learning-based federated frameworks**, where the goal is to train a "model initialization" that can quickly adapt to any client's unique data using a limited number of local updates. Techniques such as Model-Agnostic Meta-Learning (MAML) have shown promise in this context, enabling global models to exhibit rapid adaptability to local non-IID data distributions [9]. These approaches, while computationally intensive during meta-training, outperform traditional fine-tuning in client environments with rapidly varying datasets, such as personalized medical diagnostics and mobile applications.

Another highly effective method is the **layer-wise personalization strategy**, which strategically personalizes only specific parts of the model architecture. For example, shallow layers (e.g., feature extractors) are often kept generalized across clients, while deeper layers closer to the application output (e.g., task-specific classifiers) are personalized. This hierarchical personalization strikes a balance by maintaining shared foundational knowledge while catering to client-specific nuances [50]. Furthermore, this approach reduces both computation and communication burdens compared to fully client-specific models.

To address systemic challenges, such as varying computational capabilities and heterogeneous client priorities, **personalized federated optimization algorithms** like FedPer and FedAvgM have gained traction. These algorithms incorporate additional regularization terms or multi-model optimization schemas to align the global model objectives with personalized local performance goals [10]. While these methods enhance personalization robustness, they require careful tuning to balance trade-offs between privacy, utility, and computational overhead—underscoring a key tension in personalization systems.

Another innovative development involves **multi-task learning paradigms** within federated systems, where related but distinct client tasks are jointly modeled. These frameworks rely on client clustering, where subgroups with similar data distributions are identified dynamically, allowing for localized personalized federated learning within those clusters [10]. Though effective in improving task convergence, dynamic clustering methods raise challenges in secure communication and computational scalability, which are critical in practical FL deployments.

Despite these advancements, challenges persist in achieving personalization at scale with robust privacy guarantees. **Differential privacy mechanisms** can conflict with fine-grained personalization by reducing the utility of noise-regularized client updates. Similarly, **cross-device personalization** in large-scale environments, such as Internet of Things (IoT) or edge computing networks, remains constrained by device heterogeneity and communications bottlenecks [21]. Emerging hybrid approaches, such as those combining **federated transfer learning** with privacy-preserving models, offer compelling opportunities to address these limitations. These frameworks enable secure adaptation of models to individual domains by leveraging smaller client-specific datasets within the overarching FL environment [10].

Future directions in personalized FL research are likely to involve advances in automated personalization, leveraging **reinforcement learning for dynamic adaptations**, and developing benchmarks for evaluating personalization performance under realistic privacy and resource constraints [10]. Furthermore, integrating personalization with adaptive privacy guarantees, and addressing fairness across heterogeneous client populations, will be crucial for fostering widespread, trustworthy FL adoption. These innovations hold the potential to bridge the gap between federated generalization and meaningful client-specific optimization in real-world applications.

5.4 Handling Heterogeneity in Client Systems

Heterogeneity in client systems is a fundamental challenge in federated learning (FL), arising from the diversity in hardware capabilities, connectivity, and the non-independent and identically distributed (non-IID) nature of client data. This subsection examines key approaches and recent advancements in mitigating these challenges to ensure consistent and scalable model performance across heterogeneous clients, while also evaluating their trade-offs and identifying avenues for improvement.

Client devices in FL often exhibit vast differences in hardware capabilities, ranging from resource-constrained IoT sensors to high-performance edge computing nodes. These disparities result in varying processing power, memory availability, and energy limitations. To address these challenges, adaptive resource-aware strategies have been developed. Techniques such as model pruning and quantization enable the deployment of lightweight models, tailored to the constrained resources of smaller devices, without significant loss in performance [98], [99]. These methods dynamically reduce model complexity by removing redundant parameters during local training or global aggregation. Similarly, parameter-efficient fine-tuning focuses on updating specific model components to lower computation and communication overheads, offering an effective pathway for resource-constrained scenarios [100].

Variability in network connectivity amplifies the heterogeneity dilemma, as clients may experience intermittent or inconsistent communication performance. Asynchronous aggregation frameworks have been proposed to alleviate this issue by enabling federated servers to incorporate updates at varying intervals from clients, regardless of their network conditions [60]. This framework is often complemented by efficient communication methods, such as sparsified updates and selective parameter sharing, which reduce bandwidth consumption while retaining critical model information for effective aggregation [59], [101].

Beyond hardware and connectivity constraints, data heterogeneity—manifesting as non-IID distributions across clients—is a deeply pervasive issue in FL. Non-IID data can lead to local models diverging from the global objective, slowing convergence and impairing generalization. Optimization algorithms such as FedProx and Scaffold mitigate these effects by introducing regularization terms that align local and global updates, improving convergence and stability under heterogeneous data settings [35], [39]. Moreover, novel strategies like FedFA enhance feature representation alignment and classifier consistency during local training, ensuring robust aggregation across clients [73]. Another promising solution, embodied in approaches such as FedDM, uses synthetic datasets to approximate the global loss landscape, allowing for more consistent updates in environments with high data variability [102].

Personalization also serves as a vital mechanism for tackling heterogeneity, offering a more flexible alternative to enforcing a unified global model. Personalized federated learning (PFL) methods focus on tailoring client-specific models that balance shared representations with localized optimizations. By decoupling global components, such as feature extractors, from more task-specific components like classifiers, PFL enables clients to benefit from shared knowledge while optimizing for their individual data distributions [39], [103]. Advanced aggregation mechanisms, leveraging similarity-based client clustering or model interpolation, further refine this balance, enhancing both generalization and personalization [104].

However, these solutions are not without trade-offs. Ensuring lightweight models remain accurate despite computational compromises, balancing fairness across clients with differing resource profiles, and mitigating biases introduced by non-representative or dominant clients remain ongoing challenges. Regularization-based methods like FedProx, for instance, demand careful hyperparameter tuning, which

adds complexity to heterogeneous systems [35]. Meanwhile, the rising demands of multi-modal federated learning and federated foundation models exacerbate systemic heterogeneity, as these approaches often require managing large-scale, complex models in environments with inconsistent hardware capabilities [74].

To navigate these constraints, future research should emphasize integrated strategies that combine complementary approaches. For example, unifying asynchronous aggregation schemes with feature-aligned personalization techniques or pairing lightweight models with dynamic resource-monitoring systems could enhance real-time adaptability while maintaining performance. Developing robust benchmarks tailored to heterogeneous FL environments would further support the optimization of these strategies, clarifying trade-offs and fostering a more systematic understanding of their application. Additionally, interdisciplinary advancements in distributed optimization and edge computing architectures present promising opportunities to address scaling challenges, paving the way for inclusive and adaptive federated learning systems. By addressing these multifaceted needs, the FL community can ensure consistent performance and equity across increasingly diverse client ecosystems.

5.5 Resource-Aware Implementation for Energy Efficiency

Resource-aware implementation in federated learning (FL) is critical for ensuring energy efficiency, particularly in environments with resource-constrained devices such as smartphones, Internet of Things (IoT) devices, and edge computing nodes. This subsection addresses strategies and methodologies aimed at minimizing energy consumption while preserving the efficacy of federated learning workflows. With the proliferation of FL deployments across diverse hardware ecosystems, optimizing computation, communication, and energy use has emerged as one of the most pressing system challenges.

A primary mechanism for improving energy efficiency is minimizing the computational overhead of local training tasks. For resource-limited devices, employing lightweight model architectures and inference schemes has shown promise in reducing processing power requirements. Techniques such as model pruning, quantization, and compact neural architectures have been widely employed to reduce training complexity [78]; for instance, pruning techniques remove non-essential parameters from local models while preserving task accuracy, making them suitable for low-power devices. These methods not only reduce the computational footprint but also extend battery life in energy-constrained environments. More advanced hybrid methods, such as combining sparsified gradient updates with model quantization [77], enhance computational efficiency while adapting to non-IID client data.

Another critical area of focus is a reduction in the number of communication rounds required for model updates. FL typically incurs significant energy use due to frequent exchanges of model parameters between the central aggregator and clients. Adaptive update strategies, such as those dynamically selecting communication intervals based on model convergence or client energy states, can alleviate this burden [66]. Techniques like asynchronous updates, which allow clients to upload parameters at varying frequencies, have demonstrated the potential to conserve energy without severely compromising system performance [35]. Additionally, improving communication efficiency through techniques such as federated distillation or sending synthetic data representations instead of full gradient updates can significantly reduce data exchange sizes, as illustrated in [77].

The scheduling and orchestration of client participation can also make profound impacts on energy efficiency. Selecting participants based on resource availability and energy constraints—referred to as energy-aware client selection [66]—ensures that devices with sufficient energy resources prioritize model training. Similarly, leveraging renewable energy-powered devices for participation in FL workloads has been proposed as a sustainable strategy for lowering the carbon footprint [105]. By dynamically matching client workloads to devices with lower operational energy costs, such approaches incorporate ecological considerations into FL system design.

At a system level, clustering and hierarchical aggregation provide another promising mechanism for improving energy efficiency. Layered architectures, where edge servers handle intermediate aggregations before sending updates to a central server, have significantly reduced long-distance transmissions in federated IoT deployments [21]. These frameworks enable energy savings at the global level by limiting the number of directly participating devices per communication round.

In addition, energy-efficient scheduling techniques consider both computational workloads and network states. For example, Carbon-efficient scheduling prioritizes updates from clients in low-latency environments or devices connected to reliable, low-energy communication networks [50]. Beyond static optimizations, dynamic monitoring systems can detect fluctuations in device energy levels in real time and allocate training tasks to balance cost and performance. Tools for resource monitoring, addressed in [106], enable fine-grained control over power usage during federated training.

Despite these advancements, several challenges persist. Balancing trade-offs between energy efficiency, accuracy, and convergence rates remains a contested area in research, as overly aggressive optimization for energy savings can degrade model performance. Additionally, as FL becomes applicable to an increasingly diverse array of devices, tailoring resource-aware optimizations to heterogeneous hardware configurations is imperative. Real-world experiments on large-scale, energy-diverse FL infrastructures are still limited, leaving gaps in understanding how solutions generalize across scenarios with extreme heterogeneity.

Emerging directions, including integrating reinforcement learning for adaptive control of energy-efficient training schedules and better quantifying the energy-accuracy trade-offs under varying system conditions, are promising. Furthermore, aligning FL with green computing objectives—such as integrating renewable energy grids into edge systems—could redefine sustainability in large-scale machine learning deployments. Future innovations will likely

combine layered optimizations across computation, communication, and scheduling, ensuring that resource-aware FL systems can operate efficiently without compromising model fidelity or client participation incentives.

5.6 Enhancing Fault Tolerance and Robustness

Federated learning (FL) systems operate in highly dynamic and often unreliable network environments, making fault tolerance and robustness integral for enabling efficient, scalable, and energy-conscious deployments. This subsection examines strategies to address key challenges, such as client dropouts, unreliable connectivity, and unforeseen disruptions, emphasizing how these mechanisms interact with resource-aware FL implementations to ensure robustness without compromising privacy or efficiency.

One critical avenue for improving fault tolerance is addressing client dropouts, a common occurrence in FL due to the diverse operating environments of client devices. Incremental update techniques, which incorporate partial local updates when clients cannot complete full training cycles, mitigate the effects of abrupt dropouts [107]. These approaches align well with energy-aware strategies by reducing the need for redundant communication while maintaining progress in the aggregation process. Alternatively, redundancy-based techniques employ duplicate client participation or proxy models to replicate updates from dropped clients [108]. While effective in preserving model accuracy, these methods can impose additional resource demands, requiring careful balance to align with the energy efficiency goals discussed earlier.

Lenient synchronization protocols are particularly relevant in non-uniform network conditions, relaxing the requirement for strict synchronization among clients. Asynchronous FL frameworks allow aggregation of available updates without waiting for all participants, inherently improving robustness in scenarios with intermittent connectivity. However, these protocols can introduce potential convergence challenges due to stale updates, which may impair global model accuracy. Techniques like weighted asynchronous updates, which assign greater importance to fresher updates, help mitigate these issues but require calibrated mechanisms to balance robustness, energy efficiency, and model performance [88].

Secure and transparent log management enhances fault tolerance, particularly in partially trusted environments. Blockchain technology, with its decentralized and tamper-proof characteristics, provides a promising solution by ensuring immutable participation records and flagging potential disruptions, such as data corruption or non-cooperation. However, as highlighted in resource-aware FL discussions, the computational and energy overhead associated with blockchain integration must be carefully addressed to maintain scalability on resource-constrained devices.

Another pivotal aspect lies in robust aggregation mechanisms, which mitigate the cascading effects of disruptions while preserving the system's integrity. Methods like Multi-KRUM and Adaptive Federated Averaging filter outliers and suspicious updates, ensuring the robustness of the global model against adversarial behaviors such as Byzantine faults [108]. These strategies complement energy-efficient communication techniques by minimizing redundant computations caused by untrustworthy updates, thereby optimizing both robustness and resource consumption. Nonetheless, trade-offs in computational complexity and scalability must be carefully evaluated.

Backup replication systems offer additional safeguards to prevent catastrophic data or model loss during unexpected server downtimes. By employing secondary servers or local caching mechanisms, these systems enable seamless recovery of state information. Such strategies align well with energy-conscious hierarchical aggregation techniques, as they distribute workloads and recovery tasks in scalable ways [109].

Emerging trends in FL robustness highlight hybrid frameworks that combine predictive analytics with fault recovery mechanisms, further strengthening system resilience. Predictive modeling, for example, anticipates potential client dropouts or network failures and redistributes computation to robust nodes preemptively [2]. When integrated with dynamic noise injection, these approaches maintain privacy guarantees while concurrently advancing fault tolerance [79]. These innovations, when combined with resource-aware strategies, promise greater flexibility in maintaining operational stability without incurring prohibitive energy costs.

Despite significant progress, open challenges remain at the intersection of fault tolerance and energy-efficient FL. Practical implementations must address resource constraints that limit the ability of clients to participate in fault tolerance mechanisms optimally. Furthermore, trade-offs between robustness and communication efficiency, especially in large-scale, heterogeneous environments, require additional exploration. Future research can advance distributed self-healing mechanisms, wherein client nodes collaboratively detect and mitigate disruptions, fostering a balance between resilience, energy efficiency, and privacy.

In summary, ensuring fault tolerance and robustness in FL systems requires a comprehensive suite of strategies—from incremental updates and asynchronous protocols to robust aggregation mechanisms and predictive fault recovery. By integrating these efforts with resource-aware designs, FL systems can achieve scalable, energy-efficient, and privacy-preserving deployments, ultimately advancing the reliability of federated learning ecosystems.

6 Security Challenges and Defense Mechanisms in Federated Learning

6.1 Poisoning Attacks and Defense Mechanisms

Federated learning (FL) introduces unique vulnerabilities to poisoning attacks, where adversarial participants aim to compromise the global model by injecting malicious updates. These poisoning attacks can be categorized generally into two types: data poisoning and model poisoning. Each approach exploits the distributed nature of FL to degrade model integrity or embed backdoors, often without being easily detected due to the aggregation processes and data privacy constraints intrinsic to FL. This subsection examines these attacks and the existing defense mechanisms, focusing on their strengths, limitations, and technical trade-offs

to provide a comprehensive understanding of this critical security challenge.

In data poisoning attacks, adversaries manipulate their locally held datasets to generate misleading gradients during training, which are then injected into the global aggregation process. Such attacks can be untargeted, aiming to degrade overall model performance, or targeted, designed to create specific, hidden backdoors triggered by particular inputs. For instance, adversarial clients might carry out label-flipping attacks by intentionally mislabeling training data to mislead the model's learning process. Simpler yet damaging methods, such as optimizing local training to maximize divergence from honest updates, can also amplify poisoning effects [32], [44].

Model poisoning attacks, on the other hand, do not rely on altering datasets but rather directly manipulate model updates sent to the server during each communication round. Techniques in this category include crafting malicious gradients that bias model behavior or embedding backdoor triggers, enabling adversaries to exploit the global model under specific conditions. A notable challenge for detecting model poisoning arises from the non-IID (non-independent and identically distributed) nature of client data, which can naturally lead to significant variation in updates. This variability provides attackers a natural disguise, especially in targeted backdoor attacks [32], [50].

To mitigate these risks, a variety of defense strategies have been proposed, with robust aggregation protocols frequently emerging as a foundational mechanism. Approaches such as Multi-KRUM exclude a subset of outlier model updates based on distance metrics, reducing the influence of malicious clients during aggregation. Other robust methods, such as Trimmed Mean and Bulyan, leverage statistical techniques to neutralize outlier gradients and achieve improved resilience against Byzantine failures. However, these defenses face scalability challenges when the number of adversarial parties grows relative to the total client base or when adversaries deliberately remain within statistical thresholds to evade detection [34], [50].

Additionally, anomaly-detection frameworks employ machine learning and statistical methods to identify malicious updates by assessing patterns in gradients or model contributions, often using features such as variance or similarity to past updates. For instance, solutions like FedCC integrate clustering-based approaches to discern normal versus anomalous client behavior, effectively filtering out suspicious updates without compromising benign contributions. Although promising, these strategies may yield higher computational costs, potentially limiting their feasibility in low-resource or high-client settings [2].

Emerging trends highlight hybrid solutions that combine multiple defense mechanisms to enhance robustness while maintaining flexibility for diverse FL scenarios. For instance, integrating secure multi-party computation (SMPC) with robust aggregation facilitates privacy-preserving defenses that can thwart both data and model poisoning attacks. Similarly, blockchain-enabled federated learning frameworks are gaining traction as they provide decentralized trust and immutable logging capabilities, reducing risks of collusion or tampering by adversarial participants [15], [110].

Nevertheless, significant challenges persist. Coordinating these defenses in large-scale, heterogeneous client systems remains difficult, especially as clients often operate under resource constraints or unreliable network conditions. Further, dynamic poisoning attacks, where adversarial strategies evolve across training rounds, highlight the need for adaptively responsive defense mechanisms. Future work must explore scalable defense frameworks leveraging advanced anomaly detection, adaptive aggregation protocols, and hybrid cryptographic techniques to balance robustness, computational efficiency, and model utility. Continued research into quantifiable trade-offs between security, privacy, and system scalability will be essential for establishing federated learning environments resilient against poisoning attacks.

6.2 Privacy Leakage and Mitigation Strategies

Privacy leakage in federated learning (FL) has emerged as a critical security concern, stemming primarily from the risk of sensitive user data being inferred through shared model updates. While FL preserves data locality by design, adversaries can exploit vulnerabilities in gradient information or aggregated updates to reconstruct private data or infer sensitive attributes. This subsection explores these privacy risks in the context of FL, presenting mitigation strategies and discussing the trade-offs, technical innovations, and emerging challenges at the intersection of privacy, robustness, and scalability.

A key privacy leakage threat in FL is the gradient inversion attack, where adversaries leverage gradients shared by clients to reconstruct input data with alarming accuracy. This is possible because gradients encode correlations with the input features of neural networks, serving as unintentional carriers of sensitive information. For instance, Zhao et al. demonstrated the feasibility of reconstructing high-resolution images solely from communicated gradients, underscoring the vulnerability of standard FL protocols to this attack vector [85]. The issue becomes particularly acute when fine-grained updates are exchanged, such as in high-dimensional datasets frequently encountered in fields like image processing, text analysis, or healthcare [4], [37].

Another significant privacy threat lies in membership inference attacks, wherein adversaries aim to deduce whether specific data points were part of a client's training dataset. Such attacks exploit the overfitting tendencies of FL-trained models, especially in non-IID (non-independent and identically distributed) settings where the global model adapts excessively to outlier distributions. Membership inference risks are heightened by the varying generalization capabilities of the global model across clients, meaning some users may face disproportionately elevated privacy risks [7]. This underscores the challenge of ensuring an equitable level of privacy protection across all FL participants.

Mitigating privacy risks in FL primarily involves differential privacy (DP), secure aggregation protocols, and techniques for abstracted model sharing. DP mechanisms introduce calibrated noise to gradients or model updates, obscuring the contributions of individual data points. For example, client-level DP techniques in cross-device FL systems inject Gaussian noise during gradient submission, adhering to formal $((\varepsilon, \delta))$-DP guarantees [11]. While DP

provides strong theoretical privacy protections, it entails a trade-off: excessive noise can degrade model accuracy, particularly in large-scale systems or resource-constrained deployments [37], [111].

Secure aggregation protocols aim to prevent direct data reconstruction by ensuring that only aggregated updates are visible to the central server. Technologies such as secure multi-party computation (MPC) and homomorphic encryption (HE) are at the forefront of these approaches. For instance, HE facilitates computations on encrypted updates, bolstering data privacy during training without compromising functionality [33]. However, the computational overhead associated with such methods poses significant challenges for low-resource devices like IoT sensors or mobile phones [78].

Emerging hybrid approaches aim to maximize the strengths of existing techniques. For example, frameworks combining DP noise addition with HE-secured aggregation, such as PRECAD, offer resilience against diverse threat vectors while maintaining operational efficiency [26], [37]. Similarly, representation masking strategies, which share obfuscated intermediary representations rather than raw gradients or updates, are gaining traction as a promising method to mitigate feature leakage risks [101]. These innovations reflect a growing trend toward multi-layered defense mechanisms in FL privacy research.

Despite progress, significant challenges remain unresolved. Privacy-preserving mechanisms such as DP and secure aggregation are often less effective in non-IID settings, where data heterogeneity creates additional complexities, such as uneven degradation in model performance [112]. Additionally, adversaries with advanced computational resources can exploit residual correlations in encrypted updates or denoised gradients, necessitating more nuanced obfuscation techniques [44]. Furthermore, scalability issues, exacerbated by dynamic participation and client dropout scenarios, pose a continuing obstacle to deploying effective privacy-preserving FL systems in real-world environments [24].

Looking forward, integrating adaptive noise injection techniques with data-aware privacy thresholds offers a pathway to balance privacy and utility more effectively. Innovations in federated simulation platforms, such as PrivacyFL, provide valuable tools to study privacy attack patterns and assess countermeasure effectiveness under realistic conditions [51]. Additionally, advancements in hardware-accelerated cryptographic solutions may reduce computational overhead, enabling scalable and accessible privacy-preserving FL at the edge.

In summary, privacy leakage in FL represents a multifaceted challenge requiring sophisticated, multi-modal strategies. While foundational techniques like differential privacy and secure aggregation have laid the groundwork, their inherent trade-offs in efficiency and accuracy highlight the need for continued innovation. Hybrid methods, adaptive frameworks, and resource-aware approaches will play a pivotal role in achieving scalable, robust, and privacy-preserving FL systems capable of meeting the demands of diverse applications and stakeholders.

6.3 Robust Aggregation Protocols

Robust aggregation protocols are critical to defending federated learning (FL) systems against adversarial behaviors that aim to manipulate model updates during the global aggregation phase. These protocols mitigate the impact of malicious participants, data aberrations, and Byzantine failures, ensuring the integrity and robustness of the global model. The heterogeneity of clients' data, coupled with the distributed nature of FL, presents unique challenges in identifying abnormal updates without impairing overall model performance. This subsection examines computationally efficient and secure aggregation approaches, evaluates their effectiveness, and explores emerging trends in the field.

Among the foundational methods, Federated Averaging (FedAvg), while widely used for its simplicity and effectiveness in non-adversarial contexts, is highly vulnerable to model poisoning attacks where adversaries inject harmful gradients to bias the global model [23]. Consequently, robust aggregation strategies such as Krum, Trimmed Mean, and Median-based methods have garnered significant attention. Krum identifies updates that minimize the Euclidean distance to the majority of other updates, effectively mitigating the influence of outliers. However, it performs poorly in terms of scalability and computational complexity, especially in high-dimensional parameter spaces [10]. Trimmed Mean and Median-based methods, which aggregate model updates after removing extreme values, address data heterogeneity to some extent but may struggle under high levels of coordinated adversarial collusion [34]. These methods also assume that the majority of participants are honest, which limits their applicability in scenarios with widespread byzantine clients.

More advanced approaches leverage redundancy and redundancy-aware defenses, such as Multi-Krum, which builds on Krum by iteratively selecting multiple candidate updates. While Multi-Krum improves fault tolerance, its inefficiency in large client populations has prompted further exploration of scalable yet robust protocols [44]. In parallel, robust weighted aggregation methods like FoolsGold use contribution weighting to suppress disproportionately influential updates by estimating the similarity of client update directions. Such strategies efficiently discourage free-riding or adversarial behavior in heterogeneous setups [36].

Blockchain-based aggregation methods have emerged as a promising avenue, introducing decentralized trust mechanisms to track and validate updates without relying on a central server. For instance, blockchain-backed frameworks such as BFLC apply consensus mechanisms and distributed audit trails, ensuring tamper-proof integrity while simultaneously mitigating centralized vulnerabilities [38]. However, issues such as computational overhead and increased latency in real-time applications persist, necessitating further optimization.

Differentially private aggregation strategies provide an additional layer of resilience by introducing controlled noise into aggregated outputs, thus masking individual updates. Although these approaches align well with privacy-preserving requirements, they often face challenges related to balancing privacy guarantees with model utility, particularly in adversarial contexts with highly skewed gradient

distributions [11].

An increasing number of hybrid aggregation methods seek to integrate the strengths of established approaches. For example, crypto-aided methods such as those combining secure multiparty computation (SMPC) with robust aggregation significantly enhance security guarantees while maintaining computational efficiency at moderate communication costs [52]. Similarly, dynamic aggregation mechanisms adapt their robustness thresholds based on real-time inference of adversarial activity or underlying data distributions, providing contextual customization to improve both resilience and convergence rates [50].

Despite these advancements, significant challenges remain. Current robust aggregation protocols often assume fixed adversarial models and tentative collusion bounds, limiting their adaptability to evolving attack strategies. Future research must focus on harmonizing robustness with privacy-preserving requirements, investigating decentralized aggregation strategies that can scale effectively, and reducing computational overhead in edge-device constraint settings. Emerging paradigms such as federated unlearning also demand incorporation, ensuring that robust aggregation frameworks can handle data removal or contribution invalidation requests without compromising overall model integrity [113].

In summary, designing robust aggregation protocols that efficiently balance robustness, computational scalability, and privacy remains an open challenge in federated learning. Hybrid approaches combining statistical, cryptographic, and decentralized consensus techniques hold promising potential to address the multi-faceted adversarial landscape in FL. Sustained innovation in this domain will be central to achieving reliable, resilient, and secure federated systems.

6.4 Communication Security Threats and Countermeasures

The communication infrastructure in federated learning (FL) represents a critical attack surface, as model updates and auxiliary information are exchanged across potentially insecure channels. Ensuring secure and reliable communication is paramount to protecting sensitive information and maintaining system integrity, particularly in decentralized and large-scale FL deployments. This subsection examines common communication threats, evaluates existing mitigation strategies, and highlights promising research directions to bolster the resilience of FL communication pipelines.

Eavesdropping on transmitted updates remains one of the most prominent threats to FL communications, as adversaries can intercept model parameters, gradients, or auxiliary metadata to infer private details about client datasets. Although encryption protocols like Transport Layer Security (TLS) provide a basic defense by preventing direct interception, they are often insufficient against more advanced attacks. For instance, metadata leakage or side-channel vulnerabilities, such as analysis of packet timing and size, can allow adversaries to extract high-precision information about sensitive client attributes. Gradient inversion attacks have further demonstrated that even encrypted updates may expose label-level details, underscoring the limitations of conventional encryption in the face of sophisticated threats [53]. Therefore, defenses that extend beyond standard cryptographic methods are urgently needed to ensure comprehensive protection.

A related vulnerability involves message interception and tampering, where adversaries exploit transmitted updates to manipulate communication flows. This tampering can lead to backdoor attacks or model poisoning, especially in partially asynchronous FL systems. Advanced cryptographic techniques, such as homomorphic encryption (HE) and secure multiparty computation (SMPC), address these concerns by enabling computation on encrypted data without accessing raw updates [41]. While effective, these cryptographic methods often face scalability and efficiency challenges, making them less practical for large-scale FL systems involving resource-constrained devices.

To complement cryptographic measures, dropout-resilient aggregation protocols are increasingly critical for managing client unreliability and dynamic participation in FL systems. LightSecAgg, for instance, introduces a lightweight and robust design for secure aggregation, leveraging encoded aggregate masks to circumvent dependency on dropped clients' seeds. This approach reduces computational and communication overhead while maintaining strong privacy protections, thereby enhancing the feasibility of asynchronous FL under real-world conditions [55]. However, research has identified implementation pitfalls in existing secure aggregation protocols that adversaries can exploit. Misuse or incorrect configuration of these protocols can lead to inference attacks or privacy violations, as demonstrated in [56]. To address these risks, frameworks like EIFFeL enhance communication integrity by incorporating systematic validation mechanisms to detect and rectify aggregation inconsistencies.

Beyond cryptographic and aggregation-based safeguards, decentralization offers an alternative paradigm to strengthen the security of FL communications. Decentralized FL systems forgo reliance on a central server by adopting peer-to-peer communication models, which leverage techniques like Shamir's secret sharing and consensus algorithms to collaboratively aggregate updates [60]. These systems not only remove single points of failure but also distribute trust across participants, complicating the attack surface for adversaries. Nonetheless, implementing decentralized FL introduces challenges in maintaining efficiency and scalability, particularly given the dynamic nature of communication graphs in real-world settings.

In addition, adaptive encryption and obfuscation-based techniques are emerging as complementary solutions for secure FL communications. For example, selective masking methods obscure sensitive weight indices in transmitted updates, thereby diminishing the potential gain for adversaries even if the communication channel is compromised [59]. While effective in enhancing security, such methods must carefully balance obfuscation levels to avoid adverse impacts on model convergence rates or performance. Lightweight aggregation protocols, such as FedDM, further optimize communication efficiency without compromising security by selectively transmitting the most informative updates, thereby reducing bandwidth demands while maintaining robust protection [102].

Addressing the inherent trade-offs between privacy, sys-

tem performance, and scalability is another key challenge in FL communication security. Recent advancements have integrated differential privacy mechanisms into communication pipelines, adding controlled noise to protect individual updates from probabilistic inference attacks [58]. While this enhances privacy, finding the optimal balance between noise levels and model utility remains an open question, particularly in highly adversarial environments. Blockchain-backed decentralized frameworks also hold promise for mitigating communication vulnerabilities by maintaining immutable and auditable logs of client contributions [60]. However, as with other solutions, scalability and energy efficiency remain significant obstacles to widespread adoption.

In summary, communication security in federated learning requires a multi-pronged approach that integrates cryptographic defenses, adaptive frameworks, and decentralized architectures to mitigate an evolving spectrum of threats. The development of scalable, energy-efficient designs that blend secure aggregation, lightweight cryptographic protocols, and innovative techniques like selective masking offers a promising pathway toward enhancing communication security in FL. As FL continues to expand into diverse real-world settings, addressing these challenges will play a pivotal role in ensuring robust, efficient, and privacy-preserving collaboration among participants.

6.5 Trust Management in Federated Learning

Trust management in federated learning (FL) represents a critical challenge due to the decentralized and collaborative nature of the paradigm, where diverse and untrusted participants, such as edge devices or institutions, contribute to model training. This distributed architecture inherently creates vulnerabilities, as malicious participants can degrade model performance, undermine system integrity, or compromise privacy. Identifying and mitigating such threats through robust trust mechanisms is therefore essential to ensure fair participation and effective collaboration.

One of the most widely studied approaches to trust management in FL involves reputation-based systems, which assign trust scores to participants based on their historical behavior. These systems monitor contributions across multiple training rounds, assessing the alignment of submitted updates with the expected distribution of gradients or parameters. For instance, methods that calculate cosine similarity between client-submitted updates and the aggregated model update have been employed to detect outliers and untrustworthy updates. Advanced frameworks explore dynamic trust modeling, incorporating temporal trends and patterns in participation. However, trust score systems can face limitations, such as their inability to differentiate adversarial updates masked as legitimate ones. Moreover, while effectively identifying anomalous updates, these systems can disproportionately penalize participants operating under legitimate constraints, such as data heterogeneity or sparse connectivity [6].

Anti-collusion mechanisms target scenarios where groups of participants may conspire to bias the global model or influence the outcome of federated tasks. Techniques such as clustering participant updates based on similarity metrics have demonstrated potential for isolating suspicious collusions, preventing adversarial alliances from corrupting the aggregation process [50]. Recently, zero-trust systems, such as Federated Bayesian aggregation frameworks, have emerged as promising methods for combating participant collusion. These frameworks probabilistically estimate the reliability of client contributions, thus improving robustness. However, these approaches often entail higher computational or communication overhead, limiting their practicality in resource-constrained deployments.

Verification frameworks leveraging cryptographic principles offer another promising avenue for trust management. Zero-knowledge proofs (ZKPs) enable participants to demonstrate the correctness of their updates without revealing sensitive data, ensuring honest behavior during collaborative training. By verifying the statistical properties of updates (e.g., adherence to client-side data distributions) before model aggregation, such frameworks enhance the integrity of the FL process. Secure multiparty computation (SMPC)-based trust mechanisms further operationalize the federated framework by enforcing rigorous cross-validation of updates. Despite their merits, cryptographic solutions are frequently accompanied by computational inefficiencies, particularly as the number of participants scales [64], [114].

An emerging trend in trust management involves integrating machine learning techniques with FL-specific heuristics. Anomaly detection systems, powered by reinforcement learning or deep learning models, enable dynamic identification of malicious behaviors even in the absence of pre-defined adversarial patterns. For instance, systems employing clustering-based anomaly detection or outlier-resistant aggregation protocols such as Multi-KRUM have proven effective in mitigating the risk of Byzantine attacks and detecting poisoned updates [21]. However, challenges persist in balancing false positive rates (identifying benign clients as malicious) against tolerating adversarial updates.

A persistent challenge in trust management is adaptability to data and system heterogeneity. Participants differ significantly in computational capabilities, data distributions, and update patterns, adding complexity to distinguishing intentional adversarial behavior from genuine constraints. For example, standard deviation thresholds applied to update magnitudes may unfairly exclude participants with naturally high-gradient updates due to skewed data [8]. Future research should focus on hybrid trust frameworks capable of dynamically adjusting trust evaluation metrics to context-specific factors, considering not only gradient distributions but also local data characteristics and participant resource profiles.

While notable progress has been made, trust management in FL must continue to evolve. The integration of blockchain technology for immutable logging of client contributions has shown potential for decentralized trust enforcement, though scalability and performance bottlenecks remain challenges. Additionally, designing incentive mechanisms that reward consistent, high-quality participation could complement trust mechanisms and promote fair engagement across diverse participants [66].

In conclusion, the multifaceted nature of trust management in FL necessitates leveraging a combination of reputation systems, cryptographic frameworks, anomaly detec-

tion tools, and adaptive mechanisms. Future advancements should aim at balancing robustness, scalability, and inclusivity, while minimizing performance trade-offs to establish trust in diverse and dynamic federated ecosystems.

6.6 Advanced Hybrid Techniques for Enhanced Security

Hybrid techniques have emerged as a crucial frontier in addressing the multifaceted security challenges inherent to federated learning (FL), particularly in light of the dynamic and adversarial environments in which FL systems operate. By combining cryptographic protocols, differential privacy (DP), and advanced anomaly detection mechanisms, hybrid frameworks strive to provide a balanced approach that enhances both robustness and privacy without severely compromising computational efficiency. These techniques extend the foundational concepts of trust management and security addressed in preceding sections by proposing an integrated strategy that fortifies FL against sophisticated and evolving threat landscapes. This subsection offers a critical exploration of state-of-the-art hybrid approaches, discussing their design principles, strengths, limitations, and practical considerations, while identifying key directions for future innovations.

A prominent hybrid approach involves integrating cryptographic protocols, such as Secure Multiparty Computation (SMPC) and homomorphic encryption (HE), with DP to establish a multi-layered defense mechanism. For instance, frameworks like PRECAD leverage SMPC to securely aggregate client updates, while simultaneously injecting DP-calibrated noise to obscure sensitive information [12]. This dual-layer approach ensures that even if attackers gain access to encrypted data, the randomized noise obfuscates meaningful insights, delivering stronger privacy guarantees. However, the computational and communication overhead associated with encryption and noise addition often limits the scalability and practicality of such systems, particularly in resource-limited and large-scale deployments.

Another path for hybrid defenses lies in dynamic noise injection mechanisms, where the level of added noise adapts to factors like the adversarial model, data sensitivity, or the stage of model training [92], [115]. High noise levels during the early training iterations can obscure private details effectively, with noise gradually reduced as the model approaches convergence and requires fine-tuned adjustments for utility. This dynamic tuning directly addresses the privacy-utility trade-off intrinsic to FL. Nevertheless, ensuring adaptability across diverse FL architectures and threat models remains a significant challenge, often requiring extensive empirical validation before deployment in real-world systems.

Emerging hybrid solutions also highlight the integration of domain-specific simulators, such as PrivacyFL, to proactively evaluate FL systems under various threat scenarios. These simulators play a pivotal role in stress-testing hybrid frameworks, identifying vulnerabilities to sophisticated attacks like adaptive gradient inversion or model manipulation [116]. By equipping hybrid frameworks with real-time feedback mechanisms, simulators facilitate dynamic protocol adjustments, such as triggering secure aggregation protocols or enhancing DP parameters when signs of adversarial behavior are detected. However, building simulators that can generalize effectively across heterogenous data distributions and FL architectures poses a technical challenge, necessitating further advancements in simulation fidelity.

Hybridization has also extended to blockchain technology, where its inherent transparency and immutability enhance trust and accountability in FL systems. When combined with Byzantine-resilient aggregation protocols such as Multi-KRUM, blockchain-based hybrid frameworks can ensure auditable logs of aggregation operations, bolstering defenses against adversarial manipulation [108]. Yet, the energy and latency demands associated with blockchain mechanisms remain significant hurdles, especially for deployment at scale. Developing lightweight alternatives is therefore crucial for making such solutions feasible for practical, large-scale FL ecosystems.

Proactive adversarial modeling represents another innovation in hybrid frameworks, where generative models are used to mimic potential attacks and assess vulnerabilities in real time. For example, methods like Generative Gradient Leakage (GGL) leverage surrogate attacks to identify privacy weak points during training, enabling preemptive modifications to aggregation strategies or the activation of additional obfuscation mechanisms [87]. While effective in mitigating unknown attack vectors, these approaches often demand intensive computational resources and sophisticated model-training pipelines, which may limit their applicability to FL systems operating on resource-constrained devices.

The strength of hybrid approaches lies in their unified defense strategy, which combines orthogonal solutions like cryptographic methods, probabilistic privacy protections, and anomaly detection tools. This diversified approach reduces the "attack surface" by introducing multiple layers of defenses against both known and emerging threats. However, these techniques inevitably introduce trade-offs, including increased computational complexity, communication overhead, and potential reductions in model accuracy. For instance, protocols that blend SMPC with advanced DP mechanisms often consume excessive resources, making them impractical for large-scale, heterogeneous FL networks [93].

Looking ahead, the future of hybrid techniques must focus on overcoming the juxtaposed challenges of scalability, efficiency, and adaptability. Distributed privacy architectures that enable localized intelligence—such as client-side noise mitigation or lightweight secure aggregation schemes—offer a promising direction [79]. Additionally, the development of adaptive hybrid frameworks that dynamically activate different defenses based on real-time risk assessments could significantly enhance their utility in handling adversarial environments. Such innovations will require interdisciplinary research efforts to ensure seamless integration with regulatory standards, diverse FL architectures, and real-world constraints.

In conclusion, hybrid techniques represent a transformative pillar in federated learning security, marking an evolutionary step toward robust, privacy-preserving collaborative learning systems. By synthesizing advancements in cryptography, privacy theory, and adversarial modeling, these

frameworks pave the way for scalable solutions capable of addressing the complex adversarial dynamics of FL. Building upon the trust management strategies discussed earlier, hybrid defenses offer promising opportunities to safeguard FL systems and unlock the full potential of collaborative machine learning in diverse and adversarial scenarios.

7 Applications and Domain-Specific Deployments of Federated Learning

7.1 Healthcare and Life Sciences

The healthcare and life sciences domain stands out as one of the most critical application areas for federated learning (FL), where privacy-preserving collaborative machine learning methodologies have immense transformative potential. In this data-intensive field, sensitive patient information, fragmented across institutions, poses significant challenges for traditional centralized approaches, making FL a promising solution to enable cross-institutional cooperation without compromising data privacy. By maintaining local data storage and only sharing aggregated model updates, FL aligns with stringent privacy regulations such as HIPAA and GDPR, thus addressing inherent legal, ethical, and technical obstacles [4], [67].

One of the prominent applications of FL in healthcare is in clinical risk prediction, leveraging distributed electronic health record (EHR) data for tasks such as predicting in-hospital mortality or disease progression. For example, FL implementations that aggregate patient data from diverse institutions enable training robust predictive models capable of handling heterogeneous patient populations without exposing specific individuals' records. This is important because local patient data often exhibit inherent non-IID distributions due to institutional differences in demographics, healthcare protocols, and diagnostic technologies. Studies indicate that models combining such diverse data sources under the FL framework achieve performance on par with centralized models while maintaining stringent privacy guarantees [4], [8]. However, challenges persist in addressing both statistical heterogeneity and learning dynamics under limited communication budgets.

FL has also been extensively employed for medical image analysis tasks, including brain tumor segmentation, chest radiograph classification, and dermatological condition detection. Traditional approaches to centralizing medical imaging often face ethical opposition due to the sensitive nature of these data types. FL mitigates these concerns by enabling distributed training across facilities. Techniques such as federated averaging (FedAvg) and advanced privacy mechanisms like homomorphic encryption ensure both confidentiality and utility. For instance, federated implementations that integrate differential privacy mechanisms during gradient exchange have further bolstered robustness against adversarial inference attacks while preserving model accuracy [29], [45]. Despite these advances, scaling up FL deployment to larger imaging datasets remains resource-intensive, with communication overhead, limited computing power at local sites, and synchronization challenges being key constraints.

The application of FL in personalized medicine exemplifies its ability to tailor global models to individual patients. By combining FL with techniques such as meta-learning, researchers have developed frameworks that personalize drug recommendations and optimize treatment protocols for diseases such as cancer and diabetes. For example, models trained collaboratively across pharmaceutical datasets and clinical centers can leverage diverse patient responses to tailor drug regimens. While these approaches demonstrate improved model effectiveness, personalization during FL often increases the risk of re-identifiability, necessitating the integration of advanced anonymization and differential privacy techniques [9], [117].

FL's flexibility has also enabled addressing healthcare-specific challenges such as data silos and institutional inconsistencies in data formats. For example, vertical FL allows multi-organization collaborations by aligning feature spaces across institutions. This is particularly relevant for scenarios where institutions possess complementary datasets (e.g., feature-rich patient medical histories versus demographic indicators). Vertical FL solutions using secure multiparty computation and entity resolution protocols, such as SecureBoost, have demonstrated high scalability and strong privacy guarantees, with applications ranging from insurance risk modeling to genetic data analysis [4], [14].

Despite its numerous applications, FL in healthcare reveals several open challenges. Efficiently handling skewed and non-IID data distributions, balancing privacy-utility trade-offs when applying privacy-preserving mechanisms, and mitigating adversarial threats remain at the forefront of research. Furthermore, real-world deployments necessitate addressing operational complexities such as reliable internet connectivity for client updates and robust synchronization mechanisms across distributed healthcare infrastructures [7], [34]. Future research must also explore the integration of evolving technologies such as blockchain for decentralized model aggregation and generative adversarial networks (GANs) for data synthesis to enhance diversity while preserving data utility [26], [29].

In summary, federated learning has emerged as a game-changing paradigm in healthcare, offering privacy-preserving solutions to longstanding collaborative challenges. The development and adoption of FL in this domain are shaped by its ability to respect regulatory frameworks, enable secure cross-institutional collaborations, and address the data heterogeneity intrinsic to healthcare. Future efforts must focus on scalable system optimization, robust defenses against data leakage, and interdisciplinary collaborations to unlock the full potential of FL in advancing medical research and improving global health outcomes.

7.2 Financial Services and Banking

The financial services and banking sector exemplifies a critical application domain for federated learning (FL), driven by the dual demands of harnessing distributed data and complying with stringent privacy regulations such as the General Data Protection Regulation (GDPR). FL's decentralized, privacy-preserving paradigm addresses significant challenges in this space, enabling institutions to collaboratively train machine learning models for tasks like fraud detection, credit risk assessment, customer segmentation, and cross-border financial crime investigation. By facilitating secure data collaboration without requiring raw data

sharing, FL ensures confidentiality while fostering innovation in data-driven financial ecosystems.

One of the most impactful applications of FL in finance is fraud detection. Financial institutions often encounter limitations in centralizing transactional data due to privacy concerns and regulatory restrictions. FL overcomes these hurdles by enabling collaborative training on siloed and heterogeneous datasets, leveraging diverse patterns of fraudulent activity across entities. Instead of sharing raw data, participating institutions exchange model updates, preserving data privacy without compromising predictive performance. For instance, federated training frameworks integrating secure aggregation protocols can detect fraud-related anomalies across institutions while ensuring client confidentiality [24]. Nonetheless, fraud detection in FL faces challenges stemming from the non-IID (non-independently and identically distributed) nature of financial datasets and adversarial attacks designed to manipulate shared updates. Techniques such as Byzantine-resilient aggregation have been proposed to counter these risks and ensure model robustness against adversarial threats [1].

Credit scoring and risk modeling represent additional applications where FL has demonstrated substantial utility. In financial networks with overlapping yet incomplete datasets—such as lending consortia or joint ventures—institutions often need to collaboratively evaluate client risk profiles. Vertical federated learning (VFL) addresses this by enabling feature-partitioned datasets from multiple parties to collectively train predictive models without violating data privacy agreements. For example, one institution may contribute demographic details while another offers transactional data, jointly enhancing risk evaluation models without sharing sensitive information. Advanced secure computation techniques, such as homomorphic encryption, further safeguard sensitive attributes like income levels or credit histories throughout the training process [19], [33]. However, achieving an optimal balance between cryptographic overhead and model accuracy remains an ongoing challenge, particularly as the complexity of feature sets increases.

Moreover, FL has significant potential in the context of open banking frameworks, wherein third-party service providers access consumer banking data with explicit consent. These frameworks demand strong privacy guarantees during collaborative machine learning efforts. FL aligns well with open banking initiatives by enabling multi-institution analytics while ensuring compliance with data sovereignty laws. Applications such as personalized financial recommendations and customer behavior analysis benefit from the privacy-preserving nature of FL. Federated transfer learning (FTL), for example, allows global models to adapt to local data distributions, achieving personalization without exposing shared data [30]. However, heterogeneities in data formats and distributions across institutions, coupled with fairness concerns, present key challenges. These issues must be addressed to unlock the full potential of FL in open banking contexts [68].

Emerging use cases such as cross-border financial intelligence and anti-money laundering (AML) efforts further illustrate FL's transformative potential in finance. These tasks often involve collaborative analysis of transaction data across international jurisdictions, where privacy laws and data-sharing restrictions complicate centralized approaches. FL enables such global collaborations by allowing distributed training without violating local privacy regulations. Federated anomaly detection techniques incorporating differential privacy have shown promise for identifying suspicious transaction patterns at scale, enhancing capabilities for AML monitoring [11]. Nevertheless, high-frequency applications like AML face challenges related to computational demands and communication overhead. Innovations in lightweight model training and communication-efficient secure computation remain crucial for ensuring scalability and practicality [118].

Looking forward, the adoption of FL in financial systems hinges on overcoming several persistent hurdles. Ensuring resilience against model poisoning attacks and addressing collusion risks among financial entities will be vital for safeguarding the reliability and integrity of collaborative models. Additionally, balancing privacy, efficiency, and predictive accuracy in real-world deployments will require adaptive techniques capable of navigating the diverse operational constraints within the financial sector. Integrating FL with technologies such as blockchain could further enhance transparency, providing mechanisms for participation auditing and incentivization that bolster trust in collaborative frameworks [26]. As FL methodologies continue to evolve, they hold unparalleled potential to revolutionize secure, privacy-preserving, and collaborative analytics within the financial services sector.

7.3 Internet of Things (IoT) and Edge Computing

The integration of Federated Learning (FL) with Internet of Things (IoT) and edge computing is reshaping the landscape of decentralized machine learning, enabling data-driven solutions at the network periphery while preserving user privacy and optimizing system efficiency. The IoT ecosystem, comprising billions of interconnected devices across diverse domains such as smart homes, industrial automation, healthcare wearables, and intelligent transportation systems, produces vast amounts of sensitive, geographically distributed data. Leveraging FL in such environments ensures on-device training of machine learning models, thereby eliminating the need to centralize raw data and mitigating privacy risks and communication overheads [21] [70].

FL's deployment in IoT scenarios addresses several unique challenges, primarily non-IID data distributions across edge devices, resource constraints, and energy efficiency. Devices in IoT networks often generate highly heterogeneous data due to variations in location, usage patterns, sensor quality, and environmental factors. This data heterogeneity can degrade model convergence and global performance. Techniques such as personalized federated learning and cluster-based aggregation have emerged to address these discrepancies, allowing subsets of devices with statistically similar data to collaboratively train specialized models [9] [8]. Additionally, algorithms like FedProx and Scaffold have been shown to enhance model robustness under non-IID conditions by introducing proximal terms or variance reduction techniques to stabilize updates during training [35]. These advancements are crucial in IoT

environments where devices can exhibit highly skewed or imbalanced data distributions.

Bandwidth and energy efficiency represent another critical concern for FL in IoT and edge computing, given the limited computational and transmission capabilities of many devices. Communication-efficient methods, such as gradient sparsification, model compression, and structured updates, play a pivotal role in minimizing data exchanges between devices and central servers without significantly compromising model accuracy. Studies have shown that techniques like sparse communication and quantized gradient updates can reduce communication costs by orders of magnitude, making FL feasible even for resource-constrained IoT systems [31] [52]. Incorporating low-power optimization techniques further ensures that devices can sustain learning operations without depleting battery life, a key requirement for long-term IoT deployments [21].

The convergence of FL with IoT has also spurred domain-specific applications. In smart homes and cities, FL enables devices to collaboratively predict energy consumption patterns, automate processes, and enhance user experiences while safeguarding privacy [21]. In industrial IoT contexts, FL facilitates predictive maintenance and process optimization by aggregating insights across factories and sensors without compromising proprietary data. Moreover, transportation networks utilize FL models to forecast passenger demand, optimize traffic flows, and coordinate autonomous vehicles. These applications leverage hierarchical federated learning frameworks, wherein edge devices perform localized aggregations before forwarding updates to a central server, thereby addressing scalability issues and reducing latency in large-scale IoT deployments [119] [17].

Despite its transformative potential, several challenges remain unresolved. Privacy threats, such as model inversion and gradient-based reconstruction attacks, are particularly concerning for IoT due to the sensitive nature of the data generated. Privacy-preserving techniques, including homomorphic encryption and differential privacy, have been integrated into FL to provide added security during model updates, but their adaptation for resource-constrained IoT environments requires further refinement [44] [34]. Another emerging challenge lies in ensuring fault tolerance and addressing the high variability in device participation due to intermittent connectivity or hardware failures. Blockchain-assisted decentralized FL frameworks hold promise for mitigating single-point failures and improving trust within distributed networks [16].

Future directions in this domain focus on the seamless integration of FL with IoT-specific constraints. Lightweight FL algorithms tailored to ultra-low-power devices, adaptive aggregation protocols for dynamic topologies, and personalized strategies for highly diverse settings are critical areas of exploration [17] [10]. Additionally, cross-modal federated learning, which combines IoT data streams from diverse modalities such as video, audio, and sensor readings, represents a promising avenue for building comprehensive, context-aware intelligent systems. As these innovations advance, FL's penetration into IoT-powered edge computing ecosystems is poised to drive significant improvements in privacy-preserving analytics, efficiency, and functionality.

7.4 Collaborative Governance and Public Policy

Federated Learning (FL) has emerged as a transformative paradigm for fostering collaborative governance and public policy, addressing critical societal challenges—ranging from public health to infrastructure development—through secure and privacy-preserving cross-institutional data analysis. This framework is particularly valuable in contexts where large-scale, decentralized data is siloed across multiple stakeholders, hindering effective data-driven decision-making. Leveraging FL's capabilities, institutions can collectively train machine learning models across distributed datasets while upholding data sovereignty and complying with privacy regulations.

In public health, FL facilitates the aggregation of epidemiological data for robust predictive modeling, enabling applications such as forecasting disease outbreaks and optimizing vaccination strategies. By securely aggregating geographically and organizationally distributed datasets, FL enables the identification of nationwide trends without breaching patient confidentiality. Privacy-enhancing techniques like secure multi-party computation (SMPC) and homomorphic encryption ensure the confidentiality of sensitive data during collaboration [41]. Additionally, approaches integrating differential privacy protect model training against inference attacks, preserving data security even under adversarial conditions [58]. However, healthcare contexts often face challenges associated with non-IID (not independently and identically distributed) data, as patient and organizational characteristics vary significantly across institutions. To address this, domain-specific adaptations like hierarchical FL and personalized federated learning (PFL) have been proposed, allowing the global model to generalize effectively while providing localized insights [29]. Striking a balance between equitable prediction accuracy for different stakeholders and the coherence of the global model remains an open challenge in such applications.

In urban management, FL has shown promise in enabling cooperative infrastructure planning and smart city analytics by securely synthesizing data from municipalities, private enterprises, and other organizations. For example, federated frameworks can analyze traffic patterns for improved urban mobility, integrate sensor data for smarter grid management, and enhance disaster preparedness and response strategies. Clustering-based FL techniques [39] allow for localized knowledge sharing while aligning with broader regional goals, fostering adaptive and efficient decision-making. Nonetheless, large-scale deployments are often hampered by communication bottlenecks and the computational limitations of edge devices. Lightweight solutions such as gradient sparsification and dynamic masking have demonstrated the potential to reduce communication overhead while preserving model performance [59].

Cross-institutional research and policy modeling further highlight FL's potential for addressing interdisciplinary challenges, supporting evidence-based policymaking in areas like climate change, economic strategy, and education systems. FL enables secure collaboration between governmental agencies, academic institutions, and private entities, facilitating joint initiatives without data centralization. For instance, FL has been employed in federated recom-

mendation systems tailored to regional requirements for educational resources [43]. Furthermore, the integration of FL with foundation models—pre-trained on extensive datasets—redefines cross-institutional workflows, enabling fine-tuning approaches that minimize data transfer and enhance personalization [74]. However, incorporating the diverse objectives of participating institutions introduces a complexity in designing federated optimization strategies that balance global performance metrics with localized priorities and needs.

Despite its promise, deploying FL in governance contexts requires navigating several regulatory and ethical considerations. For instance, compliance with data localization laws like the General Data Protection Regulation (GDPR) necessitates the alignment of FL architectures with jurisdictional mandates [120]. Moreover, emerging security vulnerabilities—such as label recovery or inference attacks—call for enhanced defenses to preserve FL's foundational privacy assurances [53]. Establishing trust within collaborative ecosystems is equally crucial, with reputation-based aggregation protocols proposed to mitigate risks of adversarial or dishonest participants [54].

To fully realize FL's potential for collaborative governance and public policy, future research must emphasize interdisciplinary innovation by coupling technical advancements with legal and ethical solutions. Blockchain-integrated FL systems, for instance, could autonomously log contributions to ensure transparency and facilitate tamper-proof audits. Similarly, advancements in communication-efficient, decentralized FL architectures support scalability, making them better suited for global governance applications. By integrating advanced cryptographic techniques with adaptive FL frameworks, it is possible to transcend existing data-sharing barriers, enabling collective action that addresses complex and urgent societal challenges with both efficiency and privacy intact.

7.5 Emerging Domains and Interdisciplinary Applications

Federated Learning (FL) continues to make strides beyond traditional applications, finding utility in emerging and interdisciplinary domains where challenges like privacy, scalability, and diverse data integration are particularly pronounced. The adaptability of FL to address these issues positions it as a critical enabler in fostering collaboration across different sectors while adhering to strict privacy and regulatory requirements. This subsection examines the applications of FL in emerging fields such as environmental science, education, federated recommendation systems, and multimodal data integration, emphasizing their unique demands and intersections with FL's technical capabilities.

A prominent interdisciplinary domain where FL has demonstrated substantial potential is environmental analytics and climate modeling. These applications require collaboration among institutions across geographical boundaries while addressing the constraints posed by sensitive environmental data distributed across various siloed repositories. FL enables global collaborations for tasks such as climate change predictions, biodiversity mapping, and large-scale environmental modeling without data centralization, thus preserving both regional autonomy and confidentiality. Recent work highlights methods that allow representations of distributed data to be aggregated efficiently, even under highly heterogeneous conditions [7]. However, challenges remain in accommodating the domain's reliance on complex, high-volume spatiotemporal data, which necessitates scalable communication protocols and robust model adaptation techniques.

Education, particularly personalized learning systems, represents another domain ripe for FL adoption. Institutions, constrained by privacy laws and student data sensitivity, can benefit from FL-powered collaborative frameworks to develop models that personalize educational resources and assessments across diverse learner populations. For example, personalized federated learning (PFL) approaches, such as those leveraging shared feature representations combined with local customization [121], have shown promise in developing tailored learning experiences without requiring direct access to institutional data. However, integrating FL techniques to manage the heterogeneous educational datasets and maintaining representational fairness among diverse demographic groups remain open challenges.

Federated recommendation systems provide another compelling use case for FL, enabling privacy-aware personalization in industries like e-commerce, digital streaming, and social media. These systems often aggregate insights from user interactions across multiple platforms to improve the relevance of recommendations without exposing individual user data. Novel methods, such as ensemble distillation techniques [72], mitigate FL's constraints by enabling robust aggregation of heterogeneous models. Despite their potential, federated recommendation systems face difficulties in balancing personalization with privacy, particularly under resource-constrained environments and non-IID data distributions, which can amplify spurious correlations or biases in recommendations.

Cross-modal and multimodal federated learning frameworks are emerging as transformative tools in integrating diverse data types such as text, image, and sensor data. These applications are especially relevant in fields requiring a synthesis of information from disparate sources—for instance, healthcare, autonomous systems, and multimedia analytics. Recent advancements incorporate techniques like multi-modal data fusion and alignment using contrastive representation learning [76], which enhances generalization across heterogeneous data streams. However, integrating multimodal data in FL systems introduces computational overhead and necessitates fine-grained synchronization mechanisms to manage intermodal dependencies.

While interdisciplinary applications of FL hold remarkable promise, several challenges must still be addressed. Data heterogeneity, communication bottlenecks, and model adaptability are recurring hurdles across these domains. Adding to these are domain-specific complexities, such as the need for fairness in educational setups, energy efficiency in environmental modeling, or modality-awareness in healthcare applications. Emerging trends point towards hybrid solutions that combine encryption-based privacy mechanisms with adaptive learning architectures [33], thus expanding FL's utility without compromising performance

or privacy guarantees.

Looking forward, further research is needed to improve FL's ability to handle interdisciplinary contexts with higher scalability, accuracy, and inclusivity. Developing domain-specific optimization strategies and model architectures tailored to emerging fields can help address such requirements. Additionally, innovative techniques, such as highly adaptive representation learning and federated meta-learning frameworks, can empower FL systems to generalize better across domains. These efforts, coupled with interdisciplinary collaborations, will be critical for realizing FL's transformative potential in solving complex global challenges.

8 Open Challenges and Future Directions in Federated Learning

8.1 Scalability and Infrastructure Optimization

Scalability remains a paramount challenge in federated learning (FL), especially as uptake expands to millions of diverse clients such as IoT devices, smartphones, and institutional silos. Achieving scalability in FL necessitates addressing systemic constraints in communication, computation, and infrastructure while preserving privacy, robustness, and model utility. Foundationally, FL systems must handle heterogeneous client participation without degradation in training efficiency or reliability, necessitating intelligent design in computational workflows and infrastructural architectures.

One of the key barriers to scalability is the communication overhead imposed by frequent exchange of model updates between the central server and clients. Emerging techniques such as gradient sparsification and model compression provide viable solutions to this issue. Gradient sparsification reduces the size of transmitted updates by retaining only the most significant gradients, while discarding near-zero components, significantly alleviating bandwidth constraints [94]. Model weight quantization, which compresses model weights using low-precision representations, further reduces the communication footprint, though it risks degradation in model accuracy if not carefully implemented [17]. Techniques combining these strategies have shown promise, balancing model accuracy and communication efficiency.

Infrastructure limitations in centralized architectures also pose a bottleneck, particularly when accommodating massive client participation. A promising approach lies in decentralized and hierarchical FL architectures. Decentralized frameworks enable clients to collaborate directly without relying on a central server, thus eliminating a single point of failure and reducing aggregation delay. For instance, blockchain-assisted frameworks like BLADE-FL integrate blockchain for secure, decentralized coordination of model updates, enhancing system fault tolerance and scalability [15]. On the other hand, hierarchical FL introduces intermediate aggregators, such as regional servers or edge devices, which perform partial aggregation before communicating with the central server. This reduces the central server's workload and enhances scalability in geospatially distributed client networks [21].

Client heterogeneity, particularly variations in device resources, connectivity, and data distributions, compounds the difficulties of scalability. Advances such as adaptive client selection frameworks prioritize clients with stable connections, adequate resources, or representative data subsets, ensuring robust training even under resource-limited scenarios [35]. Techniques like cluster-based FL, which segment clients into groups with similar characteristics, localize training processes and reduce global communication traffic. Additionally, frameworks like FedProx and Scaffold specifically address non-IID data while ensuring efficient convergence, making them suitable for deployment at scale [8].

Systemic constraints such as energy consumption and latency must also be tackled for large-scale deployments. Approaches prioritizing energy-efficient computation on battery-constrained edge devices, for instance, propose lightweight neural networks composed of fewer parameters to reduce local computational requirements [122]. To further minimize local and global resource usage, asynchronous communication protocols have been explored, allowing updates to occur irregularly without synchronous aggregation rounds [28]. These protocols achieve faster training with fewer communication rounds but may introduce challenges in ensuring model convergence under dynamically changing client participation.

An emerging avenue for scalable FL infrastructure is the integration of multi-level aggregation with dynamic resource allocation. For example, frameworks leveraging elastic resource scheduling dynamically allocate computation and bandwidth based on real-time system loads and network conditions [17]. Such elastic infrastructures could optimize training by adapting aggregation frequencies and client engagement rates to mitigate resource bottlenecks.

Despite these advances, significant open questions remain in scaling FL to its full potential. The tension between communication efficiency, privacy guarantees, and fairness calls for more sophisticated trade-off optimization, possibly through hybrid cryptographic protocols combining secure multiparty computation (SMPC) and differential privacy [11]. Additionally, addressing systemic failures—such as Byzantine clients injecting malicious updates—requires the development of resilient aggregation protocols capable of identifying and mitigating unreliable contributions without increasing aggregation complexity [44].

Future research should also explore the intersection of scalability and federated unlearning, where accommodating the right to be forgotten will introduce additional computational and communication challenges as systems scale [93]. Moreover, effective simulation frameworks for benchmarking FL algorithms under large-scale, multi-modal environments will be crucial in developing robust solutions [35].

In summary, achieving scalability in FL necessitates a multi-faceted approach, integrating communication-efficient updates, decentralized architectures, adaptive client optimization, and robust fault-tolerant mechanisms. By addressing these challenges, federated learning can achieve the reliability, efficiency, and inclusivity needed for deployment across massive and diverse client ecosystems.

8.2 Bias, Equity, and Fairness in Federated Learning

The challenges of bias, equity, and fairness in federated learning (FL) are critical yet underexplored issues, which are intrinsically tied to the decentralized and heterogeneous nature of FL systems. By design, FL operates across a diverse range of clients, each contributing data that is highly variable, non-IID (non-independent and identically distributed), and often imbalanced [7], [8]. This decentralization can inadvertently amplify systemic inequities, favoring participants with larger, higher-quality, or more representative datasets while marginalizing those with limited resources or underrepresented data. Addressing these disparities is crucial for ensuring that federated learning systems are both equitable and broadly applicable.

Fairness concerns in FL can be categorized along two primary dimensions: (1) inter-client fairness and (2) intra-population fairness. Inter-client fairness relates to ensuring that all participating clients—whether organizations or individual users—benefit equitably from the global FL model. Models often lean towards the distributions of clients with dominant datasets, leading to suboptimal performance on data from smaller or more diverse contributors [8], [39]. On the other hand, intra-population fairness evaluates whether FL models themselves perpetuate or exacerbate biases present in the underlying client datasets, such as demographic, geographic, or socio-economic inequities [117].

These fairness challenges originate from data heterogeneity, both statistical (variances in data distributions) and systemic (biases entrenched in data collection processes) [112]. Global aggregation strategies often fail to address these disparities, leading to trade-offs where benefits to some subpopulations come at the expense of others. Research has shown that global models in FL frequently underperform compared to personalized models tailored to individual data distributions, particularly when datasets are highly heterogeneous [75], [76]. However, personalization strategies, while boosting accuracy for specific clients, may inadvertently deepen inequities across groups by prioritizing local adaptations over collective fairness.

Several methodologies have been proposed to tackle these fairness issues. Fairness-aware optimization has received significant attention, with modified aggregation algorithms designed to promote equity. For instance, weighted federated averaging assigns greater importance to updates from underrepresented or underperforming clients, aiming to balance global model performance across diverse participants [10], [68]. While these approaches can enhance fairness, they often trade off with convergence rates and overall accuracy [22]. Similarly, fairness-focused client selection mechanisms prioritize balanced client representation during training, accounting for demographic diversity or data distribution variability. However, these mechanisms need careful design to avoid inadvertently excluding clients with limited resources or inconsistent participation [95].

In privacy-preserving FL systems, promoting fairness without explicit demographic information poses an additional layer of complexity. Emerging techniques leverage trust-aware optimization and fairness-driven loss functions to ensure fairer outcomes across clients and subpopulations without relying on sensitive demographic labels [69]. Federated regularization techniques, which integrate fairness-enhancing penalties directly into local training objectives, have also demonstrated promising results in reducing bias while maintaining compliance with privacy requirements [1].

Despite these advancements, numerous challenges persist. The lack of standardized fairness evaluation metrics within the FL paradigm makes it difficult to measure or validate equitable outcomes effectively. Existing metrics often fail to capture the intricate interplay between client participation variability and data representation. Metrics that evaluate both global utility (model accuracy) and equity (performance consistency across clients) are critically needed [117]. Additionally, the scalability of fairness-aware algorithms presents ongoing hurdles, particularly in large-scale FL deployments involving millions of heterogeneous devices and data sources [123]. Furthermore, balancing fairness with privacy-preserving mechanisms, such as differential privacy, remains an open research problem, as the added noise introduced by privacy techniques often disproportionately impacts smaller or marginalized datasets [11].

To advance fairness in federated learning, future research must explore adaptive algorithms that dynamically balance fairness and utility in the face of diverse data distributions and resource constraints. Furthermore, interdisciplinary efforts are essential to integrate ethical and regulatory frameworks into the design of fairness-aware FL systems [20]. As FL continues to be deployed in critical sectors such as healthcare and finance, ensuring ethically and equitably distributed benefits across all participants is both a technical and societal imperative. By fostering innovation at the intersection of algorithmics, domain-specific expertise, and social principles, FL systems can meet the dual demands of inclusivity and performance in real-world applications.

8.3 Federated Unlearning and the Right to be Forgotten

Federated unlearning has emerged as a critical area of research in federated learning (FL) to address the pressing challenges of regulatory compliance, privacy preservation, and user autonomy. The "right to be forgotten," enshrined in legal frameworks like the General Data Protection Regulation (GDPR), necessitates mechanisms for the effective removal of individual contributions from machine learning models upon request. However, implementing this in FL introduces unique complexities stemming from its decentralized, privacy-preserving nature, where local data remains inaccessible to the central aggregator. This subsection delves into the technical underpinnings of federated unlearning, examines state-of-the-art approaches, and highlights the open challenges and opportunities for innovation.

The core objective of federated unlearning is to remove the influence of specific user data or updates on the global model without requiring full retraining, which would be computationally prohibitive and might unintentionally result in privacy violations due to analysis of remaining model updates. Techniques for federated unlearning can broadly be classified into exact unlearning and approximate unlearning. Exact unlearning entails mechanisms to simulate the

removal of client contributions such that the updated global model resembles one that would have been trained without those contributions. However, as many FL models rely on iterative updates, maintaining reversibility while adhering to constraints such as privacy, communication efficiency, and heterogeneity remains a significant challenge [124].

Exact federated unlearning mechanisms often employ tree-based retraining or gradient subtraction techniques. Tree-based retraining systematically excludes specific client updates from prior training rounds to recompute the model. Despite the theoretical accuracy, such methods are computationally expensive in large-scale FL systems, especially under non-independent and identically distributed (non-IID) data settings [8]. Gradient subtraction, on the other hand, focuses on constructing the inverse of previously shared gradients to "subtract" their contributions. While computationally more efficient compared to full retraining, gradient-based approaches risk reduction in model integrity or performance when handling high-dimensional or redundant parameter spaces [125]. Furthermore, ensuring that gradient subtraction aligns effectively for heterogeneous clients and models requires more sophisticated estimators, giving rise to trade-offs between computational tractability and unlearning accuracy.

Approximate federated unlearning techniques have been proposed to balance efficiency and compliance requirements. These methods rely on techniques such as influence function estimation or residual learning mechanisms. Influence functions approximate the contribution of a particular client to the global objective function, allowing for selective modification or removal. Estimations, however, suffer from errors that may accumulate through multiple rounds of aggregation, calling into question the guarantees of compliance with legal standards like GDPR [1], [124]. Residual learning introduces an intermediate deletion model trained on perturbed data that mimics the absence of the target client. While scalable, residual-based approaches may not generalize well to scenarios with high data heterogeneity or dynamic client participation [70].

Security concerns further complicate federated unlearning. Malicious or adversarial unlearning requests, whereby attackers seek to disrupt model performance or inject biases through repeated unlearning demands, pose a significant threat. Defense mechanisms such as unlearning request verification and trust evaluation protocols can mitigate these risks [44]. Yet, designing robust detection frameworks within FL ecosystems is non-trivial due to the distributed nature of training and limited access to client-level data.

Emerging directions in federated unlearning research include hybrid techniques that integrate cryptographic protocols and differential privacy for iterative influence removal. For instance, differential privacy encourages aggregate-level obfuscation, reducing the sensitivity of the model to individual contributions and thus diminishing the need for precise unlearning [11]. Concurrently, blockchain-based FL frameworks offer enhanced auditability and secure logging, ensuring traceable unlearning operations [16]. Federated representation learning, where abstract representations are shared instead of raw gradients, also presents a promising approach to minimize the impact of data-specific reversibility, thus simplifying unlearning tasks [71].

Looking ahead, a central research challenge lies in achieving verifiable unlearning with guaranteed compliance to regulatory demands while maintaining operational efficiency. Developing adaptive unlearning frameworks capable of dynamically handling new client participation or model evolution is critical in ensuring scalability. Moreover, benchmarking frameworks that standardize the evaluation of federated unlearning techniques—considering metrics such as time complexity, privacy leakage, and legal compliance—will drive progress in the field [36].

In conclusion, federated unlearning is a nascent but rapidly evolving domain that encapsulates fundamental questions at the intersection of machine learning, privacy law, and distributed systems. By addressing its current limitations, federated unlearning promises to bolster the ethical and practical deployment of federated learning systems in real-world scenarios while safeguarding user privacy.

8.4 Advancing Multi-modal and Cross-modal Federated Learning

The increasing complexity and diversity of real-world data necessitate advancements in federated learning (FL) architectures capable of integrating multi-modal and cross-modal data, such as text, images, sensor readings, and audio. Multi-modal federated learning (MMFL) represents a transformative leap forward in the evolution of FL systems, enabling the development of more holistic and generalized models to address complex tasks in domains such as healthcare, autonomous systems, and cross-device intelligence. This subsection explores the challenges, methodologies, and future directions shaping MMFL and its realization in practical applications.

Multi-modal FL introduces unique complexities due to the heterogeneous data representations and varying learning architectures encountered across distributed clients. Unlike unimodal FL, where all clients contribute to training on a single data modality (e.g., text or images), MMFL must accommodate clients with individual modalities, overlapping subsets of modalities, or even missing modalities entirely. This disparity leads to challenges in synchronization and scalability, as well as inefficiencies in collaborative learning. For instance, clients lacking access to specific modalities pose obstacles when aligning shared information across participants. To address this issue, methods like modality-specific feature alignment through shared latent spaces or joint embeddings have been investigated [126], enabling more cohesive information sharing while maintaining the integrity of decentralized learning.

The fusion of cross-modal data within FL further demands advanced techniques to extract complementary insights from heterogeneous data types, enabling robust learning. Contrastive learning frameworks that align embedding spaces across modalities and federated representation learning approaches have shown potential for facilitating cross-modal aggregation [126]. Despite their promise, these methods contend with limitations such as communication overhead and computational cost, as aligning complex data structures often requires iterative optimization across diverse clients. To mediate these difficulties, layer-wise alignment strategies have emerged, allowing specific model

layers to be decoupled for handling distinct modalities while maintaining consistent feature extraction in the global model [126]. Such designs have shown encouraging results in merging modalities like text-vision and sensor-audio while remaining sensitive to the constraints of decentralized systems.

One of the critical challenges in MMFL is ensuring model utility and robustness under resource-constrained settings, particularly when edge devices with limited computational power are involved. Lightweight strategies, such as parameter-efficient fine-tuning methods, have demonstrated potential by enabling incremental inclusion of modalities into pre-trained global models while significantly reducing communication and computation demands [43], [100]. These methods focus on fine-tuning a subset of parameters tailored to each client's modality, making scalable multi-modal deployment feasible for cross-device environments and personalized applications.

Empirical findings from federated simulations further underscore the trade-offs and opportunities in MMFL. Personalization techniques, such as local fine-tuning or clustering specific clients based on modality, have consistently outperformed generic aggregation strategies in diverse use cases. Similarly, federated transfer learning with modality-specific sub-models has proven integral in addressing feature heterogeneity in applications like federated medical imaging, which integrates radiology reports and diagnostic scans [29]. However, despite these advances, achieving fully generalized multi-modal models continues to be constrained by the lack of effective techniques capable of systematizing fusion and generalization across highly non-IID data modalities.

Emerging methodologies present promising solutions to these challenges. For instance, cross-modal attention mechanisms that dynamically weigh modality contributions during aggregation can bolster global model robustness by prioritizing high-confidence modalities [72]. In addition, hierarchical FL architectures, where early-stage aggregations are performed among modality-similar clients, show potential to alleviate the scaling issues inherent to MMFL [17]. These approaches not only improve computational efficiency but also facilitate modality-specific learning in large-scale settings.

A particularly exciting avenue for future work lies in Federated Foundation Models (FFMs), which focus on integrating cross-modal knowledge at the pre-training stage, reducing the need for redundant fusion during downstream applications [74]. Such models offer the potential to enable end-to-end MMFL workflows that seamlessly operate across domains while adhering to privacy-preserving principles. Furthermore, cross-disciplinary research incorporating approaches from fields like neuro-symbolic learning and multimodal cognitive modeling could inspire novel paradigms for representation learning tailored to distributed multimodal systems.

In conclusion, MMFL represents a significant step toward achieving holistic intelligence in federated learning. Foundational work on modality alignment, model fusion, and resource-efficient training has laid the groundwork for progress, but critical challenges remain in scaling these frameworks while addressing non-IID heterogeneity and equitable client participation. Bridging these challenges and advancing innovative architectural designs will pave the way for MMFL's transformative impact across diverse industries, ensuring its applicability in real-world scenarios while upholding FL's privacy-preserving ethos.

8.5 Integrating Ethical, Legal, and Interdisciplinary Dimensions

As federated learning (FL) continues to gain prominence as a transformative machine learning paradigm, integrating ethical, legal, and interdisciplinary dimensions is critical to fostering trust, fairness, and societal alignment. This integration is particularly vital given the distributed and privacy-preserving nature of FL, which amplifies challenges regarding compliance with regulatory frameworks, alignment with ethical principles, and cross-disciplinary applicability.

At the forefront, compliance with data protection regulations such as the GDPR in the European Union and the CCPA in the United States emphasizes the need for strong privacy guarantees within FL frameworks. Mechanisms like differential privacy (DP) and secure aggregation have emerged as foundational tools in this regard [11], [24]. While DP provides formal mathematical guarantees to restrict data leakage, its integration in FL is often constrained by trade-offs between privacy and model utility. Additionally, achieving regulatory compliance in diverse jurisdictions is complicated by overlapping yet distinct policy mandates, such as the varying thresholds for acceptable re-identification risks. The prospect of "federated unlearning" as a practical solution to implement the "right to be forgotten" stipulated by GDPR is gaining attention. Though unlearning frameworks leveraging retraining minimization or incremental updates offer promise, their computational overhead in dynamic ecosystems remains a significant bottleneck. Research must explore lightweight yet rigorous mechanisms to streamline compliance across global deployments.

Ethically, FL inherits challenges commonly associated with artificial intelligence (AI) systems, such as ensuring fairness and avoiding systematic biases. The heterogeneity of client datasets in FL presents unique challenges in achieving algorithmic equity. Systems designed without explicit consideration of fairness risks can inadvertently marginalize underrepresented demographic groups or amplify existing inequities, as demonstrated by disparities in global model performance under non-IID conditions [112], [127]. Recent efforts have investigated fairness-aware aggregation techniques, client selection protocols, and trust-weighted optimization strategies designed to mitigate bias even under data heterogeneity [50]. However, these approaches often struggle with scalability and degrade overall utility in environments with extreme dataset imbalances. Moreover, frameworks for fairness that obfuscate sensitive client demographics raise further questions about their ability to address equity without explicit auditing of demographic parity.

Interdisciplinary contributions from fields such as law, sociology, and ethics have further refined the conceptual boundaries of FL. Legal scholars highlight tensions between decentralized learning and liability, raising questions

of attribution and accountability for erroneous or non-compliant model outputs. Ethical AI frameworks, prevalent in the domain of centralized machine learning, must now evolve to cater to the distributed architecture of FL. For instance, participatory design approaches involving stakeholders—including client institutions, regulators, and end-users—can help ensure socially acceptable deployments [1]. Additionally, sociological perspectives propose viewing FL deployment contexts through community-oriented lenses to reflect local social norms, but translating these norms into globally applicable system designs remains complex.

Transparency is another critical aspect to fostering trust in FL systems. Trust-building mechanisms such as auditable aggregation protocols, explainable AI models, and decentralized trust management are under exploration. Blockchain-based solutions for tamper-proof contribution logs offer potential pathways forward, though such systems must address operational inefficiencies and scalability concerns [47]. Moreover, the adoption of federated simulation platforms has emerged as a useful interdisciplinary tool for testing the impact of ethical principles and regulatory constraints under different deployment scenarios [76].

Future directions should emphasize holistic frameworks that simultaneously align FL with evolving legal and ethical standards while leveraging interdisciplinary synergies. For instance, advancing adaptive, jurisdiction-aware privacy mechanisms that balance compliance and efficiency will be key. Furthermore, ethical FL solutions must incorporate cultural differences, ensuring that the models support inclusivity in geographically dispersed communities. Finally, forging collaborations between policymakers, technologists, and ethicists will be pivotal in codifying global governance frameworks tailored to FL—a step critical to its responsible adoption at scale. By addressing these dimensions, federated learning has the potential to redefine AI's role in fostering equitable, privacy-preserving, and ethically sound machine learning ecosystems.

8.6 Future Innovations in Security and Adversarial Robustness

The ever-evolving landscape of federated learning (FL) faces persistent challenges in ensuring security and robustness against both adversarial and inadvertent risks, necessitating innovative approaches to strengthen the framework. This subsection examines recent advancements and future opportunities in mitigating adversarial threats, securing communication, and fortifying the robustness of FL systems, while situating these efforts within broader interdisciplinary and technological trajectories.

Federated learning's decentralized architecture offers intrinsic privacy advantages but simultaneously introduces unique vulnerabilities, such as model poisoning, backdoor attacks, and gradient leakage. Adversarial clients can exploit the framework by injecting malicious updates that impair global model performance or introduce backdoors. For instance, model replacement approaches have demonstrated how a single malicious client can embed backdoors in FL models with near-perfect success across minimal attack rounds [128]. Traditional defenses, such as anomaly detection and Byzantine-resilient aggregation algorithms, have shown utility in combating such threats. However, methods like Multi-KRUM and Adaptive Federated Averaging face significant computational overhead and lack resilience against increasingly sophisticated attacks in dynamic, heterogeneous environments [129]. To keep pace with adversarial advancements, scalable and adaptive defense mechanisms capable of evolving alongside the threat landscape are necessary.

Beyond adversarial inputs, vulnerabilities in communication channels remain a significant concern. Techniques like secure aggregation and encryption protocols (e.g., homomorphic encryption) have demonstrated efficacy in preventing direct exposure of individual updates during transmission. Still, more advanced attacks, such as inference by malicious servers exploiting aggregated updates, present additional risks [56]. The integration of hybrid encryption protocols, possibly layered with blockchain-based solutions, emerges as a promising direction for enhancing the confidentiality and integrity of client-server communication [130]. Such innovations will be critical in bolstering secure inference mechanisms and ensuring resilient collaboration across disparate nodes in FL systems.

Another pressing issue involves defending against gradient leakage and data reconstruction attacks, wherein adversarial actors exploit shared model updates to recover sensitive input data. Studies illustrate how generative techniques such as GANs enable adversaries to reconstruct raw data, even at the batch level, from model gradients [86], [131]. Existing defenses, including differential privacy (DP), can mitigate these threats by perturbing gradients with noise. However, striking a balance between maintaining model utility and ensuring robust privacy guarantees remains particularly difficult in non-IID and highly imbalanced client settings [80]. To address these challenges, research into adaptive DP frameworks and innovative representation-level defenses is gaining momentum. Approaches like federated representation learning, which shift the focus from task-specific parameter exchanges to shared feature representations, show promise in reducing attack vectors while preserving prediction performance [132]. Adaptive noise mechanisms and frameworks like PRECAD, which dynamically modulate noise injection based on attack scenarios, represent forward-looking strategies for scaling defenses to diverse real-world contexts [92], [133].

Enhancing adversarial robustness further necessitates proactive and holistic strategies that prioritize secure-by-design principles. For instance, the incorporation of zero-knowledge proof systems enables integrity verification for client updates without compromising confidentiality [61]. Such mechanisms align well with frameworks like Cronus, which reduce the dimensionality of shared updates to minimize poisoning risks while maintaining computational efficiency [88]. Detection-oriented methods, such as clustering-based update filtering and collaborative anomaly detection models, are also gaining traction in safeguarding global model integrity against sophisticated poisoning or backdoor insertion efforts [129]. By localizing and distributing detection via client-side models, decentralized robustness can be improved through enhanced trust-based aggregation protocols [108].

Looking ahead, embedding resilience into federated

systems extends beyond addressing immediate security challenges. Integrating FL into broader AI ecosystems, such as edge computing and autonomous applications, presents opportunities for robust, context-aware learning while mitigating security challenges in distributed settings [134]. Moreover, interdisciplinary approaches—leveraging insights from cryptography, adversarial ML, and ethical AI—are essential in crafting comprehensive solutions for robust and secure FL deployments. For example, aligning representation learning techniques with ethical guidelines could further mitigate risks of data misuse and privacy infringement.

In conclusion, fortifying the security and adversarial robustness of federated learning requires coordinated efforts that embrace adaptiveness, scalability, and interdisciplinary innovation. As threats to federated systems grow more sophisticated, integrating proactive defense mechanisms, secure communication protocols, and robust learning frameworks will be instrumental in sustaining FL's promise as a privacy-preserving, trustworthy, and globally adaptable machine learning paradigm. These advancements lay the groundwork for developing resilient FL systems that harmonize technical innovation with functional and ethical imperatives, complementing the broader cross-disciplinary advancements explored in subsequent sections.

8.7 Cross-disciplinary and Technological Synergies

Federated Learning (FL), given its decentralized architecture and privacy-preserving possibilities, presents unparalleled potential when synergized with advancements in adjacent technological fields and cross-disciplinary approaches. These integrations can not only address the prevailing operational challenges in FL but also propel its applications into untapped domains, ultimately redefining its societal impact.

One critical avenue of synergy exists between FL and Internet of Things (IoT) networks, which involve vast ecosystems of edge devices generating heterogeneous and often privacy-sensitive data. Integrating FL with IoT can distribute computational workloads across geographically dispersed nodes, reducing central processing demands while safeguarding data privacy [44]. However, challenges such as resource limitations and communication constraints in IoT devices demand optimization strategies, such as model compression and adaptive client selection. Techniques like fragmented federated learning [57] show promise in addressing these issues by mixing update fragments across participants for increased robustness against poisoning attacks while maintaining IoT energy efficiency.

In autonomous and adaptive systems such as robotics, FL's ability to enable collaborative learning in dynamic environments is increasingly being explored. Autonomous systems, such as fleets of self-driving vehicles, require models that continuously adapt to heterogeneous data distributions arising from diverse environmental contexts. However, achieving low-latency real-time updates while mitigating poisoning attacks remains a substantial challenge. Robust strategies such as Byzantine-resilient aggregation mechanisms [135] and ensemble-based FL frameworks [136] could form the bedrock of secure, adaptive learning in these systems by ensuring model integrity despite adversarial interference or uncertainty in communication graphs.

Emerging applications of FL in healthcare and financial services further underline its cross-disciplinary importance. In healthcare, FL fosters secure multi-institutional collaborations by enabling shared model development while adhering to data protection regulations [34]. For instance, brain tumor segmentation in radiology or personalized medicine outcomes are achievable without direct data sharing between hospitals [137]. Conversely, the financial sector demands equitable access to federated credit-scoring models while maintaining privacy. Vertical FL tailored to disjoint feature collaboration [138] can solve such sector-specific needs but introduces complexities in managing inter-party trust and cross-feature alignment. As pointed out in recent evaluations [139], cross-disciplinary strategies must solve operational barriers—such as inferring secure knowledge across silos—without increasing collusion risks during inter-organizational cooperation.

Moreover, FL is poised to enable multimodal learning, where cross-modal data integration (e.g., combining text, images, and audio) unlocks holistic applications such as intelligent recommendation systems and biometrics-based security protocols [140]. Building effective multi-modal architectures to process incomplete or imbalanced modalities in client data presents a distinguished challenge. Contrastive learning techniques and model alignment strategies are being actively explored [134], but their computational costs and susceptibility to inference attacks demand further innovation. Cryptographic methods like homomorphic encryption or zero-knowledge proofs [61] can provide robust operational guarantees in such resource-heavy multimodal deployments.

From a broader perspective, cross-disciplinary synergies extend to legal, ethical, and societal dimensions. For example, integrating FL with sociological and legal frameworks is crucial in ensuring compliance with privacy legislation like GDPR and CPRA, while also fostering trust through enhanced transparency. Initiatives such as differential privacy mechanisms [13] provide essential mathematical guarantees of privacy but must be adapted contextually for real-world regulations, as they often involve trade-offs in model accuracy. Further, ethical concerns in algorithmic biases necessitate fairness-driven loss functions and equitable data representation frameworks [20].

Finally, FL's integration with cutting-edge fields like generative adversarial networks (GANs) and data-efficient AI opens novel research directions. GANs, despite their adversarial exploitation in privacy attacks [141], can potentially be leveraged to build synthetic yet high-utility datasets to improve FL performance while maintaining strong privacy guarantees [81]. Similarly, federated reinforcement learning holds promise for complex adaptive systems but requires foundational enhancements to FL's current optimization and bandwidth management frameworks [137].

In conclusion, the intersections between federated learning and emerging technologies or disciplines present unparalleled opportunities, but effectively leveraging them will require meticulously balancing technical performance, privacy guarantees, scalability, and ethical compliance. Future research must embrace truly interdisciplinary perspectives to unlock both the operational potential and broader societal benefits of federated learning.

9 Conclusion

Federated learning (FL) has emerged as a cornerstone paradigm in privacy-preserving collaborative machine learning, addressing fundamental challenges in decentralized model training while adhering to strict data privacy requirements. This survey has provided a comprehensive exploration of the core architectures, privacy-preserving mechanisms, security challenges, and diverse applications that underpin the FL paradigm. In this concluding subsection, we synthesize the lessons learned, contextualize the significance of FL's advancements, and identify promising future research trajectories.

The key contributions of federated learning lie in its ability to balance utility and privacy by relegating data processing to local devices and sharing only model updates for aggregation. Its foundational mechanisms—such as federated averaging (FedAvg) for model optimization—have demonstrated practical efficacy across diverse domains while minimizing privacy risks. However, significant trade-offs persist, including the tension between resource-constrained environments and communication overheads [17], as well as the challenges introduced by non-IID data distributions, particularly in robustness and fairness [7], [8]. These findings underscore the necessity of innovative system designs that can dynamically adapt to heterogeneity in both client data and device capabilities.

Our analysis identifies that privacy-preserving techniques like differential privacy (DP), secure multiparty computation (SMPC), and homomorphic encryption (HE) are integral components of FL for protecting sensitive data in adversarial contexts. Differential privacy introduces formalized guarantees, but its effectiveness must be carefully balanced against the introduction of noise, which can hinder model utility, particularly in complex machine learning scenarios [11]. By contrast, SMPC and HE-based approaches ensure cryptographic integrity with minimal data transparency, albeit at the expense of significant computational overheads [142]. Emerging hybrid methodologies combining cryptographic and non-cryptographic techniques exemplify the multidisciplinary innovations driving FL forward [34].

While FL has experienced rapid advancements in practical deployments, as evidenced in fields such as healthcare [4] and finance [6], challenges related to scalability, bias mitigation, and adversarial robustness remain unresolved. For instance, the limitations of centralized aggregators reveal vulnerabilities to single points of failure, motivating the study of decentralized and blockchain-integrated federated systems, which offer fault tolerance and transparency [15]. However, such decentralized frameworks introduce new privacy and trust challenges, particularly in environments with untrusted participants [27].

The future of FL research lies at the confluence of several emerging trends. First, personalized federated learning (PFL) represents a transformative approach for client-specific optimization, addressing issues of disparity in model performance across heterogeneous datasets [9]. Similarly, federated unlearning stands out as a nascent domain critical for aligning FL with regulatory frameworks like GDPR by enabling the selective removal of individual contributions from trained models [93]. Moreover, cross-modal federated learning, which integrates diverse data types such as text, images, and sensors, has the potential to expand FL's applicability across interdisciplinary contexts, though it necessitates new methods for multimodal alignment and integration [143].

Interdisciplinary collaboration will be pivotal as FL continues to gain traction. Addressing scalability requires infrastructural advancements such as hierarchical aggregation strategies and efficient communication protocols to reduce latency and resource consumption [1]. Simultaneously, fairness-aware methodologies that disaggregate client-specific biases without requiring access to sensitive demographic information are vital for fostering equitable FL systems [117].

In conclusion, federated learning represents a paradigm shift in collaborative machine learning, redefining possibilities for privacy-preserving technologies while opening avenues for societal, industrial, and academic applications. Yet, its transformative potential will be fully realized only through the resolution of its inherent challenges. Future research must focus on synergizing theoretical advancements and practical implementations, ensuring that FL evolves into a robust, secure, and equitable framework. As industries continue to embrace FL, its role as a cornerstone of privacy-respecting AI will be further solidified, underscoring its importance in the era of data ubiquity.

References

[1] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, pp. 50–60, 2019. 1, 6, 25, 29, 30, 32, 34

[2] N. Rodr'iguez-Barroso, D. J. L'opez, M. V. Luz'on, F. Herrera, and E. Martínez-Cámara, "Survey on federated learning threats: concepts, taxonomy on attacks and defences, experimental study and challenges," *ArXiv*, vol. abs/2201.08135, 2022. 1, 5, 6, 14, 18, 19

[3] T. Yang, G. Andrew, H. Eichner, H. Sun, W. Li, N. Kong, D. Ramage, and F. Beaufays, "Applied federated learning: Improving google keyboard query suggestions," *ArXiv*, vol. abs/1812.02903, 2018. 1

[4] J. Xu, B. Glicksberg, C. Su, P. B. Walker, J. Bian, and F. Wang, "Federated learning for healthcare informatics," *Journal of Healthcare Informatics Research*, vol. 5, pp. 1 – 19, 2019. 1, 4, 6, 7, 8, 11, 19, 24, 34

[5] Z. Sun, Y. Xu, Y. Liu, W. He, Y. Jiang, F. Wu, and L. zhen Cui, "A survey on federated recommendation systems," *IEEE transactions on neural networks and learning systems*, vol. PP, 2022. 1

[6] P. Mammen, "Federated learning: Opportunities and challenges," *ArXiv*, vol. abs/2101.05428, 2021. 1, 2, 6, 22, 34

[7] Q. Li, Y. Diao, Q. Chen, and B. He, "Federated learning on non-iid data silos: An experimental study," *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pp. 965–978, 2021. 1, 2, 6, 9, 15, 19, 24, 27, 29, 34

[8] H. Zhu, J. Xu, S. Liu, and Y. Jin, "Federated learning on non-iid data: A survey," *ArXiv*, vol. abs/2106.06843, 2021. 1, 2, 3, 6, 7, 8, 22, 24, 25, 28, 29, 30, 34

[9] A. Tan, H. Yu, L. zhen Cui, and Q. Yang, "Towards personalized federated learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, pp. 9587–9603, 2021. 1, 5, 12, 14, 15, 24, 25, 34

[10] V. Kulkarni, M. Kulkarni, and A. Pant, "Survey of personalization techniques for federated learning," *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*, pp. 794–797, 2020. 1, 11, 12, 16, 20, 26, 29

[11] J. Fu, Y. Hong, X. Ling, L. Wang, X. Ran, Z. Sun, W. H. Wang, Z. Chen, and Y. Cao, "Differentially private federated learning: A systematic review," *ArXiv*, vol. abs/2405.08299, 2024. 1, 2, 4, 5, 6, 7, 8, 9, 10, 11, 12, 19, 21, 25, 28, 29, 30, 31, 34

[12] A. Bhowmick, J. C. Duchi, J. Freudiger, G. Kapoor, and R. M. Rogers, "Protection against reconstruction and its applications in private federated learning," *ArXiv*, vol. abs/1812.00984, 2018. 1, 5, 13, 23

[13] M. Naseri, J. Hayes, and E. D. Cristofaro, "Local and central differential privacy for robustness and privacy in federated learning," *Proceedings 2022 Network and Distributed System Security Symposium*, 2020. 1, 9, 33

[14] K. Cheng, T. Fan, Y. Jin, Y. Liu, T. Chen, and Q. Yang, "Secureboost: A lossless federated learning framework," *IEEE Intelligent Systems*, vol. 36, pp. 87–98, 2019. 1, 2, 24

[15] C. Ma, J. Li, M. Ding, L. Shi, T. Wang, Z. Han, and H. Poor, "When federated learning meets blockchain: A new distributed learning paradigm," *IEEE Computational Intelligence Magazine*, vol. 17, pp. 26–33, 2020. 1, 2, 19, 28, 34

[16] J. Liu, C. Chen, Y. Li, L. Sun, Y. Song, J. Zhou, B. Jing, and D. Dou, "Enhancing trust and privacy in distributed networks: A comprehensive survey on blockchain-based federated learning," *ArXiv*, vol. abs/2403.19178, 2024. 1, 4, 9, 26, 30

[17] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konecný, S. Mazzocchi, H. B. McMahan, T. V. Overveldt, D. Petrou, D. Ramage, and J. Roselander, "Towards federated learning at scale: System design," *ArXiv*, vol. abs/1902.01046, 2019. 1, 2, 3, 4, 11, 14, 26, 28, 31, 34

[18] L. Witt, M. Heyer, K. Toyoda, W. Samek, and D. Li, "Decentral and incentivized federated learning frameworks: A systematic literature review," *IEEE Internet of Things Journal*, vol. 10, pp. 3642–3663, 2022. 1

[19] L. Yang, D. Chai, J. Zhang, Y. Jin, L. Wang, H. Liu, H. Tian, Q. Xu, and K. Chen, "A survey on vertical federated learning: From a layered perspective," *ArXiv*, vol. abs/2304.01829, 2023. 2, 5, 25

[20] Y. Zhang, D. Zeng, J. Luo, Z. Xu, and I. King, "A survey of trustworthy federated learning with perspectives on security, robustness and privacy," *Companion Proceedings of the ACM Web Conference 2023*, 2023. 2, 3, 29, 33

[21] L. U. Khan, W. Saad, Z. Han, E. Hossain, and C. Hong, "Federated learning for internet of things: Recent advances, taxonomy, and open challenges," *IEEE Communications Surveys & Tutorials*, vol. 23, pp. 1759–1799, 2020. 2, 3, 4, 5, 8, 9, 11, 14, 16, 17, 22, 25, 26, 28

[22] J. Konecný, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," *ArXiv*, vol. abs/1610.02527, 2016. 3, 29

[23] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. Bhagoji, K. Bonawitz, Z. B. Charles, G. Cormode, R. Cummings, R. G. L. D'Oliveira, S. E. Rouayheb, D. Evans, J. Gardner, Z. Garrett, A. Gascón, B. Ghazi, P. B. Gibbons, M. Gruteser, Z. Harchaoui, C. He, L. He, Z. Huo, B. Hutchinson, J. Hsu, M. Jaggi, T. Javidi, G. Joshi, M. Khodak, J. Konecný, A. Korolova, F. Koushanfar, O. Koyejo, T. Lepoint, Y. Liu, P. Mittal, M. Mohri, R. Nock, A. Özgür, R. Pagh, M. Raykova, H. Qi, D. Ramage, R. Raskar, D. Song, W. Song, S. U. Stich, Z. Sun, A. Suresh, F. Tramèr, P. Vepakomma, J. Wang, L. Xiong, Z. Xu, Q. Yang, F. X. Yu, H. Yu, and S. Zhao, "Advances and open problems in federated learning," *Found. Trends Mach. Learn.*, vol. 14, pp. 1–210, 2019. 3, 10, 11, 20

[24] Y. Zheng, S. Lai, Y. Liu, X. Yuan, X. Yi, and C. Wang, "Aggregation service for federated learning: An efficient, secure, and more resilient realization," *IEEE Transactions on Dependable and Secure Computing*, vol. 20, pp. 988–1001, 2022. 3, 20, 25, 31

[25] Z. Zhao, Y. Mao, Y. Liu, L. Song, O. Ye, X. Chen, and W. Ding, "Towards efficient communications in federated learning: A contemporary survey," *J. Frankl. Inst.*, vol. 360, pp. 8669–8703, 2022. 3, 15

[26] S. K.M., S. Nicolazzo, M. Arazzi, A. Nocera, R. R. K.A., V. P., and M. Conti, "Privacy-preserving in blockchain-based federated learning systems," *ArXiv*, vol. abs/2401.03552, 2024. 3, 10, 15, 20, 24, 25

[27] E. Hallaji, R. Razavi-Far, M. Saif, B. Wang, and Q. Yang, "Decentralized federated learning: A survey on security and privacy," *IEEE Transactions on Big Data*, vol. 10, pp. 194–213, 2024. 3, 34

[28] B. Liu, N. Lv, Y. Guo, and Y. Li, "Recent advances on federated learning: A systematic survey," *Neurocomputing*, vol. 597, p. 128019, 2023. 3, 28

[29] A. Rauniyar, D. Hagos, D. Jha, J. E. Haakegaard, U. Bagci, D. Rawat, and V. Vlassov, "Federated learning for medical applications: A taxonomy, current trends, challenges, and future research directions," *IEEE Internet of Things Journal*, vol. 11, pp. 7374–7398, 2022. 3, 5, 7, 9, 24, 26, 31

[30] W. Guo, F. Zhuang, X. Zhang, Y. Tong, and J. Dong, "A comprehensive survey of federated transfer learning: Challenges, methods and applications," *Frontiers Comput. Sci.*, vol. 18, p. 186356, 2024. 3, 12, 25

[31] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *International Conference on Artificial Intelligence and Statistics*, 2016, pp. 1273–1282. 3, 4, 11, 26

[32] L. Lyu, H. Yu, and Q. Yang, "Threats to federated learning: A survey," *ArXiv*, vol. abs/2003.02133, 2020. 3, 6, 7, 8, 19

[33] W. Jin, Y. Yao, S. Han, C. Joe-Wong, S. Ravi, A. Avestimehr, and C. He, "Fedml-he: An efficient homomorphic-encryption-based privacy-preserving federated learning system," *ArXiv*, vol. abs/2303.10837, 2023. 3, 7, 8, 10, 20, 25, 27

[34] L. Lyu, H. Yu, X. Ma, L. Sun, J. Zhao, Q. Yang, and P. S. Yu, "Privacy and robustness in federated learning: Attacks and defenses," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, pp. 8726–8746, 2020. 3, 5, 6, 7, 10, 19, 20, 24, 26, 33, 34

[35] J. Wang, Z. B. Charles, Z. Xu, G. Joshi, H. B. McMahan, B. A. Y. Arcas, M. Al-Shedivat, G. Andrew, S. Avestimehr, K. Daly, D. Data, S. Diggavi, H. Eichner, A. Gadhikar, Z. Garrett, A. M. Girgis, F. Hanzely, A. S. Hard, C. He, S. Horváth, Z. Huo, A. Ingerman, M. Jaggi, T. Javidi, P. Kairouz, S. Kale, S. P. Karimireddy, J. Konecný, S. Koyejo, T. Li, L. Liu, M. Mohri, H. Qi, S. J. Reddi, P. Richtárik, K. Singhal, V. Smith, M. Soltanolkotabi, W. Song, A. Suresh, S. U. Stich, A. Talwalkar, H. Wang, B. E. Woodworth, S. Wu, F. X. Yu, H. Yuan, M. Zaheer, M. Zhang, T. Zhang, C. Zheng, C. Zhu, and W. Zhu, "A field guide to federated optimization," *ArXiv*, vol. abs/2107.06917, 2021. 4, 9, 14, 16, 17, 25, 28

[36] D. Chai, L. Wang, L. Yang, J. Zhang, K. Chen, and Q. Yang, "A survey for federated learning evaluations: Goals and measures," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, pp. 5007–5024, 2023. 4, 7, 20, 30

[37] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller, "Inverting gradients - how easy is it to break privacy in federated learning?" *ArXiv*, vol. abs/2003.14053, 2020. 4, 5, 6, 7, 9, 10, 19, 20

[38] Y. Li, C. Chen, N. Liu, H. Huang, Z. Zheng, and Q. Yan, "A blockchain-based decentralized federated learning framework with committee consensus," *IEEE Network*, vol. 35, pp. 234–241, 2020. 4, 20

[39] F. Hanzely and P. Richtárik, "Federated learning of a mixture of global and local models," *ArXiv*, vol. abs/2002.05516, 2020. 4, 6, 11, 12, 16, 26, 29

[40] T. Yu, E. Bagdasarian, and V. Shmatikov, "Salvaging federated learning by local adaptation," *ArXiv*, vol. abs/2002.04758, 2020. 4, 11

[41] R. Kanagavelu, Z. Li, J. Samsudin, Y. Yang, F. Yang, R. Goh, M. Cheah, P. Wiwatphonthana, K. Akkarajitsakul, and S. Wang, "Two-phase multi-party computation enabled privacy-preserving federated learning," *2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID)*, pp. 410–419, 2020. 4, 8, 11, 21, 26

[42] B. Liu, Y. Guo, and X. Chen, "Pfa: Privacy-preserving federated adaptation for effective model personalization," *Proceedings of the Web Conference 2021*, 2021. 4

[43] C. Zhang, G. Long, H. Guo, X. Fang, Y. Song, Z. Liu, G. Zhou, Z. Zhang, Y. Liu, and B. Yang, "Federated adaptation for foundation model-based recommendations," *ArXiv*, vol. abs/2405.04840, 2024. 4, 27, 31

[44] Y. Chen, Y. Gui, H. Lin, W. Gan, and Y. Wu, "Federated learning attacks and defenses: A survey," *2022 IEEE International Conference on Big Data (Big Data)*, pp. 4256–4265, 2022. 5, 6, 7, 9, 10, 19, 20, 26, 28, 30, 33

[45] S. Maddock, G. Cormode, T. Wang, C. Maple, and S. Jha, "Federated boosted decision trees with differential privacy," *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2022. 5, 9, 24

[46] R. Gupta and A. K. Singh, "A differential approach for data and classification service-based privacy-preserving machine learning model in cloud environment," *New Generation Computing*, vol. 40, pp. 737 – 764, 2022. 5

[47] E. Gabrielli, G. Pica, and G. Tolomei, "A survey on decentralized federated learning," *ArXiv*, vol. abs/2308.04604, 2023. 6, 9, 10, 32

[48] S. Galbraith, C. Petit, B. Shani, and Y. Ti, "On the security of supersingular isogeny cryptosystems," in *International Conference on the Theory and Application of Cryptology and Information Security*, 2016, pp. 63–91. 6

[49] M. Ye, W. Shen, B. Du, E. Snezhko, V. Kovalev, and P. Yuen, "Vertical federated learning for effectiveness, security, applicability: A survey," *ArXiv*, vol. abs/2405.17495, 2024. 7

[50] W. Huang, M. Ye, Z. Shi, G. Wan, H. Li, B. Du, and Q. Yang, "Federated learning for generalization, robustness, fairness: A survey and benchmark," *IEEE transactions on pattern analysis and machine intelligence*, vol. PP, 2023. 7, 14, 15, 17, 19, 21, 22, 31

[51] Z. Li, S. He, Z. Yang, M. Ryu, K. Kim, and R. Madduri, "Advances in appfl: A comprehensive and extensible federated learning framework," *ArXiv*, vol. abs/2409.11585, 2024. 7, 11, 20

[52] Z. Zhang, A. Pinto, V. Turina, F. Esposito, and I. Matta, "Privacy and efficiency of communications in federated split learning," *IEEE Transactions on Big Data*, vol. 9, pp. 1380–1391, 2023. 7, 21, 26

[53] H. Chen and H. Vikalo, "Recovering labels from local updates in federated learning," *ArXiv*, vol. abs/2405.00955, 2024. 7, 21, 27

[54] Y. Fraboni, R. Vidal, and M. Lorenzi, "Free-rider attacks on model aggregation in federated learning," *ArXiv*, vol. abs/2006.11901, 2020. 7, 27

[55] J. So, C. He, C.-S. Yang, S. Li, Q. Yu, R. E. Ali, B. Guler, and S. Avestimehr, "Lightsecagg: a lightweight and versatile design for secure aggregation in federated learning," in *Conference on Machine Learning and Systems*, 2021. 7, 11, 12, 21

[56] D. Pasquini, D. Francati, and G. Ateniese, "Eluding secure aggregation in federated learning via model inconsistency," *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2021. 7, 21, 32

[57] N. Jebreel, J. Domingo-Ferrer, A. Blanco-Justicia, and D. Sánchez, "Enhanced security and privacy via fragmented federated learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, pp. 6703–6717, 2022. 7, 33

[58] M. Ryu and K. Kim, "Differentially private federated learning via inexact admm," *ArXiv*, vol. abs/2106.06127, 2021. 8, 22, 26

[59] S. Ji, W. Jiang, A. Walid, and X. Li, "Dynamic sampling and selective masking for communication-efficient federated learning," *IEEE Intelligent Systems*, vol. 37, pp. 27–34, 2020. 8, 16, 21, 26

[60] Y. Lu, Z. Yu, and N. Suri, "Privacy-preserving decentralized federated learning over time-varying communication graph," *ACM Transactions on Privacy and Security*, vol. 26, pp. 1 – 39, 2022. 8, 16, 21, 22

[61] A. Chowdhury, C. Guo, S. Jha, and L. Maaten, "Eiffel: Ensuring integrity for federated learning," *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2021. 8, 32, 33

[62] S. Vithana and S. Ulukus, "Private read update write (pruw) in federated submodel learning (fsl): Communication efficient schemes with and without sparsification," *IEEE Transactions on Information Theory*, vol. 70, pp. 1320–1348, 2022. 8

[63] Y. Liu, Y. Kang, T. Zou, Y. Pu, Y. He, X. Ye, Y. Ouyang, Y. Zhang, and Q. Yang, "Vertical federated learning: Concepts, advances, and challenges," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, pp. 3615–3634, 2022. 8

[64] J. So, B. Guler, and A. Avestimehr, "Turbo-aggregate: Breaking the quadratic aggregation barrier in secure federated learning," *IEEE Journal on Selected Areas in Information Theory*, vol. 2, pp. 479–489, 2020. 8, 22

[65] F. Fu, H. Xue, Y. Cheng, Y. Tao, and B. Cui, "Blindfl: Vertical federated machine learning without peeking into your data," *Proceedings of the 2022 International Conference on Management of Data*, 2022. 8

[66] B. Soltani, V. Haghighi, A. Mahmood, Q. Sheng, and L. Yao, "A survey on participant selection for federated learning in mobile networks," *Proceedings of the 17th ACM Workshop on Mobility in the Evolving Internet Architecture*, 2022. 8, 15, 17, 22

[67] M. Ali, F. Naeem, M. Tariq, and G. Kaddoum, "Federated learning for privacy preservation in smart healthcare systems: A comprehensive survey," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, pp. 778–789, 2022. 9, 14, 24

[68] X. Shang, Y. Lu, G. Huang, and H. Wang, "Federated learning on heterogeneous and long-tailed data via classifier re-training with federated features," in *International Joint Conference on Artificial Intelligence*, 2022, pp. 2218–2224. 10, 25, 29

[69] K. Z. Liu, S. Hu, Z. S. Wu, and V. Smith, "On privacy and personalization in cross-silo federated learning," *ArXiv*, vol. abs/2206.07902, 2022. 10, 29

[70] J. Liu, J. Huang, Y. Zhou, X. Li, S. Ji, H. Xiong, and D. Dou, "From distributed machine learning to federated learning: a survey," *Knowledge and Information Systems*, vol. 64, pp. 885 – 917, 2021. 11, 25, 30

[71] L. Wang, Y. Zhao, J. Dong, A. Yin, Q. Li, X. Wang, D. Niyato, and Q. Zhu, "Federated learning with new knowledge: Fundamentals, advances, and futures," *ArXiv*, vol. abs/2402.02268, 2024. 11, 30

[72] T. Lin, L. Kong, S. U. Stich, and M. Jaggi, "Ensemble distillation for robust model fusion in federated learning," *ArXiv*, vol. abs/2006.07242, 2020. 11, 27, 31

[73] T. Zhou, J. Zhang, and D. H. K. Tsang, "Fedfa: Federated learning with feature anchors to align features and classifiers for heterogeneous data," *IEEE Transactions on Mobile Computing*, vol. 23, pp. 6731–6742, 2022. 11, 16

[74] S. Yu, J. P. Muñoz, and A. Jannesari, "Federated foundation models: Privacy-preserving and collaborative learning for large models," in *International Conference on Language Resources and Evaluation*, 2023, pp. 7174–7184. 12, 17, 27, 31

[75] P. Liang, T. Liu, L. Ziyin, R. Salakhutdinov, and L.-P. Morency, "Think locally, act globally: Federated learning with local and global representations," *ArXiv*, vol. abs/2001.01523, 2020. 12, 29

[76] L. Collins, H. Hassani, A. Mokhtari, and S. Shakkottai, "Exploiting shared representations for personalized federated learning," *ArXiv*, vol. abs/2102.07078, 2021. 12, 27, 29, 32

[77] S. Itahara, T. Nishio, Y. Koda, M. Morikura, and K. Yamamoto, "Distillation-based semi-supervised federated learning for communication-efficient collaborative training with non-iid private data," *IEEE Transactions on Mobile Computing*, vol. 22, pp. 191–205, 2020. 12, 17

[78] K. Pfeiffer, M. Rapp, R. Khalili, and J. Henkel, "Federated learning for computationally constrained heterogeneous devices: A survey," *ACM Computing Surveys*, vol. 55, pp. 1 – 27, 2023. 12, 17, 20

[79] W. Wei, L. Liu, J. Zhou, K.-H. Chow, and Y. Wu, "Securing distributed sgd against gradient leakage threats," *IEEE Transactions on Parallel and Distributed Systems*, vol. 34, pp. 2040–2054, 2023. 13, 18, 23

[80] S. Pentyala, D. Railsback, R. Maia, R. Dowsley, D. Melanson, A. C. A. Nascimento, and M. D. Cock, "Training differentially private models with secure multiparty computation," *IACR Cryptol. ePrint Arch.*, vol. 2022, p. 146, 2022. 13, 32

[81] Z. Zhao, M. Luo, and W. Ding, "Deep leakage from model in federated learning," *ArXiv*, vol. abs/2206.04887, 2022. 13, 33

[82] D. I. Dimitrov, M. Baader, M. N. Muller, and M. T. Vechev, "Spear: Exact gradient inversion of batches in federated learning," *ArXiv*, vol. abs/2403.03945, 2024. 13

[83] A. Elkordy, J. Zhang, Y. H. Ezzeldin, K. Psounis, and A. Avestimehr, "How much privacy does federated learning with secure aggregation guarantee?" *Proc. Priv. Enhancing Technol.*, vol. 2023, pp. 510–526, 2022. 13

[84] G. Zhu, D. Li, H. Gu, Y. Han, Y. Yao, L. Fan, and Q. Yang, "Evaluating membership inference attacks and defenses in federated learning," *ArXiv*, vol. abs/2402.06289, 2024. 13

[85] W. Wei, L. Liu, M. Loper, K.-H. Chow, M. E. Gursoy, S. Truex, and Y. Wu, "A framework for evaluating gradient leakage attacks in federated learning," *ArXiv*, vol. abs/2004.10397, 2020. 13, 19

[86] B. Hitaj, G. Ateniese, and F. Pérez-Cruz, "Deep models under the gan: Information leakage from collaborative deep learning," *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017. 13, 32

[87] Z. Li, J. Zhang, L. Liu, and J. Liu, "Auditing privacy defenses in federated learning via generative gradient leakage," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 122–10 132, 2022. 13, 23

[88] H. Chang, V. Shejwalkar, R. Shokri, and A. Houmansadr, "Cronus: Robust and heterogeneous collaborative learning with black-box knowledge transfer," *ArXiv*, vol. abs/1912.11279, 2019. 13, 18, 32

[89] A. Hatamizadeh, H. Yin, P. Molchanov, A. Myronenko, W. Li, P. Dogra, A. Feng, M. G. Flores, J. Kautz, D. Xu, and H. Roth, "Do gradient inversion attacks make federated learning unsafe?" *IEEE Transactions on Medical Imaging*, vol. 42, pp. 2044–2056, 2022. 13

[90] J. C. Zhao, A. Dabholkar, A. Sharma, and S. Bagchi, "Leak and learn: An attacker's cookbook to train using leaked data from federated learning," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12 247–12 256, 2024. 13

[91] J. Zhang, Y. Chen, and H. H. Li, "Privacy leakage of adversarial training models in federated learning systems," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 107–113, 2022. 13

[92] K. Yue, R. Jin, C.-W. Wong, D. Baron, and H. Dai, "Gradient obfuscation gives a false sense of security in federated learning," in *USENIX Security Symposium*, 2022, pp. 6381–6398. 13, 23, 32

[93] N. Romandini, A. Mora, C. Mazzocca, R. Montanari, and P. Bellavista, "Federated unlearning: A survey on methods, design guidelines, and evaluation metrics," *IEEE transactions on neural networks and learning systems*, vol. PP, 2024. 13, 23, 28, 34

[94] J. Shao, Z. Li, W. Sun, T. Zhou, Y. Sun, L. Liu, Z. Lin, and J. Zhang, "A survey of what to share in federated learning: Perspectives on model utility, privacy leakage, and communication efficiency," *ArXiv*, vol. abs/2307.10655, 2023. 14, 28

[95] L. Fu, H. Zhang, G. Gao, M. Zhang, and X. Liu, "Client selection in federated learning: Principles, challenges, and opportunities," *IEEE Internet of Things Journal*, vol. 10, pp. 21 811–21 819, 2022. 15, 29

[96] W. Yang, N. Wang, Z. Guan, L. Wu, X. Du, and M. Guizani, "A practical cross-device federated learning framework over 5g networks," *IEEE Wireless Communications*, vol. 29, pp. 128–134, 2022. 15

[97] K. Singhal, H. Sidahmed, Z. Garrett, S. Wu, K. Rush, and S. Prakash, "Federated reconstruction: Partially local federated learning," in *Neural Information Processing Systems*, 2021, pp. 11 220–11 232. 15

[98] Z. Zhu, Y. Shi, J. Luo, F. Wang, C. Peng, P. Fan, and K. Letaief, "Fedlp: Layer-wise pruning mechanism for communication-computation efficient federated learning," *ICC 2023 - IEEE International Conference on Communications*, pp. 1250–1255, 2023. 16

[99] R. Lin, Y. Xiao, T.-J. Yang, D. Zhao, L. Xiong, G. Motta, and F. Beaufays, "Federated pruning: Improving neural network efficiency with federated learning," in *Interspeech*, 2022, pp. 1701–1705. 16

[100] G. Sun, M. Mendieta, T. Yang, and C. Chen, "Conquering the communication constraints to enable large pre-trained models in federated learning," 2022. 16, 31

[101] I. Tenison, S. A. Sreeramadas, V. Mugunthan, E. Oyallon, E. Belilovsky, and I. Rish, "Gradient masked averaging for federated learning," *Trans. Mach. Learn. Res.*, vol. 2023, 2022. 16, 20

[102] Y. Xiong, R. Wang, M. Cheng, F. Yu, and C.-J. Hsieh, "Feddm: Iterative distribution matching for communication-efficient federated learning," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16 323–16 332, 2022. 16, 21

[103] O. Marfoq, G. Neglia, L. Kameni, and R. Vidal, "Personalized federated learning through local memorization," *ArXiv*, vol. abs/2111.09360, 2021. 16

[104] J. Tan, Y. Zhou, G. Liu, J. H. Wang, and S. Yu, "pfedsim: Similarity-aware model aggregation towards personalized federated learning," *ArXiv*, vol. abs/2305.15706, 2023. 16

[105] H. Ochiai, Y. Sun, Q. Jin, N. Wongwiwatchai, and H. Esaki, "Wireless ad hoc federated learning: A fully distributed cooperative machine learning," *ArXiv*, vol. abs/2205.11779, 2022. 17

[106] W. Bao, H. Wang, J. Wu, and J. He, "Optimizing the collaboration structure in cross-silo federated learning," *ArXiv*, vol. abs/2306.06508, 2023. 17

[107] L. Zhao, Q. Wang, Q. Zou, Y. Zhang, and Y. Chen, "Privacy-preserving collaborative deep learning with unreliable participants," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1486–1500, 2018. 18

[108] Z. Xing, Z. Zhang, Z. Zhang, J. Liu, L. Zhu, and G. Russello, "No vandalism: Privacy-preserving and byzantine-robust federated learning," *ArXiv*, vol. abs/2406.01080, 2024. 18, 23, 32

[109] J. So, B. Guler, A. Avestimehr, and P. Mohassel, "Codedprivateml: A fast and privacy-preserving framework for distributed machine learning," *IEEE Journal on Selected Areas in Information Theory*, vol. 2, pp. 441–451, 2019. 18

[110] Y. Liu, J. J. Q. Yu, J. Kang, D. Niyato, and S. Zhang, "Privacy-preserving traffic flow prediction: A federated learning approach," *IEEE Internet of Things Journal*, vol. 7, pp. 7751–7763, 2020. 19

[111] N. Bastianello, C. Liu, and K. H. Johansson, "Enhancing privacy in federated learning through local training," *ArXiv*, vol. abs/2403.17572, 2024. 20

[112] D. Gao, X. Yao, and Q. Yang, "A survey on heterogeneous federated learning," *ArXiv*, vol. abs/2210.04505, 2022. 20, 29, 31

[113] J. Yang, Y. Zhao, and L. Wang, "A survey of federated unlearning: A taxonomy, challenges and future directions," *ArXiv*, vol. abs/2310.19218, 2023. 21

[114] M. Ryu, Y. Kim, K. Kim, and R. K. Madduri, "Appfl: Open-source software framework for privacy-preserving federated learning," *2022 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pp. 1074–1083, 2022. 22

[115] G. Andrew, P. Kairouz, S. Oh, A. Oprea, H. B. McMahan, and V. M. Suriyakumar, "One-shot empirical privacy estimation for federated learning," *ArXiv*, vol. abs/2302.03098, 2023. 23

[116] L. H. Fowl, J. Geiping, W. Czaja, M. Goldblum, and T. Goldstein, "Robbing the fed: Directly obtaining private data in federated learning with modified models," *ArXiv*, vol. abs/2110.13057, 2021. 23

[117] T. H. Rafi, F. A. Noor, T. Hussain, and D.-K. Chae, "Fairness and privacy-preserving in federated learning: A survey," *ArXiv*, vol. abs/2306.08402, 2023. 24, 29, 34

[118] N.-H. Nguyen, T. Nguyen, T. Nguyen, V. T. Hoang, D. D. Le, and K.-S. Wong, "Towards efficient communication and secure federated recommendation system via low-rank training," *Proceedings of the ACM on Web Conference 2024*, 2024. 25

[119] S. Savazzi, M. Nicoli, M. Bennis, S. Kianoush, and L. Barbieri, "Opportunities of federated learning in connected, cooperative, and automated industrial systems," *IEEE Communications Magazine*, vol. 59, pp. 16–21, 2021. 26

[120] H. Ludwig, N. Baracaldo, G. Thomas, Y. Zhou, A. Anwar, S. Rajamoni, Y. J. Ong, J. Radhakrishnan, A. Verma, M. Sinn, M. Purcell, A. Rawat, T. Minh, N. Holohan, S. Chakraborty, S. Whitherspoon, D. Steuer, L. Wynter, H. Hassan, S. Laguna, M. Yurochkin, M. Agarwal, E. Chuba, and A. Abay, "Ibm federated learning: an enterprise framework white paper v0.1," *ArXiv*, vol. abs/2007.10987, 2020. 27

[121] J. Xu, X.-Y. Tong, and S.-L. Huang, "Personalized federated learning with feature alignment and classifier collaboration," *ArXiv*, vol. abs/2306.11867, 2023. 27

[122] S. A. Osia, A. S. Shamsabadi, S. Sajadmanesh, A. Taheri, K. Katevas, H. Rabiee, N. Lane, and H. Haddadi, "A hybrid deep learning architecture for privacy-preserving mobile analytics," *IEEE Internet of Things Journal*, vol. 7, pp. 4505–4518, 2017. 28

[123] Y. Shi, J. Liang, W. Zhang, V. Y. F. Tan, and S. Bai, "Towards understanding and mitigating dimensional collapse in heterogeneous federated learning," *ArXiv*, vol. abs/2210.00226, 2022. 29

[124] H. Jeong, S. Ma, and A. Houmansadr, "Sok: Challenges and opportunities in federated unlearning," *ArXiv*, vol. abs/2403.02437, 2024. 30

[125] X. Xie, C. Hu, H. Ren, and J. Deng, "A survey on vulnerability of federated learning: A learning algorithm perspective," *Neurocomputing*, vol. 573, p. 127225, 2023. 30

[126] T. Zhou and E. Konukoglu, "Fedfa: Federated feature augmentation," *ArXiv*, vol. abs/2301.12995, 2023. 30, 31

[127] U. Milasheuski, L. Barbieri, B. C. Tedeschini, M. Nicoli, and S. Savazzi, "On the impact of data heterogeneity in federated learning environments with application to healthcare networks," *2024 IEEE Conference on Artificial Intelligence (CAI)*, pp. 1017–1023, 2024. 31

[128] E. Bagdasarian, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," *ArXiv*, vol. abs/1807.00459, 2018. 32

[129] N. Jebreel, J. Domingo-Ferrer, D. Sánchez, and A. Blanco-Justicia, "Defending against the label-flipping attack in federated learning," *ArXiv*, vol. abs/2207.01982, 2022. 32

[130] J. Liang and R. Wang, "Fedcip: Federated client intellectual property protection with traitor tracking," *ArXiv*, vol. abs/2306.01356, 2023. 32

[131] Z. Wang, J. D. Lee, and Q. Lei, "Reconstructing training data from model gradient, provably," *ArXiv*, vol. abs/2212.03714, 2022. 32

[132] J. Sun, A. Li, B. Wang, H. Yang, H. Li, and Y. Chen, "Soteria: Provable defense against privacy leakage in federated learning from representation perspective," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9307–9315, 2020. 32

[133] Z. Shen, J. Ye, A. Kang, H. Hassani, and R. Shokri, "Share your representation only: Guaranteed improvement of the privacy-utility tradeoff in federated learning," *ArXiv*, vol. abs/2309.05505, 2023. 32
[134] P. Ye, Z. Jiang, W. Wang, B. Li, and B. Li, "Feature reconstruction attacks and countermeasures of dnn training in vertical federated learning," *ArXiv*, vol. abs/2210.06771, 2022. 33
[135] J. So, B. Guler, and A. Avestimehr, "Byzantine-resilient secure federated learning," *IEEE Journal on Selected Areas in Communications*, vol. 39, pp. 2168–2181, 2020. 33
[136] X. Cao, J. Jia, and N. Gong, "Provably secure federated learning against malicious clients," *ArXiv*, vol. abs/2102.01854, 2021. 33
[137] Y. Li, Z. Guo, N. Yang, H. Chen, D. Yuan, and W. Ding, "Threats and defenses in federated learning life cycle: A comprehensive survey and challenges," *ArXiv*, vol. abs/2407.06754, 2024. 33
[138] X. Luo, Y. Wu, X. Xiao, and B. Ooi, "Feature inference attack on model predictions in vertical federated learning," *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pp. 181–192, 2020. 33
[139] E. Hallaji, R. Razavi-Far, and M. Saif, "Federated and transfer learning: A survey on adversaries and defense mechanisms," *ArXiv*, vol. abs/2207.02337, 2022. 33
[140] W. Zhang, S. Tople, and O. Ohrimenko, "Leakage of dataset properties in multi-party machine learning," in *USENIX Security Symposium*, 2020, pp. 2687–2704. 33
[141] L. Melis, C. Song, E. D. Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 691–706, 2018. 33
[142] T. Ryffel, A. Trask, M. Dahl, B. Wagner, J. V. Mancuso, D. Rueckert, and J. Passerat-Palmbach, "A generic framework for privacy preserving deep learning," *ArXiv*, vol. abs/1811.04017, 2018. 34
[143] N. R. Barroso, G. Stipcich, D. Jiménez-López, J. A. Ruiz-Millán, E. Martínez-Cámara, G. González-Seco, M. V. Luzón, M. Veganzones, and F. Herrera, "Federated learning and differential privacy: Software tools analysis, the sherpa.ai fl framework and methodological guidelines for preserving data privacy," *ArXiv*, vol. abs/2007.00914, 2020. 34