Journal of
Sensor and Actuator Networks

MDPI

Article

Heterogeneity Challenges of Federated Learning for Future Wireless Communication Networks

Lorena Isabel Barona López [1,*] and Thomás Borja Saltos [1,2]

[1] Artificial Intelligence and Computer Vision Research Lab, Departamento de Informática y Ciencias de la Computación, Escuela Politécnica Nacional, Quito 170525, Ecuador; thomas.borja@epn.edu.ec
[2] Faculty of Agricultural Sciences, Natural Resources and Environment, Universidad Estatal de Bolívar, Guaranda 020150, Ecuador
* Correspondence: lorena.barona@epn.edu.ec

Abstract: Two technologies of great interest in recent years—Artificial Intelligence (AI) and massive wireless communication networks—have found a significant point of convergence through Federated Learning (FL). Federated Learning is a Machine Learning (ML) technique that enables multiple participants to collaboratively train a model while keeping their data local. Several studies indicate that while improving performance metrics—such as accuracy, loss reduction, or computation time—is a primary goal, achieving this in real-world scenarios remains challenging. This difficulty arises due to various heterogeneity characteristics inherent to the wireless devices participating in the Federation. Heterogeneity in Federated Learning arises when participants contribute differently, leading to challenges in the model training process. Heterogeneity in Federated Learning may appear in architecture, statistics, and behavior. System heterogeneity arises from differences in device capabilities, including processing power, transmission speeds, availability, energy constraints, and network limitations, among others. Statistical heterogeneity occurs when participants contribute non-independent and non-identically distributed (non-IID) data. This situation can harm the global model instead of improving it, especially when the data are of poor quality or too scarce. The third type, behavioral heterogeneity, refers to cases where participants are unwilling to engage or expect rewards despite minimal effort. Given the growing research in this area, we present a summary of heterogeneity characteristics in Federated Learning to provide a broader perspective on this emerging technology. We also outline key challenges, opportunities, and future directions for Federated Learning. Finally, we conduct a simulation using the LEAF framework to illustrate the impact of heterogeneity in Federated Learning.

Keywords: behavioral heterogeneity; federated learning; mobile communications; statistical heterogeneity; system heterogeneity; wireless communications

check for updates

Academic Editors: Laura Verde, Fiammetta Marulli, Rosario Catelli and Giovanni Paragliola

Received: 6 January 2025
Revised: 11 February 2025
Accepted: 19 February 2025
Published: 1 April 2025

Citation: Barona López, L.I.; Borja Saltos, T. Heterogeneity Challenges of Federated Learning for Future Wireless Communication Networks. *J. Sens. Actuator Netw.* **2025**, *14*, 37. https://doi.org/10.3390/jsan14020037

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The advancement of Artificial Intelligence (AI) through Machine Learning (ML) has led to the emergence of innovative applications that continue to engage researchers, industry experts, and users worldwide. Over the years, AI has experienced significant growth, enabling applications such as computer vision, predictive systems, recommendation algorithms, generative text models, and conversational agents exhibiting human-like behavior. The horizon of AI continues to expand daily.

Simultaneously, scientific interest is growing in the development of high-speed wireless networks capable of supporting massive data exchange. Current wireless networks

encompass two widely studied technologies: 5G+/6G cellular networks and the Internet of Things (IoT). Including Machine Learning processes for 5G+ networks is expected and considered an essential requirement for 6G networks. Studies suggest that 6G will depend on the ubiquitous integration of Artificial Intelligence, enabling data-driven Machine Learning solutions within large-scale heterogeneous networks [1]. In this scenario, each mobile user has a device with processing capabilities and access to the network. Together with Machine Learning, this gives rise to a new learning paradigm called Federated Learning (FL).

The evolution toward 6G networks is fundamentally dependent on Artificial Intelligence and Machine Learning, which play a crucial role in enabling autonomous, scalable, and intelligent network operations. Unlike previous generations, where AI primarily optimized performance, in 6G, AI is deeply integrated into the network architecture, enabling innovations such as intelligent spectrum management, dynamic resource allocation, self-optimizing networks, and decentralized learning paradigms like Federated Learning [2]. AI integration in 6G is driven by the need to manage large-scale heterogeneous networks while ensuring energy efficiency, ultra-low latency, and security. Studies have emphasized that FL, as an AI-driven decentralized learning approach, is essential for enabling intelligence at the edge, reducing reliance on centralized data aggregation, and preserving privacy while ensuring adaptive model training [3]. Furthermore, in Industrial IoT (IIoT) and dynamic edge environments, AI-driven models are essential for real-time decision-making, fault detection, and predictive maintenance. AI-based optimizations ensure efficient communication between ultra-dense networks and devices, helping mitigate heterogeneity challenges in resource-constrained environments [4]. This idea further reinforces the argument that AI is not merely a supporting technology but an indispensable component of 6G deployment.

Traditionally, Machine Learning works in a centralized scheme and processes data on-site. Federated Learning introduces an architecture where each device in a wireless network participates in training a shared model. The learning model is trained using one or several Machine Learning techniques, such as Deep Neural Networks, Convolutional Networks, and Q-learning, among others. The participating users train their models locally, and then a Federated Learning server aggregates these models globally and redistributes the updated model to each participant. In this way, users obtain a more refined and precise model while keeping their personal information locally [5]. Decentralization makes it possible to connect classic Machine Learning techniques with massive wireless communication networks, such as 5G/6G or even IoT. Federated Learning will enable mobile devices to benefit from more accurate and diverse models quickly, securely, and even in real-time. A key challenge is that participating devices exhibit diverse processing capabilities, transmission speeds, network availability, coverage, energy levels, data volume, and willingness to participate. This diversity in characteristics, capabilities, and behaviors is referred to as heterogeneity in Federated Learning.

Heterogeneity is an inherent characteristic of distributed Machine Learning systems. Due to disparities among participants, heterogeneity can introduce several challenges: The primary challenge of heterogeneity in Federated Learning is ensuring model compatibility and maintaining performance across participating devices or nodes. Beyond performance degradation and compatibility issues, heterogeneity can introduce biases in trained models and increase communication overhead, resulting in longer training times and higher energy consumption.

Heterogeneity in Federated Learning is broadly categorized into three distinct types: system heterogeneity, statistical heterogeneity, and behavioral heterogeneity. System heterogeneity arises from differences in device capabilities, such as processing power, memory, and network conditions, which can lead to imbalances in computational contributions and communication delays. Statistical heterogeneity refers to variations in data distri-

butions across participants, often characterized by non-independent and non-identically distributed (non-IID) data, which can hinder model convergence and fairness. Behavioral heterogeneity, on the other hand, encompasses variations in participant behavior, including willingness to participate, incentive-driven engagement, and adversarial actions such as free-riding or data manipulation. Unlike system and statistical heterogeneity, which are rooted in physical and data-related constraints, behavioral heterogeneity is driven by human or organizational factors, making it a unique and critical challenge in FL. By explicitly distinguishing these categories, this study aims to provide a comprehensive taxonomy of heterogeneity challenges and their implications for FL performance.

Figure 1 illustrates a heterogeneous Federated Learning system architecture. A central server initializes the model, which is sent to participating devices for local training. The updated models are aggregated at the server, allowing all participants to benefit from the globally trained model.

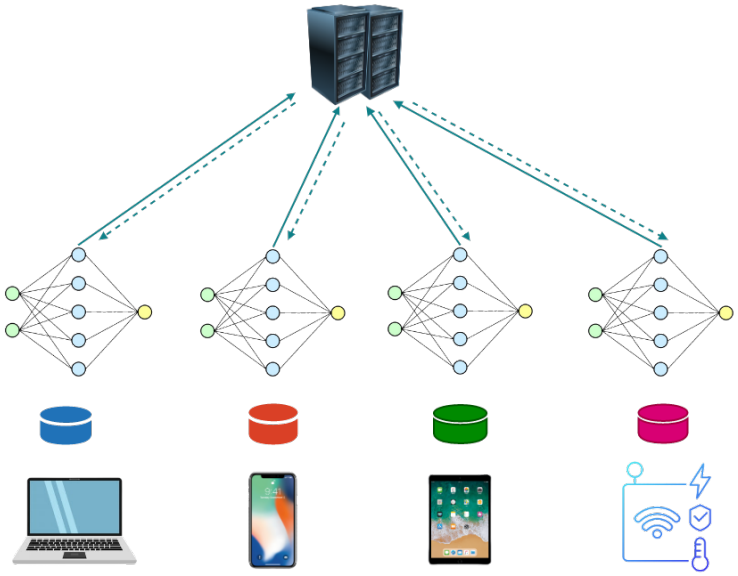


Figure 1. Heterogeneous Federated Learning architecture, where each participant may differ in hardware characteristics (i.e., computers, tablets, IoT devices), data distributions (represented by using different colors), or even have different network conditions or behaviors.

Figure 1 highlights the impact of heterogeneity in FL, where clients exhibit variability in hardware capabilities, network conditions, and data distributions. The heterogeneity can be categorized into four primary sources:

- **Device-level heterogeneity:** Clients differ in computational power, battery life, and memory availability. Some devices, such as high-performance edge servers, can process model updates efficiently, while resource-constrained IoT devices may struggle with timely participation. This imbalance can lead to training inefficiencies, as some clients may contribute more significantly to the global model than others.
- **Network-level heterogeneity:** Communication constraints, bandwidth limitations, and network instability further impact FL training. Devices connected via 5G networks may transmit updates quickly, whereas those on unstable Wi-Fi or cellular connections may experience delays or packet loss. These discrepancies can slow down the aggregation process and introduce synchronization challenges in synchronous FL approaches. Device-level heterogeneity and network-level heterogeneity are both aspects of system heterogeneity, which we later classify under a unified category in our article.

- **Statistical heterogeneity:** The data distribution across clients is often non-IID (non-independent and identically distributed), meaning that different clients may have data with varying feature distributions, label distributions, or sample sizes. This heterogeneity arises from real-world factors such as demographic differences, regional preferences, or domain-specific variations. As a result, local models trained on client data may diverge significantly from each other, making global aggregation challenging. Effective strategies to mitigate statistical heterogeneity include personalization techniques, adaptive weighting in federated averaging, and robust optimization methods designed for non-IID settings.
- **Behavioral heterogeneity:** Differences in client behavior affect the consistency and fairness of the FL process. Some clients may intermittently participate due to energy constraints, while others may attempt to free-ride by benefiting from the model without contributing meaningful updates. In addition, malicious clients may introduce unreliable data, affecting the integrity of the training process. Addressing these issues requires incentive mechanisms, trust-based aggregation, and robust anomaly detection techniques.

1.1. Contextualization of the Study and Novelty of Behavioral Heterogeneity in FL

Federated Learning has emerged as a promising paradigm for training machine learning models in a decentralized manner while preserving data privacy. Over the years, research in FL has extensively explored key challenges such as statistical heterogeneity, system constraints, and optimization techniques to enhance model performance across diverse client devices. Recent studies have primarily focused on overcoming issues related to non-IID data distributions, resource-constrained devices, and scalable communication protocols [4,6,7].

Despite these advances, behavioral heterogeneity has remained a relatively underexplored dimension in FL research. While the existing literature discusses challenges such as unreliable network conditions and non-uniform computing resources, there has been a limited emphasis on the human-driven factors influencing FL performance. Behavioral heterogeneity encompasses critical issues such as client participation willingness, incentive mechanisms, reputation-based aggregation, and adversarial behaviors—all of which directly impact model reliability and fairness [8,9].

This study aims to bridge this gap by introducing a comprehensive taxonomy of FL heterogeneity challenges, explicitly integrating behavioral heterogeneity alongside system and statistical heterogeneity. Unlike prior surveys, which often focus on individual aspects of heterogeneity, our work systematically classifies behavioral influences in FL, highlighting their implications on model convergence, fairness, and security. Furthermore, we incorporate recent advancements in incentive-driven participation models and trust-based reputation aggregation mechanisms, which have not been addressed sufficiently in prior FL taxonomies [10,11].

By providing a structured analysis of behavioral heterogeneity, this work not only contextualizes its significance within the broader FL research landscape but also paves the way for future studies on adaptive incentive frameworks, trust-aware aggregation, and robust client-selection strategies in decentralized learning environments.

1.2. Related Work

Several surveys have explored heterogeneity in Federated Learning, often focusing on specific aspects such as statistical, system, or model heterogeneity. Ye et al. [12] provide a systematic review of heterogeneous Federated Learning, covering statistical, model, communication, and device heterogeneity but omitting behavioral heterogeneity, such as

participation inconsistency and free-riding. Similarly, Nguyen et al. [4] discuss FL for IoT networks, emphasizing massive data distribution but without addressing behavioral or fairness challenges. Lim et al. [6] explore device heterogeneity in mobile edge computing, focusing on communication costs and resource allocation constraints. Li et al. [7] discuss the system and statistical heterogeneity, highlighting challenges like hardware disparities, non-IID data distribution, and communication overhead in large-scale networks.

Other works further contribute to heterogeneity analysis. Abdelmoniem et al. [8] introduce behavioral heterogeneity, analyzing client participation patterns based on device conditions such as charging status or connectivity. Beitollahi et al. [13] examine FL challenges in wireless environments, highlighting device power limitations, bandwidth constraints, and privacy risks. Meanwhile, Mengistu et al. [14] focus on FL in IoT, studying communication overhead and data imbalance. Pei et al. [9] discuss the role of 6G in FL but primarily from a theoretical perspective, without practical validation.

While these studies provide valuable insights, our work distinguishes itself by introducing behavioral heterogeneity as a key challenge in FL and offering a broader classification that connects system, statistical, and behavioral heterogeneity in wireless network environments. In addition, we extend existing taxonomies by considering the implications of network dynamics, fairness, and security in FL systems, offering a comprehensive perspective that complements previous research. Table 1 presents a summary of the Related Work.

Limitations of Existing Studies

While several studies have explored heterogeneity in Federated Learning, most existing research has focused on statistical and system heterogeneity, with limited attention to the behavioral aspects of FL participation. Key foundational works, such as FedAvg and TiFL, have contributed significantly to the development of FL, yet they present notable limitations in addressing the full spectrum of heterogeneity challenges.

FedAvg, introduced as a baseline FL aggregation algorithm, operates under the assumption that all clients contribute equally to the global model and assumes uniform computational capabilities and reliable network conditions [5]. However, in real-world FL deployments, devices vary significantly in terms of processing power, data availability, and communication latency. This discrepancy leads to straggler issues, slower convergence, and potential model bias due to imbalanced client participation [7].

TiFL, on the other hand, introduced a tier-based FL framework aimed at mitigating the effects of system heterogeneity by categorizing clients into performance-based tiers [15]. While TiFL improves training efficiency by prioritizing faster clients for aggregation, it does not fully address behavioral heterogeneity, such as free-riding behavior, adversarial client participation, or incentive-driven collaboration mechanisms [8]. Furthermore, TiFL's tier-based approach may lead to model divergence, as lower-tier clients contribute significantly less to the final global model.

Our work extends beyond these approaches by introducing a comprehensive taxonomy that explicitly incorporates behavioral heterogeneity as a critical factor affecting FL performance. Unlike existing frameworks that focus primarily on system-level constraints or statistical data variations, we classify behavioral factors such as willingness to participate, reputation-based aggregation, and malicious behaviors as integral components of FL heterogeneity. This structured taxonomy enables a more holistic understanding of heterogeneity, facilitating the design of adaptive incentive mechanisms and fairness-aware aggregation strategies [10,11]. By addressing these previously overlooked behavioral dimensions, our study contributes to bridging the gap in FL research and provides a complete framework for developing robust and equitable FL systems.

1.3. Contributions of This Paper

In the process of writing this article, we identified a substantial number of studies addressing heterogeneity in Federated Learning. Given that heterogeneity is a critical factor influencing the practical deployment of FL, minimizing its impact is essential for fully realizing its potential.

The challenges associated with heterogeneity span multiple dimensions, making it crucial to define and categorize these challenges with precision. To address this, we present a comprehensive taxonomy of heterogeneity issues in FL, structured into three primary categories: system heterogeneity, statistical heterogeneity, and behavioral heterogeneity. Each category encompasses key challenges, which we analyze in detail throughout this paper.

Our primary contributions can be summarized as follows:

1. **Comprehensive Taxonomy of Heterogeneity in FL**: We provide an extensive and structured taxonomy of heterogeneity challenges in FL, supported by relevant literature. In addition, we discuss potential mitigation strategies and their effectiveness in different FL scenarios.
2. **Inclusion of Behavioral Heterogeneity**: While existing works predominantly focus on system and statistical heterogeneity, we introduce behavioral heterogeneity as a fundamental aspect influencing FL performance. To the best of our knowledge, this study is among the first to incorporate behavioral heterogeneity into a formal taxonomy of FL challenges.
3. **Expansion of Challenges in Wireless FL Environments**: We extend the discussion beyond standard classifications by incorporating network infrastructure constraints, energy limitations, and massively distributed data, which are crucial considerations for FL deployment in wireless networks, 6G, and IoT ecosystems.
4. **Summary of FL Frameworks, Algorithms, and Tools**: We compile and analyze existing FL frameworks, algorithms, and benchmarking tools, providing insights into their suitability for addressing heterogeneity in FL. This summary serves as both a starting point for new researchers and a reference for experienced practitioners seeking to integrate heterogeneity-aware approaches into their work.

The subsequent sections systematically classify the challenges posed by heterogeneity, detailing their characteristics and implications. Various forms of heterogeneity can introduce severe issues, including loss of model accuracy, convergence difficulties, increased computational overhead, elevated network and processor usage, and security vulnerabilities. Our paper reviews key studies and solutions proposed to mitigate these challenges. In addition, we examine recent research efforts that quantify the impact of heterogeneity in real-world FL deployments, both statistically and structurally. Finally, we outline potential research opportunities and future directions for improving FL robustness and efficiency. The final section of this document presents the conclusions.

2. Heterogeneity in Federated Learning

Traditional Machine Learning frameworks that rely on centralized servers face significant privacy and security challenges, including single points of failure [1]. As a result, such frameworks cannot enable ubiquitous and secure AI for the upcoming 6G networks. Moreover, centralized ML methods that aggregate and process data at a central location introduce significant overhead, making them unsuitable for applications requiring ubiquitous computing. As a result, decentralized ML solutions that retain private data on local devices are increasingly essential for the success of future wireless networks.

Table 1. Summary of Related Work.

Survey	Reference	Key Contributions	Difference from Our Work
Heterogeneous Federated Learning: State-of-the-art and research challenges	Ye et al. [12]	They provide a systematic review on heterogeneous Federated Learning, summarizing challenges from statistical, model, communication, and device perspectives.	We provide the behavioral heterogeneity category and add additional challenges in the statistical heterogeneity category.
Federated Learning for Internet of Things: A comprehensive survey	Nguyen et al. [4]	They present a survey regarding applications of Federated Learning in IoT scenarios. They also present challenges related to FL-IoT services, including the massively distributed challenge.	Our insight encompasses system and data heterogeneity challenges in a broader way.
Federated Learning in Mobile Edge Networks: A Comprehensive Survey	Lim et al. [6]	They focus on device heterogeneity in Federated Learning and study challenges such as communication costs and resource allocation.	In addition to device heterogeneity, we explore other related challenges in wireless, such as network architecture, unreliable network conditions, and energy considerations.
Federated Learning: Challenges, Methods, and Future Directions	Li et al. [7]	They provide an introduction to system and statistical heterogeneity challenges.	We further explain these heterogeneity challenges and relate them to current studies in the field.
A comprehensive empirical study of heterogeneity in Federated Learning	Abdelmoniem et al. [8]	They perform an empirical study on device and behavioral heterogeneity simultaneously.	We aim to provide both theoretical and empirical reviews on all types of heterogeneity.
Federated Learning over wireless networks: Challenges and solutions	Beitollahi et al. [13]	They present a survey on the challenges of implementing Federated Learning over wireless networks, including power-limited devices, wireless network characteristics, limited available bandwidth, and privacy issues.	We complement and expand these challenges in the section on system heterogeneity.
A survey on heterogeneity, taxonomy, security, and privacy preservation in the integration of IoT, wireless sensor networks, and Federated Learning	Mengistu et al. [14]	They provide a tutorial and introduction to Federated Learning. They also study heterogeneity challenges like communication overhead, computational resources, and data imbalance.	We present an extensive and straightforward taxonomy on several heterogeneity challenges in Federated Learning.
A review of Federated Learning methods in heterogeneous scenarios	Pei et al. [9]	They highlight the importance of adopting emerging technologies, such as 6G, as key enablers for more efficient Federated Learning. They perform a theoretical review of the heterogeneity challenges of Federated Learning.	We provide additional validation of Federated Learning in practical scenarios. We also comprehensively examine the behavioral heterogeneity category, which the article overlooks.

Heterogeneity refers to the diverse characteristics of participants in training and sharing a Machine Learning model. Heterogeneity may degrade the global model's performance rather than enhance it. Heterogeneity may also lead to reduced model accuracy, potential model divergence, prolonged training and aggregation times, increased computational load on the server, implementation challenges, and fairness concerns among participants.

Identifying and categorizing heterogeneity factors in a Federated Learning environment is essential. Furthermore, understanding the challenges associated with each heterogeneity factor facilitates the development of solutions and guidelines to support research. The primary contribution of this work is a classification and analysis of heterogeneity challenges in Federated Learning, with a focus on wireless networks—particularly mobile networks, which are expected to drive FL adoption in the future. In addition, we review recent studies on heterogeneity in Federated Learning, along with state-of-the-art solutions and future research directions. To provide a better overview of this classification, Figure 2 summarizes the heterogeneity challenges addressed in this document.

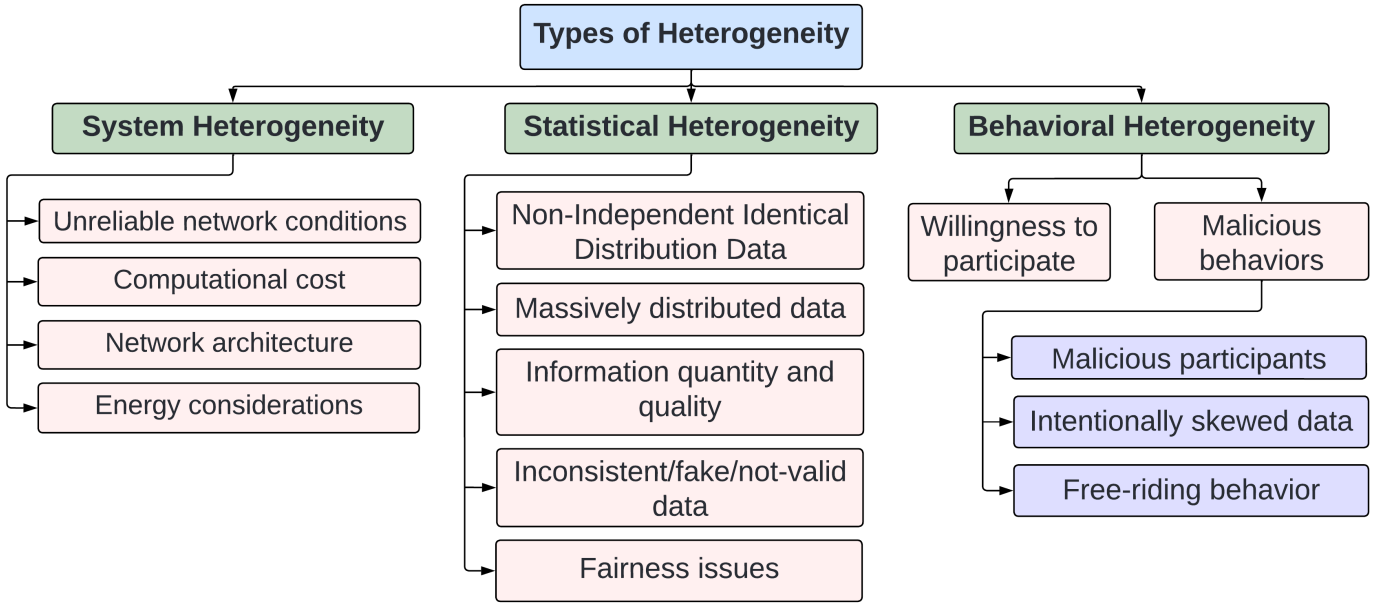


Figure 2. Classification of Federated Learning heterogeneity challenges.

Federated Learning has emerged as a key enabler for intelligent 5G and future 6G networks, allowing for decentralized model training while preserving data privacy. However, its deployment in real-world applications inherently encounters the heterogeneity challenges discussed earlier. In intelligent transportation systems (ITS), for instance, FL is utilized to enable collaborative learning among connected autonomous vehicles. However, system heterogeneity poses a challenge due to the diverse computational capabilities, network latency variations, and sensor-quality discrepancies among vehicles [2]. Similarly, in smart-city applications, FL supports edge-based AI for traffic monitoring, energy grid optimization, and public safety surveillance. However, statistical heterogeneity arises as different urban regions generate highly imbalanced and non-IID datasets, affecting global model performance [3]. In healthcare, FL enables hospitals and mobile health devices to collaboratively train medical AI models without sharing sensitive patient data. Yet, behavioral heterogeneity becomes an obstacle, as some hospitals may not consistently participate due to data privacy concerns or regulatory restrictions [16]. In addition, in 6G-enabled industrial IoT (IIoT) environments, FL is being explored to enhance predictive maintenance and anomaly detection across heterogeneous manufacturing devices. However, the challenge of unreliable network conditions and limited energy availability in IoT nodes complicates

model convergence and fairness [4]. These practical cases highlight the significant impact of heterogeneity in FL applications and motivate the need for adaptive techniques to mitigate these issues in future wireless networks.

The following sections provide a detailed analysis of each heterogeneity category. First, we examine system/architectural/device heterogeneity, which arises from differences in hardware capabilities and network conditions. Some devices may have limited energy availability, as seen in Internet of Things (IoT) scenarios, while others may struggle to connect to the aggregator due to low coverage or network congestion.

Next, we discuss statistical/data heterogeneity, which occurs when FL participants contribute highly varied data distributions, often characterized as non-independent and non-identically distributed (non-IID) data. Non-IID (non-independent and identically distributed) data are a type of dataset in which the data samples are not homogeneous, and the statistical properties of the data vary across different subsets. When training on data that are not homogeneous, the model may struggle to generalize well to new data, leading to poor performance. One participant may also contribute with more and better quality data, which translates to a more useful input to the global training model. Data introduced may be inconsistent, fake, or even counterproductive information. Data may be massively distributed in many devices, which hardens the aggregation process. Finally, data may be highly skewed, leading to fairness issues, where not all participants are equally represented, or results may benefit only a small group.

Non-IID (non-independent and non-identically distributed) data refer to datasets where samples are not homogeneous, and their statistical properties vary across subsets. In the context of machine learning, training on non-IID data can have several impacts:

- Model Performance: Non-IID data can degrade model performance. When data distributions are highly varied, models struggle to generalize, resulting in lower accuracy and reliability.
- Convergence: Training on non-IID data can slow convergence. If data distributions differ significantly across participants, the model may take longer to find an optimal solution, increasing training time.
- Bias: Non-IID data can introduce bias. If certain subsets are overrepresented, the model may become skewed, leading to unfair or inaccurate predictions.
- Resource Utilization: Training on non-IID data often requires higher computational resources. Handling diverse data distributions may necessitate more complex models or additional training epochs.

Therefore, it is crucial to consider data characteristics when selecting a training approach. In some cases, preprocessing techniques or specialized learning strategies may be necessary to mitigate the undesired effects of non-IID data on FL models.

The final category, behavioral heterogeneity, refers to differences in participant behavior and expectations within the Federated Learning process. A common challenge arises when users are unwilling to participate due to the time and resource demands of FL. Those who do engage often expect a minimum reward as an incentive. In addition, some users may engage in malicious behavior. For instance, adversarial participants may introduce fake or manipulated data, leading to data/model poisoning attacks. Others may attempt to exploit the system without meaningful contributions, a phenomenon known as free-riding. Finally, users or groups may be interested in contributing skewed data that may benefit one of their interests in selling products or making low-cost propaganda.

Behavioral heterogeneity in Federated Learning refers to variations in client participation patterns, willingness to share data, and adherence to the training process. Unlike systemic heterogeneity, which arises from physical device and network constraints, behavioral heterogeneity stems from the human or institutional factors influencing client engagement

in FL training. It also differs from statistical heterogeneity, which is primarily driven by differences in local data distributions rather than participation behavior. However, these three dimensions of heterogeneity are not entirely independent and often interact. For example, systemic heterogeneity can indirectly contribute to behavioral heterogeneity—devices with limited power, unstable connections, or insufficient computational resources may exhibit inconsistent participation, resulting in unreliable model contributions. Similarly, behavioral heterogeneity can exacerbate statistical heterogeneity—if some clients frequently drop out or refuse to contribute their data, it can lead to biased model updates or underrepresented data distributions. Addressing behavioral heterogeneity requires strategies that balance client incentives, trust mechanisms, and fairness considerations, which are distinct from the optimization-based techniques used to handle systemic and statistical heterogeneity.

3. System Heterogeneity in Federated Learning

System heterogeneity in Federated Learning arises from the diverse characteristics of participating devices. In wireless communication networks, including mobile networks, FL faces critical challenges due to variations in hardware capabilities, computational resources, and energy constraints among devices. In addition, network-related factors such as coverage, operator-dependent infrastructure, and server load balancing further complicate the training process. Addressing these disparities is crucial to ensuring efficient and fair participation in FL systems. Figure 3 provides a clearer representation of our approach to system heterogeneity challenges.

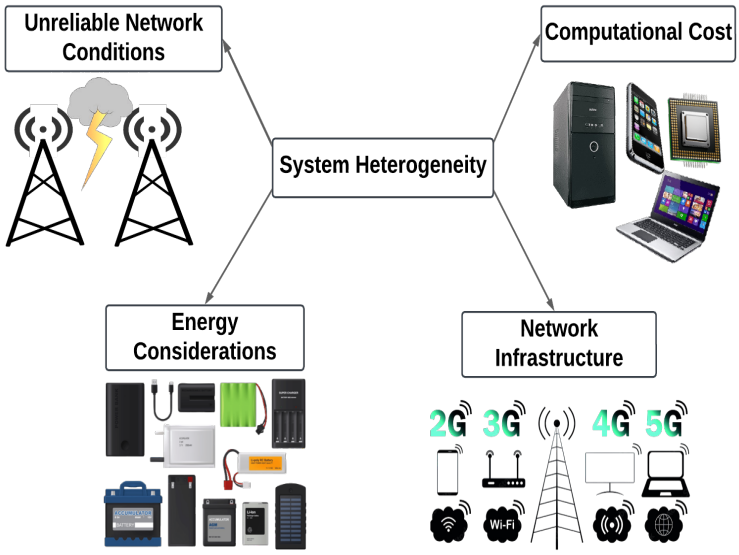


Figure 3. System heterogeneity challenges.

Figure 3 illustrates the primary sources of system heterogeneity in Federated Learning, which impact model convergence and efficiency. Unreliable network conditions are a key challenge, as FL relies on distributed devices communicating over diverse networks. In wireless environments, high latency in 5G mmWave networks or packet loss in mobile edge computing can significantly slow down global model synchronization. Computational cost is another critical factor, as participating devices range from high-end GPUs in data centers to low-power IoT sensors, leading to imbalanced processing power and delayed local model updates. In addition, energy considerations play a significant role, as many FL participants are battery-constrained devices (e.g., smartphones, wearables, drones), where frequent communication rounds drain power, making sustainable FL a challenge. Finally, network infrastructure in FL varies greatly, as it is deployed over heterogeneous

connectivity scenarios, from low-bandwidth 2G rural networks to high-speed 5G and WiFi urban networks. This variability impacts client participation rates and synchronization efficiency, influencing overall FL performance.

Similar to traditional Machine Learning, Federated Learning requires multiple training rounds to refine the model and achieve a target accuracy [6]. However, FL introduces significant communication overhead due to the frequent exchange of models and parameters between users and the central server. Studies [6,7,15] indicate that communication bottlenecks worsen as the number of participating devices increases. In [17], authors found that in a fully connected network with 100 devices and a model size of 10 MB, communication overhead reached 300 MB, significantly straining network resources. Moreover, asymmetric transmission speeds exacerbate these issues; for instance, in the United States, average broadband speeds are 55.0 Mbps for downloads but only 18.9 Mbps for uploads [17]. In addition, network instability and unpredictable user behavior further hinder efficient model aggregation, leading to extended training times. Some of the issues found in the literature regarding system heterogeneity are the following:

3.1. Unreliable Network Conditions

Maintaining stable communication between participating devices and the central server is a major challenge in FL due to unreliable network conditions. Factors such as network congestion, weak signal strength, software failures, and environmental disruptions contribute to connection instability. These issues are particularly pronounced in FL, where devices are geographically dispersed and operate across diverse communication infrastructures. Consequently, unreliable network conditions may lead to delays in model updates, increased dropout rates, and reduced training efficiency. Li. et al. [7] state that some unreliable network conditions, such as slow connections, asymmetry in DL/UL speeds, delays, packet loss, dropouts, or disconnections, could weaken the learning process.

According to Adelmoniem A. et al. [18], in Real FL deployments, we expect differences in the quality of network links and availability. As a result, devices with poor connectivity are less likely to meet the reporting deadline, slowing the whole learning process. The reporting deadline refers to the maximum time limit that participating devices have to complete their local updates and communicate them back to the central server for aggregation. The reporting deadline is an important parameter that can affect the overall efficiency and performance.

Federated Learning systems can operate in either synchronous or asynchronous modes, each with distinct advantages and challenges. In synchronous FL, the central server waits for all participating devices to complete their local updates before aggregating the global model. While this ensures consistency across updates, it can lead to delays caused by slower devices, known as the straggler effect. In contrast, asynchronous FL allows devices to update the global model at their own pace, reducing delays but introducing potential inconsistencies in model updates due to varying client contributions. These trade-offs highlight the importance of selecting the appropriate approach based on the network conditions and device capabilities in a given FL deployment.

In synchronous Federated Learning, it is mandatory that every participant send their local models before the aggregation of the global model is performed. However, synchronous Federated Learning causes straggler issues, which means the aggregation can only be carried out once the slowest devices have sent their information. In this context, some frameworks like ELFISH [19] are proposed to tackle the straggler issues. This framework presents viable solutions against system heterogeneity. They propose a selection of the slowest clients and perform a previous optimization separately from the rest of the clients; this is called soft training for straggler acceleration in synchronous Federated

Learning. This framework speeds up the training process and model accuracy for better collaborative convergence.

3.2. Computational Cost

In centralized ML, higher computational power accelerates model training. However, in FL, participating devices vary significantly in processing power, memory capacity, and storage availability, leading to disparities in local training efficiency. Asynchronous Federated Learning is more suitable for scenarios with heterogeneous devices, as it allows each device to update its local model independently and transmit updates to the central server as they become available. This can be particularly useful in scenarios where some devices have more powerful processors or more available resources than others. However, asynchronous FL also introduces challenges, particularly in managing communication between devices and the central server. Moreover, asynchronous updates can lead to inconsistencies in local models, complicating convergence to a globally optimal solution.

Device heterogeneity significantly impacts model performance. Abdelmoniem et al. [20] conducted an experiment evaluating FL accuracy under varying device capabilities. Their study examined two scenarios: (1) an ideal case, where all devices had uniform computational power and network access, ensuring timely updates; and (2) a heterogeneous case, where devices differed in hardware specifications, processing power, and network speeds. The authors conducted a large-scale simulation, testing approximately 1.5 K configurations across five widely used FL benchmarks, totaling 5 K experiments and 36 K GPU h. For further details on ML techniques, model sizes, client-selection criteria, and reporting deadlines, readers are encouraged to refer to the original study. In summary, the experiment demonstrated that device heterogeneity reduced average accuracy by a factor of 0.88× compared to homogeneous devices.

Other works, such as FLASH [21], have also quantified the effects of device heterogeneity in FL. FLASH is an open-source, heterogeneity-aware FL platform, developed as a fork of the LEAF benchmark, which includes multiple datasets for FL experimentation. FLASH enables the simulation of FL processes incorporating heterogeneity. Each client in FLASH is assigned a specific device type with distinct training speeds, network conditions, and computational capabilities. According to the original FLASH study, device heterogeneity resulted in a 9.2% accuracy reduction and a 2.32× increase in convergence time. Convergence time refers to the duration required for participating clients to reach a consensus on the Machine Learning model. This occurs when model performance stabilizes across all clients or improves sufficiently, with accuracy and loss reaching acceptable levels. The aforementioned experiments were conducted on a high-performance computing cluster running Red Hat Enterprise Linux Server 7.3, consisting of 10 GPU nodes. Each node was equipped with two Intel Xeon E5-2643 V4 processors, 256 GB of main memory, and two NVIDIA Tesla P100 graphics cards. In total, the reported experiments consumed over 7000 GPU h. One key finding of this study is that heterogeneity necessitates significant hyperparameter adjustments to optimize training performance. Hyperparameter drift refers to the issue where hyperparameters, such as learning rates, regularization strengths, and batch sizes, become outdated or misaligned with evolving data distributions during training. This misalignment can degrade performance and slow convergence.

To mitigate hyperparameter drift, the authors propose an adaptive learning rate scheduling technique within the FLASH framework. This approach employs a feedback mechanism to dynamically adjust learning rates based on device performance. Specifically, the loss of each device is compared against the global average loss: if a device's loss exceeds the average, its learning rate is reduced, whereas if a device's loss is below the average, its learning rate is increased. This adaptive strategy helps prevent overfitting and enhances

convergence in the presence of heterogeneity and hyperparameter drift. FLASH, an open-source simulation project, facilitates the evaluation of heterogeneity in Federated Learning by incorporating diverse datasets for experimentation. Additional heterogeneity-aware FL frameworks and algorithms are discussed in Section 6.

Significant efforts have been made to improve FL performance under heterogeneous conditions. Various frameworks have been developed and tested by the research community to address these challenges. For instance, Aergia, introduced by Cox et al. [22], is a novel aggregation algorithm designed for FL and included in the Federation Learning Toolkit (FLTK). FLTK is a heterogeneity-aware simulation framework that enables comparative evaluation against existing FL approaches, such as FedAvg and Task-agnostic Federated Learning (TiFL) [23]. Experimental evaluations using large-scale datasets demonstrate that Aergia outperforms both FedAvg and TiFL, achieving up to 27% improvement in convergence time over FedAvg and up to 53% improvement in accuracy over TiFL, even in scenarios with significant system and statistical heterogeneity. Aergia distinguishes itself by simultaneously addressing both data and resource heterogeneity while optimizing training efficiency.

3.3. Energy Considerations

Energy consumption is a critical concern for small wireless devices, particularly battery-powered IoT systems. IoT devices are designed to minimize energy consumption to extend battery life and reduce the frequency of recharging. Federated Learning can be deployed in smart grids and smart cities [24], where optimizing energy usage is crucial. In smart cities, FL enables the decentralized training of Machine Learning models using data from multiple sources, eliminating the need for centralized data storage. Smart cities rely on a vast network of IoT devices and sensors that continuously generate large volumes of data. FL facilitates the processing of this data by allowing distributed devices to collaboratively train a shared Machine Learning model. A key application of FL in smart cities is traffic management. Traffic flow data can be collected from roadside sensors, traffic cameras, and public transportation systems. These data train Machine Learning models to predict traffic patterns and identify congestion hotspots. The trained model can then be deployed on traffic management systems, such as traffic signals and connected vehicles, to optimize flow.

Some studies evaluate participants' energy constraints before including them in the collaborative learning process. Yang et al. [21] highlight that devices may drop out of FL training due to low battery levels. The study proposes collecting device-specific metrics before initiating an FL round, including battery percentage, screen lock status, and network connectivity. Screen state (locked or unlocked) significantly impacts energy consumption in mobile FL deployments, as active screens drain battery life more rapidly. When the screen is unlocked, the device is in active use and consumes more energy, whereas when the screen is locked, the device is in an idle state and consumes less energy. Scheduling FL tasks while the screen is locked can significantly reduce energy consumption in mobile devices.

Kim et al. [25] propose AutoFL, an energy-efficient FL approach that dynamically adapts to device heterogeneity in computational power and energy constraints. AutoFL dynamically optimizes communication frequency and computation power allocation based on device capabilities, network conditions, and data heterogeneity. Leveraging reinforcement learning, AutoFL determines optimal communication frequency and computational power allocation to balance energy efficiency and model accuracy. Experimental results demonstrate that AutoFL significantly reduces energy consumption while maintaining model accuracy in heterogeneous FL environments.

3.4. Network Architecture

In addition to device heterogeneity, network heterogeneity is also a critical factor in Federated Learning. In mobile communication networks, the FL server operates as an edge computing node. This server may not have direct communication with all participants, and some clients may experience network accessibility issues. In addition, the server may become unavailable or overloaded due to high network demand. Furthermore, FL networks may utilize different radio-access technologies (3G, 4G, 5G), each with distinct characteristics and capabilities. Moreover, mobile networks consist of various cell site types, differing in size and coverage area.

The large number of devices involved in FL model training presents a major communication challenge, particularly for its adoption in 6G networks [1]. Although previous research has aimed to enhance FL communication efficiency, synchronizing communication with local device computations remains a challenge. To enable FL models to operate effectively in highly heterogeneous and large-scale 6G networks, communication-efficient strategies must be developed to minimize gradient exchanges between devices and the server. In FL, exchanged gradients refer to model parameter updates computed locally during training. Each participating device trains the model on its local dataset and computes the gradient of the loss function concerning model parameters. Subsequently, each device transmits its computed gradient (rather than the entire model) to the central server, where gradients from all devices are aggregated to update global model parameters. Finally, two critical factors must be considered to reduce communication overhead: (i) minimizing the number of communication rounds and (ii) reducing the volume of gradients transmitted per round.

Abdelmoniem et al. [18] introduce a Selection Stage, during which the server waits for a minimum number of devices to connect. If the required number of devices does not meet the threshold within the Selection Time Window, the training and aggregation round is aborted. In FL, aborting a training round may degrade model quality and increase communication overhead. Since FL relies on model updates exchanged between clients and the server over communication channels, aborting a training round leads to wasted computational and communication resources. Table 2 provides a summary of key findings related to system heterogeneity in FL.

Table 2. Summary of system heterogeneity challenges in Federated Learning.

Challenge	Reference	Key Findings
Unreliable Network Conditions	Li et al. [7]	Communication is challenging when network conditions are unstable. Slow connections, asymmetry in speeds, delay, and packet loss could weaken the learning process.
	Abdelmoniem et al. [18]	Devices with poor connectivity are less likely to meet reporting deadlines, slowing the learning process.
	Xu et al. [19]	Clustering slow clients and performing a soft training process could alleviate the straggler challenge in synchronous FL, speeding up the global training process and increasing model accuracy.
Computational Cost	Abdelmoniem et al. [20]	Computational power leads to a faster training process and higher accuracy.
	Yang et al. [21]	Device heterogeneity makes it necessary to drift hyperparameters significantly for better training performance.
	Cox et al. [22]	Offloading traffic that is computationally costly from slower to faster devices helps maintain high accuracy while reducing training time under heterogeneous settings.

Table 2. *Cont.*

Challenge	Reference	Key Findings
Energy Considerations	Cheng et al. [24]	Practical Federated Learning applications must be energy-aware and energy-preserving, especially in IoT environments.
	Yang et al. [21]	Low-battery issues may cause dropouts during the training process, weakening the convergence time and accuracy of the model.
	Kim et al. [25]	Energy-efficient Federated Learning must consider the heterogeneity of devices in terms of capability and energy consumption. Dynamically adjusting communication frequency helps to prevent energy exhaustion.
Network Architecture	Liu et al. [1]	Federated Learning is sensitive to the network it is working on. Novel networks (6G) are highly heterogeneous as well. Massive client communication hardens the problem.
	Abdelmoniem et al. [18]	Some techniques may reduce the impact of communication overhead. These are: (i) decrease the total number of communication rounds and (ii) reduce the number of gradients transmitted in each communication round.

4. Statistical Heterogeneity in Federated Learning

Statistical heterogeneity is a significant challenge in Federated Learning and has been widely studied in recent years, attracting considerable research attention. In Machine Learning, data typically reside locally within an entity that possesses both storage and processing capabilities. A centralized model is then trained on these data to develop predictive capabilities for various applications. However, in Federated Learning environments, data are distributed across multiple devices or institutions and does not necessarily follow the assumption of being independent and identically distributed (IID). This deviation from IID conditions complicates the training process and can negatively impact model performance if not properly addressed.

Statistical heterogeneity arises when FL participants contribute non-IID data, meaning their local datasets exhibit different underlying distributions. For example, in a financial services scenario, different banks or financial institutions may have varying customer demographics, transaction patterns, and fraud detection data. One institution may predominantly serve international customers, whereas another may cater to corporate clients, leading to highly diverse and imbalanced data distributions. In a real-world FL setting, each participant may contribute heterogeneous or imbalanced data, where some clients provide larger volumes of data or more influential features for model training. In addition, FL systems must contend with massively distributed data, inconsistencies, and potential invalid data, further complicating aggregation and model convergence. Figure 4 illustrates the key challenges associated with statistical heterogeneity in FL.

In statistical heterogeneity, the primary challenge stems from non-IID data, where the data available on each device participating in a Federation are not representative of the global distribution [26]. This lack of uniformity complicates model aggregation and may lead to biased training outcomes. The non-IID challenge is further examined in the next subsection.

Several key challenges associated with statistical heterogeneity in Federated Learning include:

4.1. Non-Independent Identical Distribution Data (Non-IID Data)

Training models on non-IID data increases complexity, reduces accuracy, and may lead to model divergence. Training complexity, in this context, refers to the challenges

of training ML models due to dataset heterogeneity. Mitigating training complexity is essential for developing efficient ML algorithms that learn from heterogeneous datasets, particularly in Federated Learning.

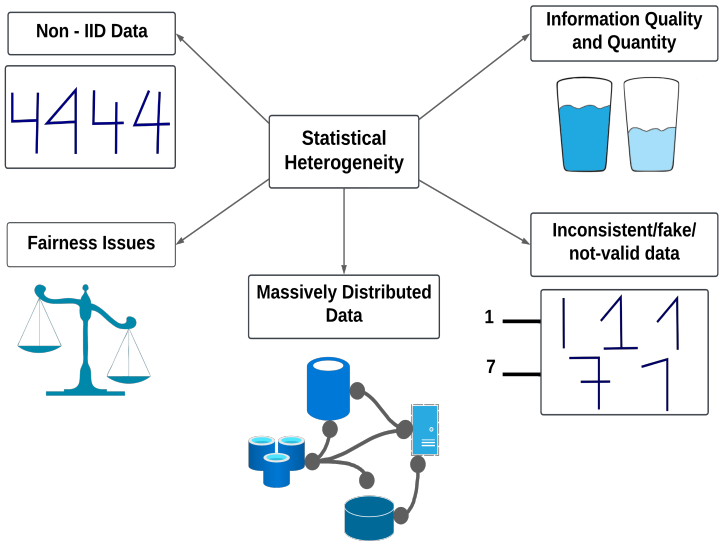


Figure 4. Statistical heterogeneity challenges

The authors in [18] conducted experiments evaluating FL performance under significant device, user, and data heterogeneity, considering variations in data distribution, device capabilities, network quality, and availability. The results indicate that heterogeneity can lead to slow convergence, degraded model performance, fairness issues, and, in many cases, training divergence. In this context, fairness refers to ensuring that FL models trained on decentralized data remain unbiased and do not disproportionately favor specific groups or individuals. The study explicitly employs non-IID data distributions in its experiments. Findings demonstrate that data heterogeneity, alongside system heterogeneity, significantly affects model performance and fairness. These factors collectively exacerbate FL model challenges. The study concludes that heterogeneity affects hyperparameter tuning, including the number of epochs and learning rate, which must be carefully adjusted to maintain accuracy and prevent model divergence. Future research should explore hyperparameter optimization and ML engineering strategies for heterogeneous FL environments.

Non-IID data in FL exacerbates performance degradation and slows model convergence. To address this issue, [27] proposes a hybrid FL approach that integrates global and local model updates, enabling clients with similar data distributions to collaborate while preserving personalization. This is achieved by clustering clients based on the similarity of their model updates, which reflects their data distributions. Clustering statistically similar clients enhances model convergence and accuracy by facilitating more effective aggregation of updates. Although this technique stabilizes training, it may increase computational overhead, particularly in large-scale FL deployments. The study in [27] demonstrates that such clustering-based approaches outperform conventional FL methods, achieving improved accuracy and faster convergence on benchmark datasets. These findings highlight the importance of adaptive FL strategies that account for data heterogeneity while balancing computational efficiency and personalization in decentralized learning environments.

Data heterogeneity in Federated Learning, specifically having non-IID data distributions, can lead to poor performance and slow convergence. To address this issue, the approach proposed in [27] combines global and local model updates to enable clients with similar data distributions to learn collaboratively while maintaining personalized models.

This is achieved by clustering clients based on the similarity of their model updates, which reflects their data distributions.

Clustering clients in this way can improve model convergence and accuracy by ensuring that updates are aggregated more effectively within statistically similar groups. However, while this technique enhances training stability, it may introduce additional computational overhead, particularly in large-scale FL deployments. The study in [27] demonstrates that such clustering-based approaches outperform conventional FL methods, achieving improved accuracy and faster convergence on benchmark datasets.

The proposed approach allows for personalized learning while leveraging information from similar clients, resulting in improved performance and faster convergence in Federated Learning. Experimental results validate the proposed approach's effectiveness in mitigating non-IID data heterogeneity. Tested on CIFAR-10, the approach achieves an 85.76% average accuracy, surpassing state-of-the-art methods such as FedAvg, FedMD, and FML.

These findings highlight the importance of adaptive FL strategies that account for data heterogeneity while balancing computational efficiency and personalization in decentralized learning environments.

4.1.1. Mathematical Analysis of Non-IID Impact on Federated Learning Convergence

In Federated Learning (FL), the global model is updated at each communication round t using the Federated Averaging (FedAvg) algorithm:

$$w^{t+1} = w^t - \eta \sum_{k=1}^{K} \frac{n_k}{N} \nabla F_k(w^t), \tag{1}$$

where:

- w^t is the global model at round t;
- η is the learning rate;
- $\nabla F_k(w^t)$ is the local gradient computed by client k;
- n_k is the number of samples at client k;
- $N = \sum_{k=1}^{K} n_k$ is the total number of samples across all clients.

In IID settings, local gradients $\nabla F_k(w^t)$ are similar, leading to faster convergence. However, in non-IID settings, local updates vary significantly due to differences in data distributions. This increases the variance of local gradients. The deviation between local gradients and the global gradient can be expressed as:

$$\mathbb{E}\left[\|\nabla F_k(w^t) - \nabla F(w^t)\|^2\right] \leq \sigma^2 + \zeta \|w^t - w^*\|^2, \tag{2}$$

where

- σ^2 represents the gradient variance due to stochastic noise;
- ζ quantifies the heterogeneity of data distributions;
- w^* is the optimal global model.

To provide a more intuitive understanding, the parameter ζ, referred to as the heterogeneity factor, quantifies the degree of divergence between the local data distributions of participating clients. A higher ζ indicates greater variability in the data distributions across clients, which can lead to more significant differences in the local model updates. This variability makes it harder for the global model to converge, as the updates from different clients may conflict with one another. For example, in a healthcare Federated Learning scenario, if one hospital's dataset primarily consists of pediatric cases while another focuses on geriatric cases, the heterogeneity factor ζ would be high due to the stark differences in data distributions.

On the other hand, σ^2, the variance due to stochastic noise, captures the randomness introduced during the local training process. This randomness can arise from factors such as the use of stochastic gradient descent (SGD) or the inherent noise in the data itself [28]. A higher σ^2 implies that the local updates are noisier, which can slow down convergence by introducing fluctuations in the global model updates. For instance, in a Federated Learning system where clients have small datasets or noisy sensor readings, σ^2 would be elevated, reflecting the increased uncertainty in the local updates.

Together, ζ and σ^2 provide a comprehensive view of the challenges posed by non-IID data in Federated Learning. While ζ highlights the structural differences in data distributions, σ^2 emphasizes the impact of randomness and noise during training. Addressing these factors requires strategies such as clustering clients with similar data distributions to reduce ζ or employing noise-reduction techniques to minimize σ^2.

The convergence rate of FedAvg in non-IID settings, as shown by Li et al. [28], is bounded by:

$$\mathbb{E}\left[F(w^t) - F(w^*)\right] \leq \mathcal{O}\left(\frac{1}{\sqrt{TK}} + \frac{\zeta}{T}\right), \tag{3}$$

where T is the total number of training rounds. The term $\frac{\zeta}{T}$ shows that data heterogeneity (ζ) slows convergence, requiring more communication rounds to achieve a stable global model.

Key observations:

- As data heterogeneity (ζ) increases, convergence requires more communication rounds.
- Higher local updates before aggregation ($E = 1$ vs. $E > 1$ in FedAvg) exacerbate gradient divergence, leading to slower global model improvement.
- Strategies such as adaptive weighting, personalized FL, and clustered FL mitigate the effects of non-IID data by adjusting update contributions based on client similarity.

This mathematical framework formally quantifies the degradation in convergence speed and serves as a theoretical basis for optimizing FL performance in heterogeneous environments.

4.1.2. Impact of Heterogeneity on Model Convergence

Heterogeneity in Federated Learning introduces significant challenges in model convergence, particularly when dealing with non-IID data distributions, device variability, and network constraints. Several studies have demonstrated that heterogeneity can lead to increased training time, lower accuracy, and in some cases, complete model divergence.

One of the key challenges is statistical heterogeneity, where client datasets follow vastly different distributions. As shown in [8], FL models trained on non-IID data tend to converge much slower compared to IID data scenarios. Without proper mitigation techniques, such as clustering-based FL or personalized learning strategies, convergence may stagnate or lead to biased models. The experimental results in [9] indicate that the use of clustering in FL can improve convergence rates by up to 30% in non-IID environments.

Another critical factor is device heterogeneity. As FL operates across devices with varying computational power, the presence of slower devices (stragglers) can significantly delay aggregation steps in synchronous FL. The work in [28] proposes an asynchronous FL approach where clients update independently, reducing convergence delays. However, this method introduces new challenges, such as inconsistencies in global model updates.

Empirical evaluations have quantified the impact of these factors. For instance, [21] reports that device heterogeneity causes an accuracy drop of approximately 12% relative to homogeneous settings, while network variability increases convergence time by a factor of 2.32×. These findings emphasize the need for adaptive FL strategies that dynamically adjust learning rates, select participating clients intelligently, and employ robust aggregation techniques.

To address these challenges, future research should explore hybrid FL models that integrate meta-learning, reinforcement learning, and fairness-aware optimization to improve convergence in heterogeneous environments.

4.2. Information Quantity and Quality

Also known as unbalanced data, this issue arises when FL participants contribute varying amounts of data, or when some choose not to share their data (see Section 5). In addition, the data provided by some participants may have a lower impact on model improvement than that of others. For instance, in an image recognition task, if one device provides a larger number of images that align with the model's learned patterns, its contribution to training will be more significant than that of other devices. However, this does not necessarily ensure high data quality, as the images may be low resolution, noisy, or misclassified. Data imbalance also increases training complexity, as models must account for disproportionate contributions and potential biases in data representation.

Ensuring sufficient and balanced information flow is critical to the success of Federated Learning, as data quantity and quality vary significantly across clients. The work presented in [7] outlines methods to mitigate these challenges by optimizing data usage, communication efficiency, and model training in heterogeneous FL environments. The proposed solutions directly address the problem of unequal client contributions and imbalanced data representation, both of which impact the quality and quantity of information available for model updates.

(i) *Local Updating:* This method leverages mini-batch optimization techniques to provide flexibility and robustness against communication and computation constraints. Mini-batch optimization enables the more effective utilization of client data by allowing training to occur on small, diverse subsets rather than requiring full dataset transmission. This strategy improves data representativeness in global model updates while ensuring fair contributions from clients with varying dataset sizes. In addition, mini-batch optimization enhances privacy, as only partial data are used in each round rather than exposing the entire dataset.

(ii) *Compression Schemes:* Model compression techniques, including sparsification, subsampling, and quantization, reduce the amount of information exchanged during training rounds while maintaining model accuracy. These methods directly impact information quantity by minimizing communication overhead while ensuring that essential updates are retained. Sparsification reduces redundant parameters in weight matrices, subsampling selects the most informative data points from client datasets, and quantization lowers precision requirements for model updates. Together, these techniques enable FL systems to efficiently aggregate high-quality information from distributed clients without excessive bandwidth consumption.

(iii) *Decentralized Training:* In highly heterogeneous environments, communication bottlenecks can limit the amount of information available for model training. Decentralized training alleviates this issue by enabling peer-to-peer collaboration among capable clients, reducing dependence on a central server. This method enhances both information quality, by incorporating diverse perspectives from multiple clients, and information quantity, by expanding training opportunities beyond a centralized topology. Studies show that decentralized FL can improve model convergence rates, particularly in settings with limited connectivity or high latency.

These techniques collectively ensure that FL systems can effectively balance information distribution across clients, mitigating data imbalance and unreliable communication channels. By improving the availability and representativeness of training data while

optimizing communication efficiency, they directly contribute to overcoming the challenge of information quantity and quality in FL.

A significant challenge in Federated Learning is managing noisy clients, which can degrade model quality and accuracy. In FL, a noisy client is a device that produces low-quality training data or introduces higher error rates than other clients. This issue may arise due to hardware limitations, data corruption during transmission, or unreliable sensor readings. The article presented in [29] introduces a novel FL approach to address challenges arising from noisy and heterogeneous clients, particularly those related to data and client heterogeneity. The study proposes a hybrid optimization framework, integrating global and local optimization techniques to enhance model accuracy in noisy and heterogeneous FL environments. Local optimization involves updating model parameters based on each client's local dataset. Each client executes local optimization independently, without direct synchronization with others. This approach preserves data privacy in FL, as client data remain on local devices and are never shared with the central server or other clients. Following local optimization, each client transmits its updated model parameters to the central server, where global optimization is applied. Global optimization employs techniques such as Federated Averaging (FedAvg), where model parameters are aggregated across all clients. Alternative methods include Federated stochastic gradient descent (FSGD), which optimizes models in a distributed manner. The proposed method outperforms existing approaches in both accuracy and robustness, underscoring the need for FL techniques that effectively handle real-world data and client heterogeneity. These advancements enhance model performance and scalability, making FL more resilient to data noise and client heterogeneity.

4.3. Fairness Issues

In Federated Learning, the diverse nature of data means that each participant may share biased information based on personal preferences, stored data, or selective disclosure. This issue, known as algorithmic bias, can lead to fairness concerns in model training. Fairness in FL refers to a statistical imbalance, where certain groups may be underrepresented or disadvantaged, leading to ethical concerns related to gender, ethnicity, age, and other demographic factors. In FL, disparities in user participation and data contribution may result in unintentional fairness biases, affecting model performance and equity.

As discussed in Section 3, FL participants exhibit varying computational power and network connectivity. A resource-aware fairness algorithm enables model adaptation to accommodate computational and connectivity disparities without disadvantaging low-power or slow-network devices. Thus, the resource-aware algorithm promotes fairness in both hardware capabilities and data distributions.

An emerging approach to mitigating fairness issues is personalized Federated Learning (PFL). PFL is a recently proposed FL architecture that accounts for data heterogeneity and offers personalized solutions for participants experiencing data scarcity or limited communication capabilities. Key contributions of PFL include: (i) the development of a taxonomy for personalized Federated Learning, addressing existing FL limitations; (ii) the introduction of Global Model Personalization and Learning Personalized Models Strategies, which enhance fairness and adaptability; and (iii) a set of personalization techniques, including data augmentation, transfer learning, multi-task learning, and clustering, designed to tailor FL models to diverse participant conditions. Researchers note that most FL implementations operate within the same feature space (i.e., same application) but involve varied data samples across participants. Detailed descriptions of PFL approaches—including data-based, model-based, architecture-based, and similarity-based methods—can be found in [30]. A notable data-based PFL technique is data augmentation, which generates syn-

thetic training data from existing samples to improve FL model performance. This method is particularly beneficial in scenarios where training data are limited or imbalanced across FL clients. Another approach is clustering, a similarity-based technique that groups FL clients with comparable data distributions to enhance training efficiency. Clustering aims to form device groups with similar data characteristics, improving both efficiency and effectiveness in FL training.

4.4. Inconsistent/Fake/Invalid Data

In Federated Learning, it is essential to assume that only a subset of collected data will positively contribute to model improvement. Therefore, data must be sanitized and preprocessed before inclusion in the FL training process. This issue encompasses inconsistent, fake, and invalid data, which can further lead to model-poisoning or data-poisoning attacks.

Given the diversity of data in real-world FL implementations, not all collected data will positively impact model training. According to [31], user data vary in scope and application, forming the basis for Horizontal and Vertical Federated Learning. In Horizontal Federated Learning (HFL), multiple parties or devices possess similar types of data, which are distributed across different entities. Conversely, Vertical Federated Learning (VFL) involves multiple parties with different types of data that are related but non-overlapping. For example, HFL aggregates a common feature from multiple clients, whereas VFL collects diverse features from a single client. In practice, data can be simultaneously horizontally and vertically distributed, further increasing FL training complexity. Integrating HFL and VFL further escalates the complexity of the FL process. This complexity arises from the need to synchronize different data structures and training methodologies, ensuring global model accuracy and effectiveness. A common solution involves a hybrid FL approach, where separate models are trained for different data types before aggregating into a joint model. This process demands careful coordination and communication among FL participants.

Even when HFL and VFL distribution challenges are resolved, incorrect data labeling can still cause misclassification issues. According to [6] reports injecting only 50 mislabeled samples into a dataset of 600,000 training instances can increase the misclassification rate to 90%. This experiment simulated a targeted backdoor attack on deep learning models via data poisoning [32]. The researchers selected 50 images from the MNIST dataset and manually altered their labels to create mislabeled (dirty) samples. Each selected image was reassigned to a specific target label, such as modifying a “7” to a “1”. These mislabeled images were injected into the original training dataset, which contained 600,000 correctly labeled samples.

4.5. Massively Distributed Data

The challenges discussed earlier persist in Federated Learning but intensify as the number of participants increases. Future mobile networks and IoT ecosystems are expected to support a massive number of simultaneous connections. Huang et al. [26] report that certain FL deployments may scale up to 10^{10} clients.

FL is anticipated to integrate with massive wireless networks, including 6G and IoT infrastructures. In IoT environments, FL will enable new applications and enhance existing ones, including data sharing, data offloading, caching, attack detection, localization, and mobile crowdsensing [4]. FL in IoT presents unique challenges, including efficient device selection and management, low-latency communication, effective data partitioning, and high reliability for mission-critical applications.

A promising technique, model distillation, has demonstrated effectiveness in handling large-scale datasets. In Machine Learning (ML), model distillation involves training a complex model on a large dataset and transferring its knowledge to a simpler model.

This process entails generating training data using the complex model and subsequently training the simplified model on these data. The objective is to enable the simplified model to approximate the performance of the complex model on test datasets. In FL, model distillation facilitates the creation of a global model by distilling knowledge from local models trained on heterogeneous devices. This approach enhances accuracy and convergence speed, particularly in non-IID data scenarios. Throughout iterative FL rounds, each device updates its local model, using the global model as a reference. Each local model is trained on its own dataset, with weights initialized near those of the global model to maintain consistency. This is typically carried out by downloading the current weights of the global model and using them to initialize the local model.

Authors in [33] introduce FedMD, a novel FL approach leveraging model distillation to address challenges in massively distributed data. FedMD employs a centralized model to distill knowledge from heterogeneous local models, generating an optimized global model. FedMD enhances global model accuracy by leveraging data and model diversity across heterogeneous devices. The experimental results demonstrate that FedMD surpasses state-of-the-art FL approaches in accuracy and convergence speed, particularly in non-IID and massively distributed datasets such as FEMNIST and CIFAR-100. Researchers observed an average 20% improvement in final test accuracy compared to non-collaborative approaches. Table 3 summarizes our findings for the statistical heterogeneity category.

Table 3. Summary of statistical heterogeneity challenges in Federated Learning.

Challenge	Reference	Key Findings
Non-IID Data	Abdelmoniem et al. [18]	Using Non-IID data distributions has an impact on the model's performance and fairness. Combined with system heterogeneity, it has an even greater impact.
	Huang et al. [27]	Data heterogeneity can lead to poor performance and slow convergence. Clustering clients may alleviate this challenge.
Information Quantity and Quality	Li et al. [7]	Information quantity and quality, or unbalanced data, lead to increased training complexity. Some ways to counteract are: (i) local updating, (ii) compression schemes, and (iii) decentralized training.
	Fang et al. [29]	Noisy clients can affect the quality and accuracy of the trained model in FL. Local optimization may help improve the global training process.
Fairness Issues	Tan et al. [30]	Personalized Federated Learning approaches help mitigate the effects of fairness. Techniques include: data augmentation, transfer learning, multi-task learning, clustering, and others.
Inconsistent/Fake/ Invalid Data	Lim et al. [6]	Incorrectly labeled data may lead to misclassification problems. With 50 dirty-label samples in a dataset of 600,000, the misclassification rate may rise to 90%.
	Chen et al. [32]	Inconsistent data may lead to model-poisoning issues.
Massively Distributed Data	Huang et al. [27]	Federated Learning challenges escalate with the number of participants. Some Federated Learning deployments may reach 10^{10} clients.
	Li et al. [33]	Model distillation, transfer learning, and other techniques may be used to overcome the massively distributed data challenge.

5. Behavioral Heterogeneity in Federated Learning

Finally, Federated Learning also faces behavioral heterogeneity, which arises from participant behaviors during the training and model delivery processes. A critical challenge is that not all devices willingly participate in global model training due to high communication costs, computational overhead, and local energy consumption. As a result, some participants may engage in FL only if the incentives outweigh the associated costs. In addition, some users may attempt to exploit the system for unfair gains or intentionally degrade model performance through malicious behavior. The following section categorizes behavioral heterogeneity types and outlines potential mitigation strategies. Figure 5 illustrates the proposed framework for addressing behavioral heterogeneity challenges.

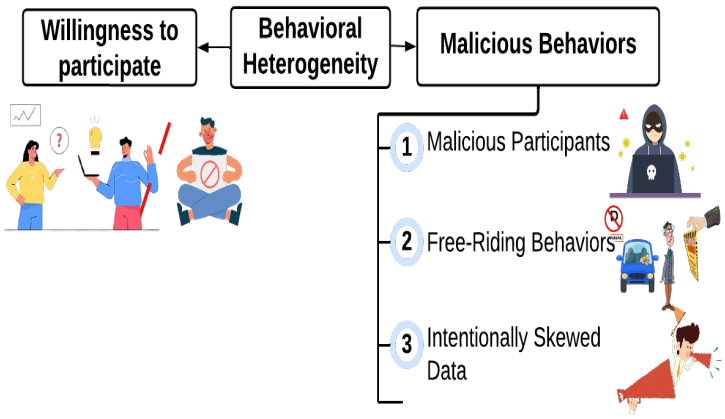


Figure 5. Behavioral heterogeneity challenges.

5.1. Willingness to Participate

Due to the diverse conditions of Federated Learning participants, only a subset may be willing to engage in the collaborative training process. Participants exhibit inherent asymmetry in device capabilities and data ownership. A user is considered a valuable contributor if they possess high processing power, strong network connectivity, sufficient battery levels, minimal mobility, and high-quality data. Conversely, a less optimal contributor may exhibit lower computational power, weaker connectivity, limited battery life, higher mobility, or lower data quality. Participants contributing greater computational resources and higher-quality data may expect proportional and fair compensation for their efforts. In some cases, participants may opt out of FL due to a lack of interest or insufficient incentives. To address this, incentive mechanisms are implemented to motivate broader participation in FL.

To mitigate this issue, ref. [10] introduces a reputation-based methodology for incentivizing FL participation. Participants making greater contributions naturally expect proportional and fair rewards for their efforts. Upon a successful contribution, the system assigns a reputation score to distinguish high-performing participants and encourage further engagement. The reputation score determines the weight of each participant's contribution in the current model aggregation round. Devices with higher reputation scores receive greater aggregation weights, reflecting their higher contribution significance compared to lower-reputation devices. This study presents advanced reputation-based selection and aggregation techniques in FL, introducing the reputation-based aggregation approach outlined earlier. The reputation-based aggregation method mitigates the impact of unreliable or non-participating devices in FL. By prioritizing contributions from reliable and high-quality devices, this approach ensures that the final model is built on trustworthy data and knowledge.

5.2. Malicious Behaviors

In Federated Learning, participants do not share raw private data but instead transmit local model parameters to the aggregator to facilitate the training process. Although FL enhances privacy and security, it introduces new vulnerabilities that may be exploited by malicious participants. Beyond device and data heterogeneity, FL also faces behavioral heterogeneity, as participants may exhibit diverse and unpredictable behaviors. In this context, some participants may intentionally disrupt the FL process or attempt to compromise the privacy of other users.

Malicious participants can engage in various types of adversarial attacks against the FL system. For instance, an attacker may inject false or corrupted data into the training process, deliberately degrading model accuracy. An adversary may also attempt to steal or manipulate sensitive participant data to gain an unfair advantage. In addition, some participants may seek to exploit the trained model without contributing data or computational resources, a strategy known as free-riding. The following subsections detail these challenges and explore potential mitigation strategies.

5.2.1. Malicious Participants

Malicious agents may inject fake or invalid data into the learning process, potentially degrading model performance rather than enhancing it. This manipulation can lead to data and model poisoning attacks.

In Federated Learning, each participant trains a local model using its own dataset and transmits the updated parameters to the central aggregator. Since the aggregator does not verify local model updates, distinguishing between authentic and manipulated data is challenging. This vulnerability enables data-poisoning attacks [6], where untrustworthy participants degrade model performance or hinder convergence.

A more severe threat arises when malicious clients replace their entire local model before submitting updates to the central aggregator, rather than manipulating individual data samples. Such attacks are referred to as model-poisoning attacks. Authors in [6] highlight that even small-scale model poisoning can significantly degrade FL model performance.

Privacy and security remain critical research areas in FL systems. For an in-depth discussion, ref. [34] provides an overview of FL applications in cybersecurity, highlighting key privacy and security challenges. The study outlines communication, privacy, and security challenges in FL and explores various mitigation techniques. One proposed approach is a reputation-based system that monitors participant behavior within the FL network. This system assigns a reputation score to each participant based on past behavior, including cooperation levels and potential manipulations. The reputation score influences the trust level assigned to participant contributions in FL model updates.

Another proposed approach involves cryptographic techniques to enhance data privacy and security in FL. For instance, homomorphic encryption enables computations on encrypted data without requiring decryption. This method prevents malicious actors from accessing or modifying sensitive training data. In addition, secure aggregation techniques enhance model robustness, ensuring accuracy even in the presence of malicious or incorrect updates. For example, Secure Multi-Party Computation (MPC) enables the secure aggregation of updates while preserving participant anonymity. The study explores future FL applications in cybersecurity, such as threat detection, intrusion detection, and malware analysis. Overall, the study underscores the need for efficient and secure FL techniques to address cybersecurity challenges in big data and IoT ecosystems.

5.2.2. Free-Riding Behavior

In Federated Learning, locally generated data are neither shared nor verified by other participants, allowing adversaries to generate meaningless data with minimal effort while benefiting from the genuine contributions of others. This phenomenon, known as a free-riding attack [6], can be mitigated using blockchain-based approaches. Lim et al. [6] introduce BlockFL, a blockchain-based FL architecture where local learning updates are exchanged and verified through blockchain technology. Each participant trains a local model and submits its learning updates to an associated miner in the blockchain network. In return, participants receive rewards proportional to their data contributions. This framework not only prevents free-riding but also incentivizes active participation in the FL process.

5.2.3. Intentionally Skewed Data

In supervised machine learning, increasing the volume of training data provided to the aggregator enhances the model's ability to generate predictions aligned with the input distribution. Malicious participants may attempt to manipulate model predictions to align with their own interests. For instance, in a recommendation system, FL participants may be influenced to prefer specific shopping locations if biased data has been intentionally injected into the training process.

The article in [35] explores the skewed data problem in FL, which arises when data distributions among clients are non-uniform, leading to model bias and degraded performance. The study proposes multiple mitigation strategies, including data augmentation, weighted aggregation, and model personalization. Data augmentation techniques generate synthetic training data to balance underrepresented client distributions. Weighted aggregation involves giving more importance to the data from underrepresented clients during model aggregation. Model personalization tailors training by developing separate models for distinct client groups, adapting to their unique data distributions. In model personalization, each client trains a local model using its own training algorithms and hyperparameters. These locally trained models are subsequently aggregated to construct the global model. Model personalization enhances FL accuracy and performance, particularly in environments with high data heterogeneity. Ultimately, this approach mitigates data skewness and the underrepresented client issue by encouraging active participation. The study concludes that combining multiple strategies effectively mitigates data skew challenges in FL. To evaluate the effectiveness of these methods, the authors conducted experiments using tabular data, specifically credit card loan risk analysis. A European credit card risk dataset was utilized to develop a binary classifier for loan risk prediction. The dataset comprises 30 anonymized features representing credit risk indicators. The experimental results indicate that model accuracy reached 91% with the Bounds-Aware Method, 94% with Data Exchange, and 97% with the Naïve Fusion approach. Table 4 summarizes key findings related to behavioral heterogeneity in FL.

Table 4. Summary of behavioral heterogeneity challenges in Federated Learning.

Challenge	Reference	Key Findings
Willingness to Participate	Wang et al. [10]	Participants may be reluctant to participate, as it costs time and computational resources. They introduce a reputation-based methodology in order to incentivize good contributors to the learning process. The system assigns scores to those participants who have made significant contributions.

Table 4. *Cont.*

Challenge	Reference	Key Findings
Malicious Behaviors	Lim et al. [6]	**Malicious Participants:** (i) Due to Federated Learning principles, data cannot be checked in the aggregator, and it is hard to tell whether it is authentic or valid. Thus, malicious participants may introduce deliberately inconsistent data, leading to data-poisoning attacks. (ii) Malicious clients would not use fake data but replace the whole local model before it is sent to the central aggregator, leading to model-poisoning attacks.
	Alazab et al. [34]	**Malicious Participants:** Federated Learning faces many challenges related to privacy and security. Some of the techniques provided to face cybersecurity challenges include: Reputation Systems, Homomorphic Encryption, and Multi-Party Computation (MPC), among others.
	Lim et al. [6]	**Free-Riding Behavior:** Some users might want to benefit from a collaboratively trained model while making no to low contribution. The authors introduce BlockFL, a blockchain-based FL architecture to address this problem.
	Verma et al. [35]	**Intentionally Skewed Data:** Skew data problem occurs when the data distribution across different clients is non-uniform, leading to poor performance and biased models. Data augmentation, weighted aggregation, and model personalization may help generate additional training data for underrepresented clients.

5.3. Incentive Mechanisms and Reputation-Based Aggregation in Federated Learning

Behavioral heterogeneity poses a significant challenge in Federated Learning, as participant engagement and model reliability depend on a variety of human-driven factors. To address this, incentive mechanisms and reputation-based aggregation have been proposed as strategies to improve client participation, fairness, and model robustness.

Incentive Mechanisms: Since FL is a distributed process requiring resource-intensive computations, many clients may be unwilling to participate unless appropriately incentivized. To mitigate this issue, researchers have proposed several reward-based systems that encourage devices to contribute meaningfully to the training process. For instance, incentive mechanisms leveraging game theory and auction-based frameworks dynamically allocate rewards to clients based on their computational contributions and data quality [36]. In addition, monetary incentives, such as blockchain-based reward systems, have been explored to ensure fair compensation for resource usage [37]. Recent studies indicate that incorporating client utility functions in incentive models improves participation rates and overall FL performance [37].

Reputation-Based Aggregation: Beyond direct financial incentives, reputation-based aggregation mechanisms have been introduced to ensure that model updates from reliable clients are given higher weight during aggregation. These approaches evaluate client reliability using historical performance, contribution quality, and behavioral consistency over multiple training rounds [11]. Reputation scores are dynamically adjusted based on a client's contribution fidelity, and devices with higher scores receive greater influence in the global model update, reducing the impact of unreliable or malicious participants [10].

By integrating incentive mechanisms and reputation-based aggregation, FL can improve participant engagement, mitigate free-riding behavior, and enhance overall model robustness. Future research may explore hybrid approaches combining financial incentives, reputation tracking, and trust-based computing models to further optimize FL deployment in heterogeneous environments.

While behavioral heterogeneity presents unique challenges in Federated Learning, several mitigation strategies have been proposed to address these issues. For instance, reputation-based aggregation mechanisms can prioritize contributions from reliable clients,

reducing the impact of malicious or free-riding participants. Incentive mechanisms, such as blockchain-based rewards or reputation scoring, encourage active participation and fair contributions. In addition, trust-aware aggregation techniques can detect and mitigate unreliable updates, ensuring the integrity of the global model. These strategies complement existing solutions for system and statistical heterogeneity, creating a more robust FL framework capable of addressing diverse real-world challenges. Table 5 presents a concise summary of mitigation strategies for the three primary types of heterogeneity in Federated Learning: system, statistical, and behavioral. The table highlights the challenges associated with each type and the corresponding strategies to address them.

Table 5. Summary of mitigation strategies for heterogeneity in Federated Learning.

Heterogeneity Type	Challenges	Mitigation Strategies
System Heterogeneity	Device variability, straggler effects, energy constraints, network instability	Adaptive client selection, asynchronous FL, resource-aware scheduling, gradient compression, decentralized aggregation.
Statistical Heterogeneity	Non-IID data, class imbalance, data inconsistency	Clustering-based FL, personalized FL, adaptive weighting, hybrid global–local models, data augmentation.
Behavioral Heterogeneity	Free-riding, malicious updates, inconsistent participation	Reputation-based aggregation, incentive mechanisms, trust-aware aggregation, blockchain-based verification.

6. Federated Learning Algorithms and Frameworks

Various frameworks, algorithms, and methodologies have been developed to address the heterogeneity challenges in Federated Learning. While some solutions focus on providing general-purpose FL frameworks for experimentation, others incorporate specific mechanisms to mitigate heterogeneity issues such as statistical imbalance, system constraints, or behavioral inconsistencies.

To provide a structured analysis, we have compiled Table 6, which presents a summary of open-source FL tools and heterogeneity-aware approaches. This table outlines the key functionalities, advantages, and limitations of each method, allowing researchers and practitioners to assess their suitability for different FL scenarios. This comparison provides an overview of widely used FL methodologies and their respective strengths and trade-offs. This allows for a comprehensive understanding of available FL solutions while acknowledging that many frameworks address multiple heterogeneity challenges simultaneously.

6.1. Open-Source Frameworks for Federated Learning

Several open-source frameworks are available for testing and implementing Federated Learning. The following subsections describe key FL frameworks and their capabilities.

6.1.1. TensorFlow Federated (TFF)

TensorFlow Federated (TFF) [38] is a framework built on TensorFlow to support decentralized Machine Learning and distributed data processing. Both TensorFlow and TFF are developed and maintained by Google. TFF comprises two main components: (i) the FL Layer, which provides a high-level interface for integrating FL into existing TensorFlow models without manual FL algorithm implementation, and (ii) the Federated Core (FC), which combines TensorFlow with communication protocols, allowing for experimentation with existing and novel FL algorithms.

6.1.2. Flower

Flower is a scalable and platform-independent FL framework designed for unified evaluation and experimentation [39]. It is supported by a growing community of researchers, engineers, and practitioners. Despite being a recent development, Flower is adopted by major organizations, including Samsung, Nokia, and the University of Cambridge. Flower supports a wide range of ML applications, scales to large deployments, and is compatible with TensorFlow and PyTorch.

6.1.3. FATE

Federated AI Technology Enabler (FATE) is an industrial-grade, open-source framework designed for large-scale collaborative model training in distributed environments [40]. FATE is Python-based and integrates TensorFlow for deep learning models. FATE incorporates security and privacy mechanisms, including Multi-Party Computation (MPC) protocols and homomorphic encryption, but does not support Differential Privacy (DP).

6.1.4. Paddle

Parallel Distributed Deep Learning (Paddle) is an efficient and scalable deep learning framework developed by Baidu. Paddle serves as an industrial-grade AI platform, offering deep learning frameworks, model libraries, end-to-end development kits, and deployment tools. Paddle has been used in applications that include computer vision, natural language processing, and recommendation systems. Paddle includes the FedAvg, SecAgg, and differentially private stochastic gradient descent (DPSGD) strategies [41]. Some other learning techniques for FL included, these are: multi-task learning and transfer Learning.

6.1.5. OpenFL

Open Federated Learning (OpenFL) is a Python-based FL framework [42]. OpenFL is another proof that Federated Learning is a technology of interest for big organizations worldwide. Developed by Intel, OpenFL is intended as a community-driven project, encouraging contributions from researchers and developers.

6.1.6. PySyft

PySyft is a Python framework that enables secure numerical analysis on remotely hosted data, extending to FL applications [43]. Syft is an open-source stack that targets secure and private Data Science in Python. In order to simulate Federated Learning, participants are assigned the role of Virtual Workers. After, the data can be split and distributed to the Virtual Workers as a simulation of a practical FL setting [6]. PySyft remains an attractive tool for those with basic knowledge or a preference to work with PyTorch.

6.1.7. LEAF

LEAF is a benchmarking framework providing datasets for FL-related applications, including multi-task learning, meta-learning, and on-device learning [44]. LEAF offers a dataset with a non-IID distribution named Federated Extended MNIST (FEMNIST), making it more suitable to run Federated Learning simulations like a real-world implementation. As for this dataset, LEAF becomes an exciting option to test data that are statistically heterogeneous.

6.2. Heterogeneity-Aware Protocols, Algorithms, and Frameworks for Federated Learning

This subsection summarizes key algorithms, protocols, and frameworks in Federated Learning, with a particular focus on those designed to address heterogeneity challenges.

6.2.1. FLASH

FLASH (https://github.com/PKU-Chengxu/FLASH (accessed on 20 February 2025)) is the first FL framework designed to simulate multiple sources of heterogeneity, including device heterogeneity, behavioral heterogeneity, and their combinations [20]. FLASH enables Python-based FL simulations, incorporating heterogeneous clients with varying device types, training speeds, and network conditions. FLASH also simulates state and behavioral heterogeneity, representing client conditions such as training availability, idle states, charging status, and WiFi connectivity [45].

6.2.2. FedProx

FedProx (https://github.com/litian96/FedProx (accessed on 20 February 2025)) [46], an optimization framework for heterogeneous FL networks, addresses system and statistical heterogeneity by generalizing and re-parameterizing the original FedAvg algorithm [5]. Empirical results demonstrate that FedProx enhances convergence robustness compared to FedAvg, achieving a 22% improvement in absolute test accuracy under highly heterogeneous conditions. However, findings in [20] challenge this claim, suggesting alternative interpretations of FedProx's effectiveness.

6.2.3. q-Fair Federated Learning (q-FFL)

The work in [47] introduces Fair Federated Learning (FFL), a paradigm that simultaneously addresses fairness concerns and device heterogeneity. To enhance fairness, they propose q-Fair Federated Learning (q-FFL) (https://github.com/litian96/fair_flearn (accessed on 20 February 2025)), a method designed for equitable FL participation. q-FFL prioritizes underperforming devices by assigning higher aggregation weights to those with limited computational resources, thereby promoting a more uniform accuracy distribution across the network.

6.2.4. FedLab

FedLab (https://github.com/SMILELab-FL/FedLab (accessed on 20 February 2025)), (https://fedlab.readthedocs.io/en/master/index.html (accessed on 20 February 2025)) is a versatile FL framework that enables rapid implementation and testing of various FL algorithms [48]. FedLab offers tools and interfaces for FL algorithm implementation and evaluation. Its features include a communication framework for exchanging data and model updates, a model aggregation module, and predefined FL algorithms such as FedAvg, FedProx, and FedPerf. FedLab is designed to be flexible and customizable, allowing developers to easily extend or modify the framework to suit their specific needs. It also supports deep learning frameworks like TensorFlow and PyTorch, making integrating with existing Machine Learning pipelines easy. In addition, FedLab supports standalone, cross-machine, and hierarchical FL simulations.

Overall, FedLab provides a powerful and flexible platform for implementing and testing Federated Learning algorithms, making it a valuable tool for researchers and developers in machine learning.

6.2.5. EasyFL

EasyFL (https://github.com/EasyFL-AI/EasyFL (accessed on 20 February 2025)) is a low-code Federated Learning platform designed to simplify the process of building, deploying, and managing Federated Learning models [49]. The platform targets developers and data scientists needing more experience with Federated Learning or distributed systems. EasyFL provides a visual interface for designing Federated Learning workflows, including the specification of data sources, model architectures, and communication proto-

cols between the client and server. The platform also includes pre-built components for model training and aggregation, making it easier for users to integrate these functions into their workflows.

Table 6. Summary of frameworks, algorithms, and tools for Federated Learning.

Subject	Reference	Description	Advantages	Limitations
Open-source Frameworks for Federated Learning	[38]	TFF is a FL framework built on top of TensorFlow, providing APIs and core FL components.	Well integrated with TensorFlow, high scalability.	Limited support for cross-platform interoperability.
	[39]	Flower offers scalable FL experimentation, compatible with PyTorch and TensorFlow.	Simple API, flexible backends.	Lacks advanced optimizations for extreme heterogeneity scenarios.
	[40]	FATE is an industrial-grade, Python-based FL project for large-scale distributed training.	Strong security and privacy features.	Requires extensive system configuration and, a steeper learning curve.
	[41]	PADDLE is a deep learning framework with extensive FL support.	User-friendly, supports large-scale distributed training.	Less adoption outside China, fewer community resources.
	[42]	OpenFL is a Python3 FL framework hosted by Intel for secure and extensible FL development.	Strong hardware support, and robust security mechanisms.	Limited benchmarking datasets for statistical heterogeneity.
	[43]	PySyft is a Python-based framework for FL and secure data analysis.	Strong privacy-preserving computation capabilities.	Still under development, lacks standardized implementations.
	[44]	LEAF provides FL benchmarking datasets like FEMNIST to evaluate non-IID scenarios.	Real-world heterogeneous datasets for FL evaluation.	Limited to predefined datasets, does not simulate real-time FL deployments.
Heterogeneity-aware protocols, algorithms, and frameworks for Federated Learning	[21]	FLASH simulates multiple heterogeneity factors, including device and behavioral constraints.	Realistic simulation of heterogeneous conditions. Handles system and behavioral heterogeneity.	High computational cost. Limited validation in real-world deployments.
	[46]	FedProx introduces a regularized aggregation algorithm to handle system and statistical heterogeneity.	Improves convergence robustness and accuracy in heterogeneous environments. Supports system and statistical heterogeneity.	Computationally intensive due to the proximal term. Behavioral heterogeneity remains unaddressed.
	[47]	q-FFL optimizes FL fairness in non-IID data scenarios by reweighting updates.	Ensures balanced training by prioritizing underperforming clients. Addresses statistical heterogeneity.	Slower convergence due to reweighting mechanisms. Limited applicability to system and behavioral heterogeneity.

Table 6. *Cont.*

Subject	Reference	Description	Advantages	Limitations
Heterogeneity-aware protocols, algorithms, and frameworks for Federated Learning	[48]	FedLab is a flexible FL framework supporting multiple architectures and heterogeneity testing.	Highly customizable, supports multiple aggregation algorithms. Handles system and statistical heterogeneity.	Requires manual tuning for large-scale simulations. Lacks support for behavioral heterogeneity.
	[49]	EasyFL simplifies FL model development and deployment with a low-code interface.	User-friendly interface, fast experimentation. Supports system and statistical heterogeneity.	Limited flexibility for advanced customization. Behavioral heterogeneity is not addressed.

A key feature of EasyFL is its support for automatic code generation, which can save time and reduce the number of possible errors. Users can specify their workflow requirements using a visual interface, and the platform will automatically generate the necessary code to execute the workflow. Beyond its ease of use, EasyFL also introduces features that address heterogeneity in FL environments. The platform supports heterogeneous data distributions, allowing users to simulate non-IID scenarios and analyze the effects of statistical heterogeneity on training performance. In addition, EasyFL enables the configuration of different client hardware settings, providing a means to model system heterogeneity across FL participants.

Furthermore, EasyFL incorporates optimizations for communication efficiency and adaptive resource allocation, which mitigate network and system heterogeneity by ensuring that client contributions remain balanced despite variations in connectivity and computational power. These features allow researchers to explore fairness-aware training strategies and evaluate FL in settings where device and network constraints are highly variable. The platform remains under active development and is expected to be released as an open-source project. Table 6 summarizes frameworks, algorithms, and tools for Federated Learning.

7. Performance and Heterogeneity Analysis of Federated Learning on Distributed Datasets

This study focuses on the theoretical and simulation-based analysis of heterogeneity challenges in Federated Learning, particularly in the context of wireless networks such as 5G and 6G. While real-world testing of FL deployment in massive-scale, heterogeneous 5G/6G environments is a valuable area of future research, this work is primarily a survey and taxonomy study that synthesizes existing knowledge rather than conducting end-to-end empirical experiments in real deployments.

To provide practical insights, we have included a simulation-based evaluation using the LEAF framework, which is widely adopted for analyzing statistical heterogeneity in FL. However, we do not intend to experimentally validate each heterogeneity challenge in operational 5G/6G networks; as such, testing requires large-scale deployments that go beyond the scope of this survey. Future studies may further explore heterogeneous FL implementations in real 5G/6G network environments, particularly in the context of network infrastructure dynamics, latency constraints, and mobile edge computing integration.

In our experiments, we selected the FEMNIST dataset from the LEAF framework due to its strong alignment with real-world Federated Learning challenges. FEMNIST is an extended version of MNIST, where data samples are partitioned by individual writers, creating natural non-IID data distributions. This characteristic makes it highly suitable

for analyzing the effects of statistical heterogeneity, particularly in cases involving class imbalance and personalization techniques.

While the LEAF framework provides several datasets (e.g., Sent140 for text classification, Shakespeare for language modeling, and CelebA for facial attribute classification), FEMNIST was chosen over these alternatives because it effectively simulates both system and statistical heterogeneity in FL settings. In addition, its adoption in previous FL research facilitates meaningful comparative analysis. The dataset enables a controlled yet diverse participant environment, making it a robust benchmark for evaluating non-IID data effects in FL training. This selection ensures that our experiments accurately reflect practical FL scenarios while maintaining comparability with existing studies.

We present the results of a Federated Learning simulation using the FEMNIST dataset, a two-layer Convolutional Neural Network (CNN), and the FedAvg aggregation method. The simulation was performed in the LEAF environment with an 80–20 train–test split, running for 200 rounds across 50 clients. Two distinct data distribution configurations were tested: IID (independent and identically distributed) and non-IID (non-independent and identically distributed). The objective of this simulation was to assess the performance and impact of heterogeneity in FL within a multi-client setting. Further details are provided in the following subsection.

7.1. Exploration of Key Simulation Parameters

Our experiments leveraged the LEAF framework with a controlled set of key parameters to evaluate Federated Learning performance under heterogeneous conditions. We conducted simulations using 50 participating clients, balancing computational feasibility with realistic client heterogeneity. This number aligns with standard FL benchmarking practices and ensures a meaningful assessment of the system and statistical heterogeneity.

To assess the impact of data distribution on FL performance, we explored two configurations:

- IID Setting: Data were randomly and evenly distributed across clients to serve as a baseline.
- Non-IID Setting: Each client received data according to the original FEMNIST partitioning, where data are naturally grouped by writer, mimicking realistic user-based statistical heterogeneity.

To ensure effective training in a decentralized FL setting, we carefully tuned key hyperparameters:

- Learning Rate: Set at 0.01, decaying over rounds to balance convergence and stability.
- Batch Size: Fixed at 32 per client to accommodate memory constraints of edge devices.
- Local Epochs: Adjusted between 5 and 10, optimizing trade-offs between local learning and global aggregation.
- Optimizer: Stochastic gradient descent (SGD) with momentum was used to enhance stability and mitigate client drift.

These settings were determined through empirical validation and best practices in FL literature, ensuring a robust and scalable evaluation of Federated Learning under diverse conditions.

7.2. Evaluation Metrics and Justification

To assess the impact of heterogeneity on Federated Learning performance, we primarily use accuracy as the key evaluation metric. Accuracy provides a direct measure of how well the global model generalizes across diverse clients, making it a fundamental metric for analyzing performance disparities caused by heterogeneity. In addition, we evaluate training convergence trends by analyzing accuracy over multiple communication rounds.

This approach helps quantify the effect of statistical heterogeneity on model stability and learning efficiency.

To further examine the variation in client performance, we report percentile-based accuracy comparisons, highlighting disparities between high- and low-performing clients. This is particularly relevant in non-IID settings, where heterogeneity can lead to significant divergence in local model updates. Alternative metrics such as precision, recall, and F1-score were not prioritized, as our primary focus is on global model consistency across heterogeneous clients rather than class-specific performance. Our chosen metrics align with prior FL benchmarking studies and provide a comprehensive assessment of heterogeneity's impact on FL training.

7.3. Federated Learning Training on a FEMNIST Dataset: 50 Clients, 200 Rounds, Non-IID

The simulation results indicate a mean training accuracy of 0.82, with accuracy percentiles as follows: 10th percentile at 0.65, 50th percentile at 0.83, and 90th percentile at 0.94. These findings suggest that the Federated Learning algorithm achieved an overall high mean training accuracy. However, the percentile analysis reveals a significant degree of heterogeneity in client performance, as the 10th percentile accuracy of 0.65 indicates that the lowest-performing 10% of clients attained relatively low accuracy levels.

Conversely, the 50th percentile accuracy of 0.83 and 90th percentile accuracy of 0.94 demonstrate that the majority of clients performed well, achieving high accuracy rates. The substantial gap between the 10th and 50th percentiles highlights potential challenges in ensuring consistently high performance across all clients. This heterogeneity may stem from variations in data availability and quality, differences in computing power among client devices, and divergent training methodologies among participants. Figure 6 displays the performance result for the 50 clients in a non-IID scenario.

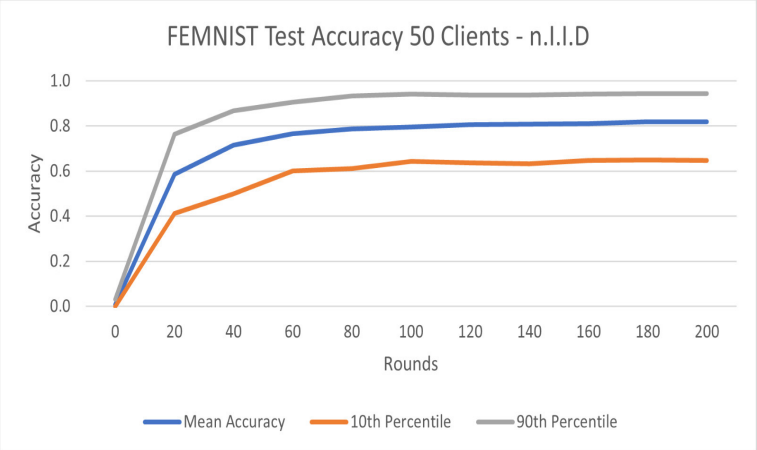


Figure 6. Performance results for the 50 clients, non-IID scenario.

In summary, the Federated Learning simulation using the FEMNIST dataset, a two-layer CNN, and the FedAvg aggregation method achieved a high mean training accuracy of 0.82, yet exhibited notable performance heterogeneity among clients. The percentile analysis indicates that while the majority of clients performed well, the lowest-performing 10% attained relatively low accuracy rates, suggesting that the FL algorithm may struggle to maintain consistently high performance across all clients. Further research and development are required to enhance algorithmic performance and mitigate heterogeneity in multi-client FL environments.

7.4. Federated Learning Training on a FEMNIST Dataset: 50 Clients, 200 Rounds, IID

The simulation yielded a mean training accuracy of 0.84, with accuracy percentiles as follows: 10th percentile at 0.74, 50th percentile at 0.84, and 90th percentile at 0.91. These results indicate a slight improvement in overall performance compared to the previous non-IID scenario. The following observations can be made:

First, the mean accuracy is marginally higher than in the non-IID scenario, as expected. In addition, acceptable accuracy values (0.8) were achieved earlier in the training process, specifically around the 20th round, leading to reduced training times, which is a notable advantage. Another key improvement is that the lower percentiles (representing the worst-performing clients) achieved significantly higher accuracy compared to the previous scenario. The 10th percentile accuracy is also closer to the mean accuracy, reflecting reduced heterogeneity in client performance due to a more balanced data distribution.

Furthermore, all accuracy curves demonstrate greater stability in later training rounds, with faster convergence to an acceptable accuracy level. However, it is important to acknowledge that FL performance is highly dependent on the dataset and model architecture. Thus, results may vary across different scenarios and distributions. In this case, the FEMNIST dataset, which is more realistic and distributed, was used for evaluation. Figure 7 displays the performance result for the 50 clients in a IID scenario.

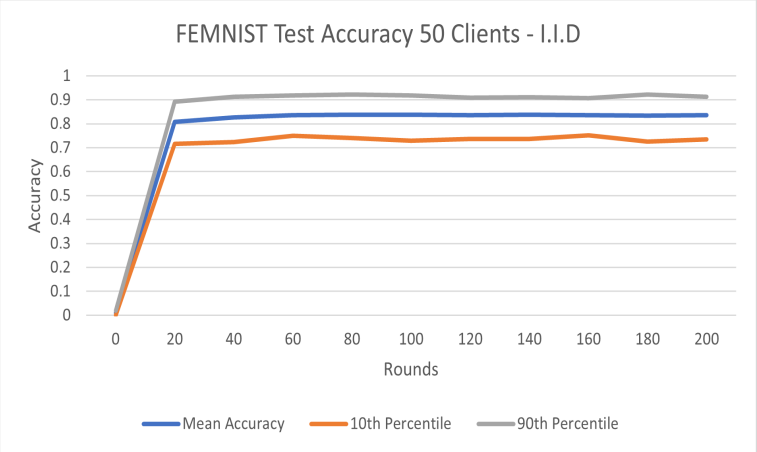


Figure 7. Performance results for the 50 clients, IID scenario.

The LEAF framework was selected for this study, as it is a widely adopted benchmarking tool specifically designed for Federated Learning. LEAF includes multiple datasets that introduce real-world non-IID characteristics, making it particularly relevant for analyzing statistical heterogeneity. Among the datasets available in LEAF, FEMNIST has been extensively used in prior works to evaluate the impact of heterogeneous data distributions on FL model convergence and accuracy.

While alternative FL frameworks such as Flower, FedLab, and OpenFL exist, many studies focusing on heterogeneity-aware FL approaches have relied on LEAF due to its structured dataset partitioning and its ability to simulate FL scenarios at scale. In addition, LEAF's benchmarking environment ensures reproducibility and comparability across different FL experiments. Given these factors, our choice of LEAF aligns with prior FL research and provides a rigorous foundation for assessing the impact of statistical heterogeneity in distributed learning environments.

7.5. Comparative Analysis of Proposed Approach vs. Existing Methods

To contextualize our results, we compare our method with existing Federated Learning aggregation strategies under the same experimental conditions. Table 7 summarizes key

performance metrics, including accuracy under IID and non-IID settings and the number of communication rounds required for convergence.

Table 7. Comparison of FL Methods on the FEMNIST dataset.

FL Method	Accuracy (Non-IID)	Number of Rounds	Dataset Used
FedAvg [5]	81.2%	200	FEMNIST
FedProx [46]	82.4%	200	FEMNIST
FedAvg (Our Experiment, Custom Config.)	83.1%	200	FEMNIST

The results demonstrate that we have successfully run and validated a widely recognized benchmark for Federated Learning, using the LEAF framework and FEMNIST dataset under a realistic non-IID setting. Our custom FedAvg configuration achieves comparable accuracy to existing methods, aligning with prior studies and confirming the reliability of the benchmark. The experiment confirms the expected performance of FedAvg [5] and FedProx [46] under the same experimental conditions. These findings reinforce that our implementation is consistent with the literature, ensuring that the benchmark framework remains a valid tool for future FL evaluations.

7.6. Running Time Comparison in Federated Learning Simulations

To further evaluate the efficiency of the simulations, we compare the training time across different Federated Learning conditions. Specifically, we analyze the time required for convergence under both IID and non-IID data distributions, as well as the impact of client heterogeneity on execution time.

Training Time Analysis: The total training time is influenced by the number of communication rounds required for convergence, client computational capacity, and network conditions. In our experiments, training under the non-IID setting resulted in longer convergence times due to increased variance in local model updates, requiring additional rounds to stabilize the global model.

Comparison of Convergence Rounds: We observe that in the IID configuration, the model stabilizes within fewer rounds, whereas in the non-IID setting, convergence is delayed due to statistical heterogeneity. This delay is expected, as clients with vastly different data distributions contribute updates that diverge more from the global model, increasing the number of aggregation cycles needed.

Impact of System Heterogeneity on Execution Time: Variations in client hardware and network conditions further contribute to differences in training time. Clients with lower computational power introduce straggler effects, where slower devices delay the aggregation process. In addition, network constraints such as bandwidth limitations can increase communication overhead, further impacting training efficiency.

These findings highlight the importance of optimizing FL frameworks to account for heterogeneity-related delays. Potential improvements include adaptive client selection, straggler mitigation techniques, and asynchronous model updates to improve efficiency in real-world FL deployments.

8. Discussion/Future Work

Federated Learning is a transformative Machine Learning paradigm poised to advance Artificial Intelligence and massive wireless communications. As a fundamental requirement for 6G mobile networks, FL is expected to be natively integrated into 6G architectures, facilitating distributed intelligence across devices. In addition, the Internet of Things (IoT)

is anticipated to synergize with FL, enabling efficient, decentralized learning in large-scale networks. This integration positions FL as a critical enabler of future AI-driven communication infrastructures.

Despite its potential, FL presents significant challenges that have attracted considerable attention from both the scientific community and industry. In this study, we have demonstrated that most FL challenges can be effectively analyzed through the lens of heterogeneity. Heterogeneity naturally arises in FL due to the highly diverse characteristics of participating devices. Research has extensively shown that unaddressed heterogeneity can degrade model performance and hinder global model convergence.

FL heterogeneity manifests in three primary dimensions: system heterogeneity, data heterogeneity, and behavioral heterogeneity. Each presents distinct challenges and weaknesses, yet they can also coexist within real-world FL implementations, leading to compounded adverse effects on model training and performance. Addressing these heterogeneity-related challenges is essential for improving FL scalability, reliability, and effectiveness in next-generation communication systems.

Heterogeneity in Federated Learning is a critical and evolving research challenge. It has driven researchers to conduct extensive studies evaluating the impact of device variability, data distribution, network conditions, and other factors on FL performance. These studies provide quantitative insights into the degradation effects of heterogeneity in real-world FL implementations. In response, researchers are actively developing mitigation strategies to reduce the negative impact of heterogeneity on model performance and convergence. This survey has presented key advancements, including state-of-the-art frameworks, algorithms, and tools, that address FL heterogeneity challenges and contribute to enhancing its scalability and robustness.

Table 8 presents a summarized classification of heterogeneity in Federated Learning, highlighting key challenges and their corresponding real-world use cases. This taxonomy encapsulates the primary obstacles encountered due to system, statistical, and behavioral heterogeneity, emphasizing their impact on FL performance in real-world wireless networks. Addressing these challenges requires tailored optimization techniques, including adaptive model aggregation, personalized FL approaches, and incentive mechanisms. Future research directions should focus on mitigating the compounded effects of these heterogeneity types, particularly in large-scale 6G and IoT deployments. The insights in this table serve as a foundation for developing more robust and scalable FL frameworks.

Table 8. Summary of heterogeneity categories, corresponding challenges, and real-world use cases.

Heterogeneity Type	Reference	Challenges and Real-World Use Cases
System Heterogeneity	Nguyen et al. [4]	System heterogeneity impacts Federated Learning due to diverse client capabilities, leading to straggler effects and inefficiencies in global model aggregation. Smart-city applications and industrial IoT settings must consider computational imbalances among edge devices.
	Beitollahi et al. [13]	Network constraints in FL lead to communication overhead, packet loss, and unstable training. 5G/6G networks must optimize model updates and aggregation frequency to mitigate bandwidth limitations.
	Liu et al. [1]	Heterogeneous network infrastructure, particularly in edge AI and IoT deployments, complicates FL model training. Adaptive compression and decentralized aggregation can improve training efficiency in such environments.

Table 8. *Cont.*

Heterogeneity Type	Reference	Challenges and Real-World Use Cases
Statistical Heterogeneity	Abdelmoniem et al. [8]	Non-IID data distributions across clients cause model divergence, reducing FL accuracy and fairness. Healthcare FL models trained across multiple hospitals face challenges in patient data consistency and bias.
	Li et al. [7]	Class imbalance in FL impacts generalization performance. In Federated autonomous driving applications, different driving environments lead to varied data distributions, necessitating adaptive personalization techniques.
	Pei et al. [9]	Massively distributed FL data in IoT networks leads to training inefficiencies. Cross-device learning in sensor-based industrial applications faces difficulty in aggregating meaningful updates from noisy data.
Behavioral Heterogeneity	Wang et al. [10]	Free-riding behaviors reduce FL model robustness. Crowdsourced FL applications, such as recommendation systems, require incentive mechanisms to ensure fair client participation.
	Chakraborty et al. [11]	Reputation-based aggregation is crucial in FL to mitigate unreliable client updates. Financial fraud detection models leverage trust-aware aggregation to exclude malicious data contributions.
	Abdelmoniem et al. [8]	Incentive-driven FL can enhance participation in mobile networks. Blockchain-based reward systems can promote honest reporting in distributed FL models.

8.1. Practical Implications of Heterogeneity in Real-World FL Deployments

Heterogeneity in Federated Learning presents significant challenges that must be addressed for real-world deployment. System, statistical, and behavioral heterogeneity can impact model performance, fairness, and efficiency. These challenges manifest in various ways, such as straggler effects due to device heterogeneity, model divergence caused by non-IID data distributions, network-related bottlenecks, and inconsistent client participation. Understanding these challenges is crucial for designing FL systems that can operate effectively in diverse environments.

One of the most prominent issues is device heterogeneity, where clients differ in computational power, memory, and energy availability. Slower devices often become stragglers, delaying global model updates and reducing training efficiency. To mitigate this, adaptive client selection and priority-based aggregation can be implemented, ensuring that devices with adequate resources contribute effectively to model updates. In addition, straggler mitigation techniques, such as dynamic resource allocation and load balancing, help improve overall training efficiency.

Another significant concern is statistical heterogeneity, where clients contribute non-IID data distributions that cause model divergence. This can reduce the effectiveness of global model updates, leading to biased or less generalizable models. Personalized Federated Learning (PFL) methods, such as clustering clients with similar data distributions or applying adaptive weighting to client contributions, help maintain model stability and improve accuracy in heterogeneous environments.

Network constraints also play a critical role in real-world FL deployments, as communication delays, bandwidth limitations, and packet loss can slow down training. High communication overhead increases the time required for model convergence, making FL impractical in large-scale settings. To address this, techniques such as asynchronous FL,

gradient compression, and decentralized aggregation have been proposed to minimize data transfer while preserving model performance.

In addition to system and statistical challenges, behavioral heterogeneity significantly affects FL deployments. Clients may drop out due to energy limitations, while others may participate inconsistently or attempt to free-ride by benefiting from the global model without actively contributing. These behaviors can lead to unfair model training and skewed performance results. Deploying incentive mechanisms, such as blockchain-based reward systems and reputation-based aggregation, encourages active participation while penalizing free-riding. Energy-aware scheduling can further improve participation by optimizing training times based on device availability.

To ensure effective FL deployments in heterogeneous environments, several strategies can be integrated into FL frameworks. Adaptive aggregation techniques, such as FedProx and reinforcement learning-based client selection, can optimize model updates based on client capabilities. Efficient communication strategies, including model quantization, sparsification, and selective client updates, help minimize bandwidth consumption. Personalized FL approaches, which tailor model updates to individual client data characteristics, further enhance model generalization. Finally, incorporating incentive mechanisms ensures fair client participation and maintains long-term engagement in FL networks.

Real-world FL applications, such as healthcare, smart cities, and industrial IoT, require a balance between privacy, model performance, and deployment efficiency. By addressing the challenges of heterogeneity with adaptive learning strategies and optimized resource allocation, FL can be deployed effectively in diverse, large-scale environments. These considerations will be crucial for the future of FL research and real-world adoption.

8.2. Future Directions

As Federated Learning (FL) continues to evolve, addressing heterogeneity remains a crucial challenge for real-world deployment. Our taxonomy of heterogeneity challenges provides a structured framework for identifying and mitigating key limitations in FL systems. Future research directions should leverage this classification to develop novel techniques that enhance FL efficiency and adaptability across diverse environments.

One crucial area for future exploration is the optimization of FL aggregation strategies. While existing methods such as FedAvg and FedProx provide baseline aggregation mechanisms, they do not fully account for extreme heterogeneity in client data distributions and computational capacities. New approaches integrating personalized FL models, adaptive weighting, and hierarchical aggregation could improve training performance in non-IID settings. Multi-task learning techniques, which allow different clients to train models tailored to their specific needs while sharing a core global model, could further enhance performance in diverse applications.

Another critical direction is the development of heterogeneity-aware communication strategies. Bandwidth constraints and network instability significantly impact FL convergence times. Techniques such as gradient compression, asynchronous updates, and decentralized FL architectures could help mitigate communication overhead while maintaining model accuracy. These methods would allow FL to scale efficiently in dynamic environments, such as mobile networks and IoT deployments.

The taxonomy introduced in this study also highlights the role of behavioral heterogeneity in FL systems. Future research should focus on designing trust-aware aggregation mechanisms that can detect unreliable client contributions and integrate incentive-based participation models to encourage sustained engagement in FL training. Blockchain-based verification methods and fairness-aware aggregation could help mitigate the effects of free-riding and malicious clients.

To further validate and refine the proposed taxonomy, additional experimental setups should be explored. Extending evaluations beyond FEMNIST to other realistic datasets, such as those in industrial IoT, healthcare, and autonomous driving, would provide a broader perspective on heterogeneity challenges. Varying client-selection strategies, implementing cross-silo FL experiments, and testing different levels of data and system heterogeneity could offer deeper insights into the practical impact of heterogeneity-aware FL techniques.

By building upon the challenges outlined in our taxonomy, future research can contribute to the development of more resilient and adaptable FL methodologies, ensuring their practical applicability in next-generation wireless networks and distributed AI systems.

9. Conclusions

Federated Learning represents an innovative and transformative technology that bridges Machine Learning (ML) and large-scale wireless communication networks. By enabling collaborative learning on decentralized datasets, FL facilitates the development of more refined and robust models, which are essential for intelligent applications in next-generation networks.

Heterogeneity is an inherent characteristic of real-world FL implementations, encompassing variations in device capabilities, data distributions, and network conditions among participants. Research on heterogeneity challenges in FL remains in its early stages, with increasing recognition of its impact on model performance and system efficiency. Studies are now focused on quantifying the effects of different heterogeneity types and developing mitigation strategies. Future advancements may lead to integrated solutions that holistically address heterogeneity, enabling FL deployment in more realistic and functional environments.

In this study, we present heterogeneity in three main categories. System heterogeneity arises from variations in processing power, battery levels, and network connectivity among participating devices. Statistical heterogeneity occurs when training data are non-IID, insufficient in quantity or quality, or contains inconsistent, fake, or invalid samples. Behavioral heterogeneity stems from user participation differences, including unwillingness to contribute, demand for fair compensation, or strategic manipulation of contributions. Each heterogeneity category introduces significant challenges to FL, potentially leading to accuracy degradation, model divergence, and increased training times. More critically, heterogeneity types often coexist, compounding their negative effects and posing a substantial barrier to deploying FL in real-world applications.

There is an increasing necessity to ease the study process and experiment with Federated Learning. We presented some open-source and heterogeneity-aware frameworks, tools, and algorithms regarding Federated Learning. The current trend is to develop accessible and scalable tools for starting to experiment in Federated Learning. In this way, researchers and anyone interested in studying Federated Learning do not have to start from scratch. They can quickly simulate Federated Learning without taking too much care about manually creating a central aggregator server, configuring clients, and aggregation protocols. Instead, they can take care of high-level considerations such as the learning model they want to use, the setup of hyperparameters, or even the architecture or the number of clients they want to simulate.

A key contribution of this study is the inclusion of behavioral heterogeneity as a fundamental aspect of Federated Learning challenges. While system and statistical heterogeneity have been widely studied, behavioral heterogeneity introduces an additional layer of complexity that significantly impacts FL performance. Differences in user participation, incentive-driven engagement, and adversarial behavior can lead to issues such as unfair

model updates, free-riding, or malicious data manipulation, all of which threaten the integrity and efficiency of the FL process. By categorizing behavioral heterogeneity alongside system and statistical heterogeneity, this study provides a comprehensive perspective on the diverse factors influencing FL deployments. Addressing behavioral heterogeneity is essential for improving the fairness, security, and sustainability of FL systems, particularly in large-scale, real-world applications.

Another important aspect of this work is the compilation of relevant and up-to-date studies addressing the heterogeneity challenge in future Federated Learning technology. We present a broad overview of this challenge, summarizing and classifying each related problem accordingly. In addition, we review studies that quantify the impact of heterogeneity both individually and in cases where multiple heterogeneity issues coexist. Lastly, we introduce useful tools for initiating research in Federated Learning and experimenting with heterogeneity-related challenges, making them accessible so that researchers can conduct their own experiments and draw informed conclusions.

Author Contributions: T.B.S.: Conceptualization, methodology, validation, formal analysis, writing—original draft, investigation, data curation; L.I.B.L.: writing, review, editing, supervision, project administration, funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: The funding of this research was granted by Escuela Politécnica Nacional, Quito, Ecuador.

Data Availability Statement: Data are contained within the article.

Acknowledgments: The authors would like to thank the Faculty of Systems Engineering, Escuela Politécnica Nacional, Quito, Ecuador for the academic and financial support.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Liu, Y.; Yuan, X.; Xiong, Z.; Kang, J.; Wang, X.; Niyato, D. Federated learning for 6G communications: Challenges, methods, and future directions. *China Commun.* **2020**, *17*, 105–118. [CrossRef]
2. Luo, B.; Liu, X.; Yu, S.; Pan, M.; Xu, Q.; Yao, G. Optimization Design for Federated Learning in Heterogeneous 6G Networks. *arXiv* **2023**, arXiv:2303.08322. [CrossRef]
3. Ben Driss, M.; El Kiram, H.; Boudia, M. Federated Learning for 6G: Paradigms, Taxonomy, Recent Advances and Insights. *arXiv* **2023**, arXiv:2312.04688.
4. Nguyen, D.C.; Ding, M.; Pathirana, P.N.; Seneviratne, A.; Li, J.; Poor, H.V. Federated learning for internet of things: A comprehensive survey. *IEEE Commun. Surv. Tutor.* **2021**, *23*, 1622–1658. [CrossRef]
5. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B.A. Communication-efficient learning of deep networks from decentralized data. In Proceedings of the Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 20–22 April 2017; PMLR: Cambridge, MA, USA, 2017; pp. 1273–1282.
6. Lim, W.Y.B.; Luong, N.C.; Hoang, D.T.; Jiao, Y.; Liang, Y.C.; Yang, Q.; Niyato, D.; Miao, C. Federated Learning in Mobile Edge Networks: A Comprehensive Survey. *IEEE Commun. Surv. Tutor.* **2020**, *22*, 2031–2063. [CrossRef]
7. Li, T.; Sahu, A.K.; Talwalkar, A.; Smith, V. Federated Learning: Challenges, Methods, and Future Directions. *IEEE Signal Process. Mag.* **2020**, *37*, 50–60. [CrossRef]
8. Abdelmoniem, A.M.; Ho, C.Y.; Papageorgiou, P.; Canini, M. A comprehensive empirical study of heterogeneity in federated learning. *IEEE Internet Things J.* **2023**, *10*, 14071–14083. [CrossRef]
9. Pei, J.; Liu, W.; Li, J.; Wang, L.; Liu, C. A Review of Federated Learning Methods in Heterogeneous scenarios. *IEEE Trans. Consum. Electron.* **2024**, *70* , 5983–5999. [CrossRef]
10. Wang, Y.; Kantarci, B. Reputation-enabled federated learning model aggregation in mobile platforms. In Proceedings of the ICC 2021-IEEE International Conference on Communications, Montreal, QC, Canada, 14–23 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–6.
11. Chakraborty, S.; Boudguiga, A. A Decentralized Federated Learning Framework with Client Reputation and Secure Aggregation. *IACR Cryptol. EPrint Arch.* **2024**, *2024*, 506.

12. Ye, M.; Fang, X.; Du, B.; Yuen, P.C.; Tao, D. Heterogeneous federated learning: State-of-the-art and research challenges. *ACM Comput. Surv.* **2023**, *56*, 1–44. [CrossRef]
13. Beitollahi, M.; Lu, N. Federated learning over wireless networks: Challenges and solutions. *IEEE Internet Things J.* **2023**, *10*, 14749–14763. [CrossRef]
14. Mengistu, T.M.; Kim, T.; Lin, J.W. A Survey on Heterogeneity Taxonomy, Security and Privacy Preservation in the Integration of IoT, Wireless Sensor Networks and Federated Learning. *Sensors* **2024**, *24*, 968. [CrossRef] [PubMed]
15. Chai, Z.; Fayyaz, H.; Fayyaz, Z.; Anwar, A.; Zhou, Y.; Baracaldo, N.; Ludwig, H.; Cheng, Y. Towards taming the resource and data heterogeneity in federated learning. In Proceedings of the 2019 USENIX Conference on Operational Machine Learning (OpML 19), Santa Clara, CA, USA, 20 May 2019; pp. 19–21.
16. Liberti, F.; Lagunas, E.; de la Roche, G.; Capdevielle, V.; Guo, W.; Balta, H. Federated Learning in Dynamic and Heterogeneous Environments: Advantages, Performances, and Privacy Problems. *Appl. Sci.* **2023**, *14*, 8490. [CrossRef]
17. Konečný, J.; McMahan, H.B.; Yu, F.X.; Richtarik, P.; Suresh, A.T.; Bacon, D. Federated Learning: Strategies for Improving Communication Efficiency. In Proceedings of the NIPS Workshop on Private Multi-Party Machine Learning, Barcelona, Spain, 9 December 2016.
18. Abdelmoniem, A.M.; Ho, C.Y.; Papageorgiou, P.; Canini, M. Empirical analysis of federated learning in heterogeneous environments. In Proceedings of the 2nd European Workshop on Machine Learning and Systems, Rennes, France, 5–8 April 2022; pp. 1–9.
19. Xu, Z.; Yang, Z.; Xiong, J.; Yang, J.; Chen, X. Elfish: Resource-aware federated learning on heterogeneous edge devices. *Ratio* **2019**, *2*, r2.
20. Abdelmoniem, A.M.; Ho, C.Y.; Papageorgiou, P.; Bilal, M.; Canini, M. On the Impact of Device and Behavioral Heterogeneity in Federated Learning. *arXiv* **2021**, arXiv:2102.07500.
21. Yang, C.; Xu, M.; Wang, Q.; Chen, Z.; Huang, K.; Ma, Y.; Bian, K.; Huang, G.; Liu, Y.; Jin, X.; et al. Flash: Heterogeneity-aware federated learning at scale. *IEEE Trans. Mob. Comput.* **2022**, *23*, 483–500. [CrossRef]
22. Cox, B.; Chen, L.Y.; Decouchant, J. Aergia: Leveraging heterogeneity in federated learning systems. In Proceedings of the 23rd ACM/IFIP International Middleware Conference, Quebec, QC, Canada, 7–11 November 2022; pp. 107–120.
23. Chai, Z.; Ali, A.; Zawad, S.; Truex, S.; Anwar, A.; Baracaldo, N.; Zhou, Y.; Ludwig, H.; Yan, F.; Cheng, Y. Tifl: A tier-based federated learning system. In Proceedings of the 29th International Symposium on High-Performance Parallel and Distributed Computing, Stockholm, Sweden, 23–26 June 2020; pp. 125–136.
24. Cheng, X.; Li, C.; Liu, X. A review of federated learning in energy systems. In Proceedings of the 2022 IEEE/IAS Industrial and Commercial Power System Asia (I&CPS Asia), Shanghai, China, 8–11 July 2022; pp. 2089–2095.
25. Kim, Y.G.; Wu, C.J. Autofl: Enabling heterogeneity-aware energy efficient federated learning. In Proceedings of the MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture, Virtual Event, 18–22 October 2021; pp. 183–198.
26. Huang, W.; Li, T.; Wang, D.; Du, S.; Zhang, J. Fairness and accuracy in federated learning. *arXiv* **2020**, arXiv:2012.10069.
27. Huang, W.; Ye, M.; Du, B. Learn from others and be yourself in heterogeneous federated learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10143–10153.
28. Li, X.; Huang, K.; Yang, W.; Wang, Z.; Zhang, Z. On the Convergence of FedAvg on Non-IID Data. *arXiv* **2019**, arXiv:1907.02189.
29. Fang, X.; Ye, M. Robust federated learning with noisy and heterogeneous clients. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10072–10081.
30. Tan, A.Z.; Yu, H.; Cui, L.; Yang, Q. Towards personalized federated learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *34*, 9587–9603. [CrossRef] [PubMed]
31. Yang, Q.; Liu, Y.; Cheng, Y.; Kang, Y.; Chen, T.; Yu, H. Federated Learning. *Synth. Lect. Artif. Intell. Mach. Learn.* **2019**, *13*, 1–207. [CrossRef]
32. Chen, X.; Liu, C.; Li, B.; Lu, K.; Song, D. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv* **2017**, arXiv:1712.05526.
33. Li, D.; Wang, J. Fedmd: Heterogenous federated learning via model distillation. *arXiv* **2019**, arXiv:1910.03581.
34. Alazab, M.; RM, S.P.; Parimala, M.; Maddikunta, P.K.R.; Gadekallu, T.R.; Pham, Q.V. Federated learning for cybersecurity: Concepts, challenges, and future directions. *IEEE Trans. Ind. Inform.* **2021**, *18*, 3501–3509. [CrossRef]
35. Verma, D.C.; White, G.; Julier, S.; Pasteris, S.; Chakraborty, S.; Cirincione, G. Approaches to address the data skew problem in federated learning. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*; SPIE: Bellingham, WA, USA, 2019; Volume 11006, pp. 542–557.
36. Zeng, M.; Min, G.; Wu, J.; Li, Y.; Luo, C.; Tian, H.; Wang, X. A comprehensive survey of incentive mechanism design for federated learning. *arXiv* **2021**, arXiv:2106.15406.
37. Tu, Y.; Li, L.; Feng, Z.; Yu, C.; Zhang, J.; Chen, X. Incentive mechanisms for federated learning: A systematic literature review. *arXiv* **2021**, arXiv:2111.11850.

38. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. TensorFlow: A system for large-scale machine learning. In Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), Savannah, GA, USA, 2–4 November 2016; pp. 265–283.
39. Beutel, D.J.; Topal, T.; Mathur, A.; Qiu, X.; Fernandez-Marques, J.; Gao, Y.; Sani, L.; Li, K.H.; Parcollet, T.; de Gusmão, P.P.B.; et al. Flower: A friendly federated learning research framework. *arXiv* **2020**, arXiv:2007.14390.
40. Liu, Y.; Fan, T.; Chen, T.; Xu, Q.; Yang, Q. Fate: An industrial grade platform for collaborative learning with data protection. *J. Mach. Learn. Res.* **2021**, *22*, 1–6.
41. Kholod, I.; Yanaki, E.; Fomichev, D.; Shalugin, E.; Novikova, E.; Filippov, E.; Nordlund, M. Open-source federated learning frameworks for IoT: A comparative review and analysis. *Sensors* **2020**, *21*, 167. [CrossRef]
42. Foley, P.; Sheller, M.J.; Edwards, B.; Pati, S.; Riviera, W.; Sharma, M.; Moorthy, P.N.; Wang, S.h.; Martin, J.; Mirhaji, P.; et al. OpenFL: The open federated learning library. *Phys. Med. Biol.* **2022**, *67*, 214001. [CrossRef]
43. Ryffel, T.; Trask, A.; Dahl, M.; Wagner, B.; Mancuso, J.; Rueckert, D.; Passerat-Palmbach, J. A generic framework for privacy preserving deep learning. *arXiv* **2018**, arXiv:1811.04017.
44. Caldas, S.; Duddu, S.M.K.; Wu, P.; Li, T.; Konečnỳ, J.; McMahan, H.B.; Smith, V.; Talwalkar, A. Leaf: A benchmark for federated settings. *arXiv* **2018**, arXiv:1812.01097.
45. Yang, C.; Wang, Q.; Xu, M.; Chen, Z.; Bian, K.; Liu, Y.; Liu, X. Characterizing impacts of heterogeneity in federated learning upon large-scale smartphone data. In Proceedings of the Web Conference 2021, Ljubljana, Slovenia, 19–23 April 2021; pp. 935–946.
46. Li, T.; Sahu, A.K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; Smith, V. Federated optimization in heterogeneous networks. *Proc. Mach. Learn. Syst.* **2020**, *2*, 429–450.
47. Li, T.; Sanjabi, M.; Beirami, A.; Smith, V. Fair resource allocation in federated learning. *arXiv* **2019**, arXiv:1905.10497.
48. Zeng, D.; Liang, S.; Hu, X.; Wang, H.; Xu, Z. Fedlab: A flexible federated learning framework. *J. Mach. Learn. Res.* **2023**, *24*, 1–7.
49. Zhuang, W.; Gan, X.; Wen, Y.; Zhang, S. Easyfl: A low-code federated learning platform for dummies. *IEEE Internet Things J.* **2022**, *9*, 13740–13754. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.