Communication Optimization Techniques in Personalized Federated Learning: Applications, Challenges and Future Directions

Fahad Sabah[a,b], Yuwen Chen[a], Zhen Yang[a,*], Abdul Raheem[a], Muhammad Azam[b], Nadeem Ahmad[b] and Raheem Sarwar[c]

[a]Faculty of Information Technology, Beijing University of Technology, Beijing, China
[b]Faculty of CS&IT, Superior University, Lahore, Pakistan
[c]OTEHM, Faculty of Business and Law, Manchester Metropolitan University, United Kingdom

Keywords:
Federated learning
Personalized federated learning
Communication optimization
Privacy preservation
Communication efficiency.

ABSTRACT

Personalized Federated Learning (PFL) aims to train machine learning models on decentralized, heterogeneous data while preserving user privacy. This research survey examines the core communication challenges in PFL and evaluates optimization strategies to address key issues, including data heterogeneity, high communication costs, model drift, privacy vulnerabilities, and device variability. We provide a comprehensive analysis of key communication optimization techniques; Model Compression, Differential Privacy, Client Selection, Asynchronous Updates, Gradient Compression, and Model Caching, by their efficiency and effectiveness under diverse PFL conditions. Our study quantitatively compares these methods, identifies limitations, and proposes enhanced strategies to improve communication efficiency, reduce latency, and maintain model accuracy. This research delivers actionable insights for optimizing PFL communication, enhancing both model performance and privacy safeguards. Overall, this work serves as a valuable resource for researchers and practitioners, offering practical guidance on leveraging advanced communication techniques to drive PFL improvements and highlighting promising directions for future research.

1. Introduction

Personalized federated learning (PFL) is a subset of federated learning (FL) that aims to develop machine learning models that are customized to the unique characteristics and preferences of individual devices or users within a federated network. In PFL, an initial global model is trained on a central server and serves as a shared starting point for all clients. Each client then fine-tunes this global model using its local data, thus creating a personalized model that reflects its specific data distribution and user preferences. This approach preserves data privacy while leveraging the collective knowledge and experiences of all clients. This interest is fueled by the growing necessity to tailor global models to individual user preferences while ensuring privacy, especially in domains such as healthcare, finance, and mobile applications [1, 2, 3, 4, 5]. Despite its promise, PFL remains an emerging research area, necessitating a thorough survey to evaluate the current state of the art, identify challenges, and suggest future research directions. This survey aims to enhance the understanding of PFL and its potential research areas, particularly concerning communication optimization.

Bhosle & Musande [6] emphasizes the effectiveness of CNN in character and digit recognition tasks, achieving high accuracy levels. They highlighted how CNN-based approaches, which handle both structured and unstructured data effectively, can inspire communication-efficient PFL models, particularly for resource-constrained devices. Sun et al. [7] introduced a GNN-DCGAN model to predict ice resistance in polar ship navigation, demonstrating improved accuracy through data augmentation and graph neural networks. Song et al. [8] proposed a dual-population evolutionary algorithm that outperforms traditional optimization techniques in maintaining diversity and convergence. This work provides insights into multi-objective optimization, which relates to optimizing client selection and communication strategies in PFL to balance resource utilization and model performance.

Akande et al. [9] explored how deep learning models can improve the accuracy and efficiency of computer-aided engineering simulations for vehicle wheels. The study by Chai et al. explored a multi-objective evolutionary algorithm (MOEA/D) to enhance communication efficiency in FL. This method focused on optimizing the structure of the global model, thereby achieving efficient evolution in neural networks, which ultimately reduces communication overhead without sacrificing accuracy. The authors confirmed that MOEA/D outperformed traditional algorithms like NSGA-II by achieving faster convergence and better scalability for model structure optimization [10]. To address the issues of personalization in FL, Fan et al. introduced MiniPFL, a hierarchical model that leverages a two-layered approach to balance client heterogeneity. MiniPFL decomposes FL into shallow and deep layers, with each layer holding common and personalized information, respectively. Experimental results indicate that MiniPFL reduces communication rounds by up to 30% while enhancing model

*Corresponding author

accuracy by 2.7% on various datasets, including CIFAR-10 and Tiny-Imagenet [11].

In addressing the statistical heterogeneity among clients, Tu et al. proposed a PFL method, pFedLT, that incorporates meta-learning strategies and layer-wise feature transformations. This approach utilizes scaling and shifting operations to capture client-specific data distributions, significantly improving accuracy and reducing communication costs under non-IID settings [12]. The practical integration of FL with edge computing has been advanced by Song et al. with their personalized federated deep reinforcement learning (PF-DRL) model. This model adapts deep reinforcement learning for trajectory optimization in multi-UAV systems, ensuring robust performance in heterogeneous environments. The authors reported that PF-DRL offers improved convergence rates and higher service quality for edge-based applications compared to traditional methods [13].

In order to improve the efficiency of client communication in FL, Wu et al. developed an adaptive client and communication optimization algorithm. This solution dynamically adjusts client selection and communication intervals based on runtime data, effectively reducing convergence time for FL models and enhancing their scalability [14]. Meanwhile, Wu et al. proposed FedKD, a knowledge distillation-based FL model, which employs gradient compression and mutual knowledge distillation to cut down on communication costs by nearly 95%, while maintaining performance comparable to centralized learning [15]. To address issues of resource limitations in edge FL, Yuan et al. proposed a lightweight FL model that utilizes pruning and masking techniques to optimize communication and computation costs. This approach allows for efficient deployment on edge devices and improves model accuracy by 9.36% over state-of-the-art techniques [16].

In the domain of wireless networks, Fan et al. introduced a PFL model that allocates local fine-tuning learning rates and optimizes communication resources, leading to a faster convergence and improved accuracy in wireless environments [17]. Complementing these efforts, Wang et al. proposed a self-knowledge distillation framework tailored for digital twins in Industrial IoT. This model addresses the challenges posed by heterogeneous industrial data and mitigates issues of historical knowledge forgetting, enabling high performance in digital twin applications [18]. Another critical area of FL research focuses on optimizing communication pathways to reduce latency and unnecessary model updates. Traditional FL systems, relying on cloud servers, often encounter high transmission delays when handling concurrent client updates, leading to inefficiencies. Wang et al. address this by introducing an edge-based communication framework that employs mobile edge nodes to act as communication hubs, thereby reducing the load on central servers. The framework also employs cosine similarity to filter out unnecessary model updates, achieving significant reductions in communication costs and faster convergence rates [19].

PFL has been proposed to address the varying data distributions across clients, yet many PFL algorithms suffer from negative knowledge transfer and high communication costs. To tackle these limitations, Wu et al. propose the FEDORA framework, which reframes PFL as a privacy-preserving transfer learning problem. The FEDORA framework incorporates adaptive parameter propagation based on client task similarity and selective regularization, effectively enhancing generalization and reducing communication costs by preventing negative knowledge transfer [20].

In this survey, we offer a detailed overview of communication optimization strategies in PFL, encompassing their definitions, algorithms, architectures, challenges, and future research directions. We begin by introducing the fundamental concepts of personalization, followed by a discussion of various communication optimization techniques, focusing on their architectures and communication methods. We emphasize future directions and open research questions in PFL strategies. Finally, we summarize the key contributions and limitations of existing PFL research and propose directions for future studies. We believe this survey will be a valuable resource for researchers and practitioners interested in PFL and will help advance the state-of-the-art (SOTA) in this exciting research field.

1.1. Definition of PFL

Figure 1 shows the basic architecture of PFL which is a type of FL in which the goal is to train machine learning models that are tailored to the individual characteristics of each device or user in a federated network [21, 22]. Generally in PFL, a global model is initially trained on a central server, which serves as a starting point for all clients. Then, each client further fine-tunes the global model using its local data, while preserving the privacy of its data. As a result, the global model is personalized for each client, based on its unique data and preferences, while still benefiting from the collective knowledge of all clients [23].

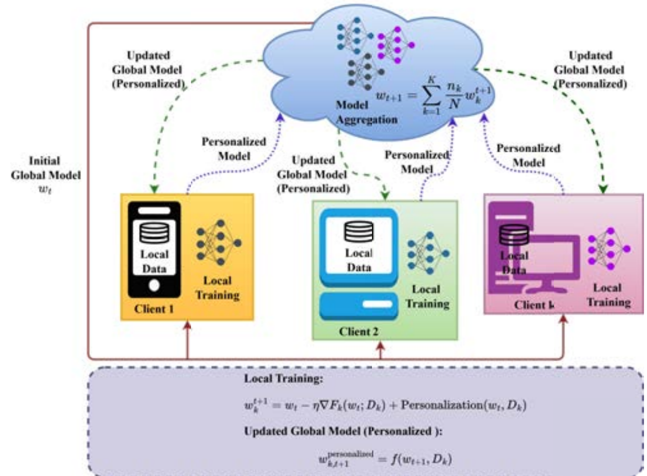

Figure 1: Basic Personalized Federated Learning Architecture

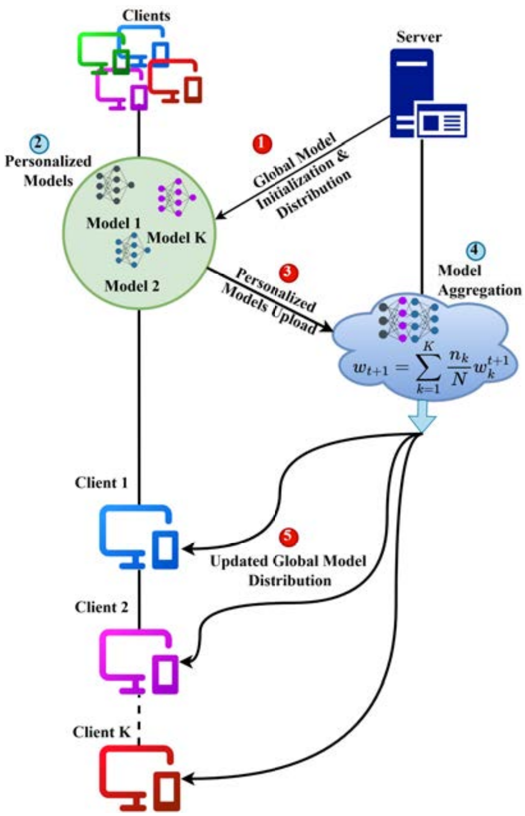

Figure 2: Communication process in PFL

Figure 2 shows the communication process of PFL, following are the details of each step involved, especially the communication steps shown in red:

Let, K be the total number of clients. $S_t \subseteq \{1, 2, \ldots, K\}$ be the subset of clients selected in round t. w_i^t represents the local model weights of client i after local training in round t. $\Delta w_i^t = w_i^t - w_{\text{global}}^t$ is the model update (i.e., the difference between the local model and the global model at round t). w_{global}^t is the global model at round t.

In each communication round t, a subset of clients S_t is selected based on certain criteria such as availability, trust level, or network bandwidth. The number of clients selected is denoted as $|S_t|$.

Each selected client $i \in S_t$ updates its local model w_i^t using its local data D_i. The update follows the optimization of the client's personalized objective:

$$w_i^{t+1} = w_i^t - \eta \nabla L_i(w_i^t, D_i) \tag{1}$$

Where, η is the learning rate, and $L_i(w_i^t, D_i)$ is the loss function for client i with respect to its local data D_i.

After local training, each client sends its model update Δw_i^t to the server. The communication cost is proportional to the size of the model parameters, denoted by $|\Delta w_i^t|$, which represents the number of parameters or bytes exchanged.

The total communication cost from the selected clients to the server in round t is:

$$C_{\text{client-server}}^t = \sum_{i \in S_t} |\Delta w_i^t| \tag{2}$$

The server aggregates the model updates from the selected clients. A common aggregation method is Federated Averaging (FedAvg), where the global model is updated as the weighted average of the client updates:

$$w_{\text{global}}^{t+1} = w_{\text{global}}^t + \frac{1}{|S_t|} \sum_{i \in S_t} \Delta w_i^t \tag{3}$$

This aggregation can also be weighted by the number of local data points $|D_i|$ on each client, if the data sizes are heterogeneous:

$$w_{\text{global}}^{t+1} = w_{\text{global}}^t + \frac{1}{\sum_{i \in S_t} |D_i|} \sum_{i \in S_t} |D_i| \Delta w_i^t \tag{4}$$

Once the global model is updated, the server broadcasts the updated global model w_{global}^{t+1} to all or a subset of the clients. The communication cost for broadcasting is given by:

$$C_{\text{server-client}}^t = |w_{\text{global}}^{t+1}| \tag{5}$$

The total communication cost per round t, considering both client-to-server and server-to-client communication, is the sum of both:

$$C_{\text{total}}^t = C_{\text{client-server}}^t + C_{\text{server-client}}^t \tag{6}$$

which expands to:

$$C_{\text{total}}^t = \sum_{i \in S_t} |\Delta w_i^t| + |w_{\text{global}}^{t+1}| \tag{7}$$

After receiving the global model, each client i personalizes the model further using its local data:

$$w_i^{t+1} = w_{\text{global}}^{t+1} - \eta \nabla L_i(w_{\text{global}}^{t+1}, D_i) \tag{8}$$

1.2. Motivations

Despite substantial progress in various applications, personalized federated learning (PFL) still faces several challenges. A primary concern is the increasing number of edge servers participating in federated learning (FL), which results in significant communication overheads in current PFL methods [24]. These challenges are due to multiple factors. In deep learning tasks typical of FL, model parameters can range from tens to hundreds of megabytes, and FL training convergence often requires hundreds or thousands of communication rounds. The frequent long-distance transmission, global model aggregation, and backhaul of model parameters collectively contribute to substantial communication overhead, limiting the scalability of PFL systems. To overcome these challenges, it is essential to design PFL systems that emphasize communication efficiency and scalability. Developing techniques that reduce communication overhead while ensuring effective collaboration and convergence in FL training processes is crucial [25, 26]. These motivations underscore the need for a research survey on communication architectures in PFL.

- **Emerging Research Area:** Personalized federated learning (PFL) is an emerging research area, as this field continues to evolve, there is a need to review and analyze the existing communication optimization strategies employed in PFL [4]. Our research survey provides a timely and comprehensive analysis that fills a gap in current literature by focusing specifically on communication efficiency.

- **Diverse Communication Optimization Approaches:** A wide range of communication optimization approaches and algorithms are employed in personalized federated learning (PFL), such as model compression, differential privacy, client selection, asynchronous updates, gradient compression, and model caching. A research survey can offer a thorough overview of these approaches, detailing their strengths, limitations, and applicability across various scenarios [27]. Our research survey is distinct in that it covers a diverse set of communication techniques, providing a comprehensive view of the current state of the art.

- **Key Insights for Future Developments:** A survey can provide the existing gaps and challenges in communication optimization strategies for personalized federated learning (PFL). By recognizing the limitations of current approaches, researchers can propose new techniques and avenues for future development [28, 29]. Our detailed analysis pinpoints key areas for future research, fostering the innovation of more efficient communication methods in PFL.

- **Practical Implications:** Personalized federated learning (PFL) has numerous applications in fields such as healthcare, finance, and other domains where privacy and personalization are critical. A research survey on communication optimization strategies can offer valuable insights and guidelines for practitioners and policymakers to effectively implement PFL systems and overcome potential communication challenges [30, 31, 32]. This research survey not only explores theoretical aspects but also provides practical guidelines and algorithms for implementing these methods.

1.3. Contributions

There are several surveys that provide an overview of the general concepts [33], methods, and applications of FL. Some of them specifically delve into FL from the perspectives of privacy [34], security [30] and robustness [35]. However, there is a lack of a comprehensive survey that focuses specifically on models and communication architectures in PFL. This survey aims to fill this gap in the current literature on PFL. The main objective of this paper is to provide a systematic perspective i.e. characteristics, graphical overview, algorithms used, advantages, disadvantages and challenges in models and communication architectures in PFL for the researchers. The contributions of this survey can be summarized as follows:

- **Comprehensive overview:** This research survey offers a comprehensive overview of the existing communication optimization techniques used in personalized federated learning (PFL). It catalogs various approaches, algorithms, and methods employed in the field, providing researchers and practitioners with a holistic understanding of the optimization landscape.

- **Communication Architectures:** This research survey explores communication architectures specifically designed for personalized federated learning (PFL). In PFL, communication challenges arise not only from standard FL constraints (e.g., model size, bandwidth limitations) but also due to the need for personalization on the client side. Thus, we discuss approaches such as model compression, differential privacy, client selection, asynchronous updates, gradient compression, and model caching, all in the context of ensuring efficient and scalable communication while enabling client-specific model training.

- **Identifying research gaps:** By reviewing the literature, this survey identifies research gaps and open challenges in the field of communication optimization in personalized federated learning (PFL). These gaps include unexplored optimization techniques, limited evaluation on specific data types or applications, and particular issues related to scalability, fairness, and privacy.

- **Inspiring future research:** This research survey inspires and stimulates further research in the field by highlighting promising directions and emerging trends. It proposes novel research avenues, such as hybrid optimization approaches, adaptive learning rate scheduling, compression schemes, or personalized aggregation methods.

In summary, this research survey contributes by providing a comprehensive overview, classification of existing approaches in PFL for communication optimization. It also identifies research gaps, offers guidelines, and inspires future research directions. These contributions collectively enhance the understanding, development, and application of these strategies in PFL.

1.4. Structure of this Paper

Introduction: This section presents an overview of personalized federated learning (PFL), its basic architecture, and algorithm. We discuss our motivation for conducting this research survey and outline our contributions. **Communication Models for Personalized Federated Learning:**

The main contributions of this survey include the classification and analysis of existing research on PFL communication optimization techniques. This section provides the research problem statement and and research questions. Then in section; **Communication Optimization Techniques** we provide detailed descriptions and algorithmic forms of each technique, highlighting their advantages and applications. **Challenges and Open Research Directions:** This section discusses potential future directions for research in PFL and identifies key challenges and limitations that need to be addressed. **Conclusion:** This section summarizes the key findings and contributions of the paper and discusses the implications of the study for future research and practical applications of PFL.

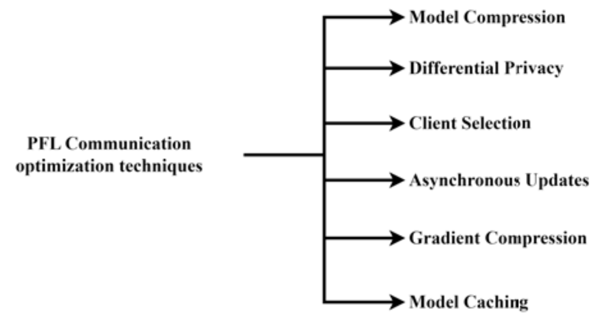

Figure 3: Classification of Communication Optimization Strategies in PFL

2. Communication Optimization Models for Personalized Federated Learning

As referenced in the Introduction section, there is still a gap of comprehensive reviews that assess communication models within the context of personalized federated learning (PFL). In response to this gap, we conducted a systematic mapping review (SMR) with the aim of delineating the prevailing research challenges and identifying discernible voids, providing a comprehensive overview of the research landscape in this domain. Owing to the limited depth of evaluation typically assigned to articles in practical application of the SMR protocol, allowing for the potential inclusion of a broader spectrum of articles, our initial approach involved using the SMR methodology to elucidate the interaction between the body of literature and pertinent categories, ultimately revealing research gaps.

2.1. Research Questions

Following are the research questions we used for our survey:

- RQ1: What are the communication optimization models in PFL?
- RQ2: What issues are addressed by these models and what mechanisms are employed?
- RQ3: On the basis research survey and gaps identified, which areas should be focused on?

The schematic representation of our review's methodological framework and proposed categorical schema is depicted in Figure 3.

3. Communication Optimization Techniques

In this section, we classify research work related to communication architectures in personalized federated learning (PFL), as shown in Figure 3. In PFL, each client has its own locally collected data, and the goal is to collaboratively train a global model that can make accurate predictions for each client's specific needs. Communication optimization techniques play a crucial role in improving the efficiency and effectiveness of PFL. Following sections discuss the techniques used to optimize communication in PFL.

3.1. Model compression

A technique which plays a crucial role in PFL by reducing the size of the models and improving communication efficiency between the central server and participating clients. By compressing the model, the amount of information transmitted during communication is reduced, resulting in lower latency and bandwidth requirements. Figure 4 shows model compression and Algorithm 1 presents the working of model compression in PFL.

Algorithm 1 Model Compression in PFL

Input(s): Global model M, Compression rate C
Output(s): Compressed global model $M_{\text{compressed}}$
Algorithm:

```
Initialization
while not converged do
    Client Selection
    for each selected client do
        Receive M from server
        Perform local model training using client data
        Compute local model update ΔM_local
    end for
    Aggregate local model updates
    Compress aggregated model update ΔM_aggregated
    Update global model M using ΔM_aggregated
end while
Compressed global model M to M_compressed with rate C
```

Let, M be the global model, M_t global model at iteration t, $\Delta M_{\text{local}}^{k}$ is the local model update for client k, $\Delta M_{\text{aggregated}}$ is the aggregated model update across clients, $\Delta M_{\text{compressed}}$ be the compressed aggregated update, $\mathcal{L}(M; \mathcal{D}_k)$ the loss function for client k on data $\mathcal{D}_k$, λ be the regularization parameter, $\mathcal{S}_t$ be the set of selected clients at iteration t, $C(\cdot, C)$ is compression function with rate C and η be the learning rate.

Local Model Update (for client k using its data $\mathcal{D}_k$):

$$\Delta M_{\text{local}}^{k} = \arg\min_{M} \left(\mathcal{L}(M; \mathcal{D}_k) + \frac{\lambda}{2} \|M - M_t\|^2 \right) \quad (9)$$

Aggregation of Local Updates:

$$\Delta M_{\text{aggregated}} = \frac{1}{|S_t|} \sum_{k \in S_t} \Delta M^k_{\text{local}} \tag{2}$$

Compression of Aggregated Update (compression rate C):

$$\Delta M_{\text{compressed}} = C(\Delta M_{\text{aggregated}}, C) \tag{3}$$

Global Model Update:

$$M_{t+1} = M_t + \eta \cdot \Delta M_{\text{compressed}} \tag{4}$$

Final Model Compression:

$$M_{\text{compressed}} = C(M, C) \tag{5}$$

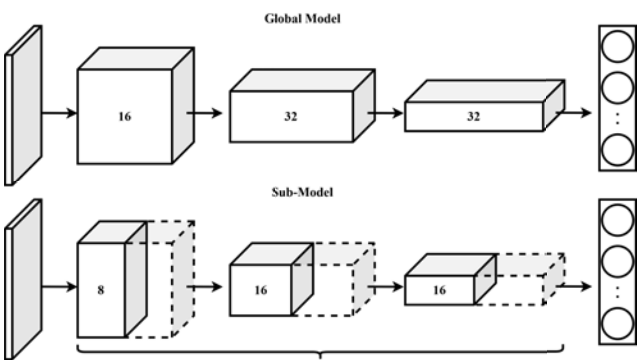

Figure 4: An Example of Model Compression

Snowball [36] is a novel approach to compression-based training within the realm of FL. The primary objective of Snowball is to significantly reduce the energy consumption associated with parameter transmission while maintaining the global model's accuracy. This is achieved through a coarse-to-fine grained compression mechanism, which reduces the volume of data that needs to be communicated between clients and server. By implementing such a mechanism, Snowball effectively optimizes the use of computational and communication resources. The framework's ability to maintain accuracy while reducing communication overhead exemplifies its potential in energy constrained environments such as mobile or edge computing scenarios. Personalized Sparsification with Element-wise Aggregation for Federated Learning (FedPSE) [37] focuses on enhancing FL performance on non-independent and identically distributed (non-IID) datasets through an innovative element wise aggregation method. Traditional FL systems often struggle with heterogeneous data distributions, which can degrade model performance. FedPSE addresses this challenge by performing aggregation at a more granular level, ensuring that the communication costs are minimized while the model's performance on non-IID data is enhanced. This element-wise approach reduces the data transmission required and also ensures that the model updates are more representative of the diverse data distributions across different clients.

Table 1 shows the summary of contributions in PFL communication optimization techniques using Model Compression with respect to their main ideas and applications. Federated Learning for Multiple Personalized Tasks (FedMPT) [41] is another technique, designed to optimize personalized models for clients engaged in multiple tasks, balancing the accuracy of each task with the efficiency of communication. In FL, handling multiple tasks simultaneously while ensuring efficient communication is a significant challenge. FedMPT tackles this by using regularization techniques to balance the model accuracy for each task against the overall communication costs. This personalized approach allows clients to achieve high accuracy across various tasks without adding excessive communication overhead. By optimizing the trade-off between accuracy and communication, FedMPT provides a solution which enhances the individual task performance and the overall efficiency of the system.

Quantized Personalization via Distillation (QuPed) [44] offers a dual-layer personalization strategy aimed at addressing both data heterogeneity and resource diversity among clients in FL. This technique integrates model compression with resource-efficient training methods to adapt models to the unique data characteristics and computational capabilities of each client. By focusing on these two critical aspects, QuPed ensures that the FL process is both efficient and scalable. The model compression aspect reduces the amount of data that needs to be transmitted, while the resource-efficient training methods ensure that even clients with limited computational resources can participate effectively. This dual-layer approach makes QuPed a versatile and robust solution for diverse FL environments.

Federated Learning for Non-intrusive load monitoring (FedNILM) [45] addresses the challenge of data privacy in F through the use of efficient model compression techniques like filter pruning and multi-task learning. Additionally, it employs unsupervised transfer learning to build personalized models. This method is particularly effective for applications such as energy disaggregation, where accurate and personalized model predictions are necessary without compromising user privacy. By combining these techniques, FedNILM not only preserves privacy but also reduces the communication overhead, making the learning process more efficient. The integration of filter pruning and multi-task learning ensures that the model remains lightweight and capable of handling multiple tasks simultaneously, further enhancing its applicability in privacy-sensitive environments.

FedSUMM [49] introduces a dynamic gradient adapter designed to provide locally tailored parameters for individualized models at the client level. This technique accelerates model convergence by dynamically adjusting gradients to better fit local data distributions. By tailoring the gradients to the specific characteristics of the local data, FedSUMM ensures that each client's model updates are more effective, leading to faster overall convergence. This approach not only

Table 1
Summary of contributions in PFL Communication Optimization Techniques using Model Compression

Technique(s)	Main Idea/Contribution	Dataset(s)	Clients	Communication Rounds	Accuracy (%)
Snowball [36]	A novel compression-based training framework for FL, to reduce the energy consumption of parameter transmission.	Fashion-MNIST [38] CIFAR-10 [39]	60	500	91.65 83.56
FedPSE [37]	An element-wise aggregation method to enhance performance on non-IID datasets while reducing communication costs.	Fashion-MNIST IMDB [40]	100	300	87.72 99.75
FedMPT [41]	Optimizing personalized models for clients with multiple tasks, balancing both the accuracy of each task and the communication efficiency.	Multi-MNIST [42] CelebA [43]	75	10 and 5	89.1 90.8
QuPed [44]	Dual-layer personalization for both data heterogeneity and resource diversity.	CIFAR-10 FEMNIST	50, 66	250	96.64 74.64
FedNILM [45]	Realized data privacy-preserving through FL, efficient model compression via filter pruning and multi-task learning, and personalized model building by unsupervised transfer learning.	REFIT [46] UK-DALE [47] REDD [48]	6	-	-
FedSUMM [49]	A dynamic gradient adapter designed to offer locally tailored parameters for the individualized model at the local level.	CSL [50] CLTS [51] LCSTS [52] THUCNEWS [53] EDUCATION [49]	50	200	48.0 49.6 37.7 33.8 49.5
L2GD [54]	Bidirectional compression mechanism to further reduce the communication bottleneck between the local devices and the server.	CIFAR-10	10	20000	82.3 (MobileNet)
SoteraFL [55]	The combination of general compression operators and local differential privacy. Applies compression directly to differentially private stochastic gradient descent and identifies its limitations.	MNIST	-	1000	85 (SGD)
Dis-PFL [56]	A decentralized sparse training approach, termed Dis-PFL, is employed in which each local model maintains a constant number of active parameters throughout the local training and peer-to-peer communication phases.	CIFAR-10 CIFAR-100 Tiny-Imagenet	100	500 500 300	85.70 53.48 16.95
LSGD-PFL, ACD-PFL, and ASVRCD-PFL [57]	Universal optimizers, rendering the design of task-specific optimizers unnecessary in many instances.	MNIST KMINIST [58] Fashion-MINST CIFAR-10	20	1000	-
CMP-Fed [59]	Combines communication compression with privacy protection, achieving agent-level differential privacy while maintaining high model accuracy.	Fashion-MNIST	100	180	77.93
iECS-FL [60]	Compressed sensing for model compression, iterative reconstruction and retraining.	MNIST Fashion-MNIST SVHN CIFAR-10	100	50 to 500	97.75 86.51 85.70 86.01

improves the efficiency of the F process but also enhances the performance of the individualized models. The ability to dynamically adapt gradients makes FedSUMM particularly suitable for environments where data distributions vary significantly across clients.

In [54] authors introduced the Loopless Gradient Descen (L2GD) algorithm for PFL, addressing data heterogeneity across edge devices. L2GD balances global and local models and reduces communication overhead with bidirectional compression. Experimental results validate its efficiency in reducing communication while maintaining convergence rates similar to those of traditional SGD. Traditional centralized approaches can be vulnerable and have high communication pressure. In [55], authors introduced the SoteriaFL framework, which addresses the challenges of large-scale machine learning in bandwidth intensive environments. It combines communication efficient FL with communication compression and privacy preservation at the client level. The framework leverages general compression operators and local differential privacy, offering better communication efficiency without compromising privacy or utility compared to other private FL algorithms. The comprehensive analysis demonstrates the advantages of SoteriaFL in terms of communication complexity, privacy, and utility. In summary, SoteriaFL provides a unified solution for communication efficient FL with compression and client level privacy preservation. Decentralized sparse training based Personnalized Federated Learning (Dis-PFL) a method proposed by [56], utilized a decentralized communication protocol and personalized sparse masks to customize local models. Reduced communication and computation costs by maintaining a fixed number of active parameters throughout the training and communication process.

In [57], the authors introduced three universal optimizers: Local Stochastic Gradient Descent for Personalized FL (LSGD-PFL), Accelerated Block Coordinate Descent for Personalized FL (ACD-PFL), and Accelerated Stochastic Variance Reduced Coordinate Descent for Personalized FL (ASVRCD-PFL). These optimizers are designed to be universally applicable, eliminating the need for task-specific optimizers in many instances. They provide the best-known communication and computation guarantees, making them highly efficient for PFL scenarios. In [59], authors introduced CMP-Fed, a new FL scheme that addresses challenges of privacy and communication efficiency. Existing approaches treat privacy and communication independently, leading to accuracy degradation and limited exploration of privacy impact on compression techniques. CMP-Fed combines communication compression with privacy protection, achieving agent-level differential privacy while maintaining high model accuracy. The key component is the compressed model perturbation (CMP) approach, which compresses shared model updates and applies random noise perturbation. Experimental results on the Fashion-MNIST dataset demonstrate CMP-Fed's superiority in model accuracy compared to existing differentially private FL schemes, while enjoying the communication benefits of model compression. In [60], authors proposed an enhanced compressed sensing FL algorithm (iECS-FL) to improve communication efficiency in FL. By leveraging compressed sensing to compress local network models trained on clients, it significantly reduces the communication bandwidth required. Through iterative reconstruction and retraining method it also improves the learning performance of compressed models.

In summary methods like snowball and FedPSE use compression to lower transmission energy and aggregation costs on datasets such as Fashion-MNIST and CIFAR-10, achieving high accuracy across numerous communication rounds. Techniques like FedMPT and QuPed focus on multi-task personalization and resource diversity, while FedNILM and FedSUMM introduce privacy-preserving, gradient-adaptive methods for tailored local models. Solutions like L2GD and SoteraFL incorporate bidirectional compression and differential privacy, addressing communication bottlenecks for large datasets. Dis-PFL and iECS-FL enhance performance with sparse decentralized training and compressed sensing. Overall, these approaches showcase the diversity in compression methods that balance data heterogeneity, client constraints, and communication efficiency across varied datasets.

3.2. Differential Privacy

Differential privacy (DP) techniques involve adding noise to model updates before sharing them. This allows updates to be shared less frequently without compromising privacy. The reduced update frequency directly decreases the amount of data transmitted [1, 4]. DP can be combined with techniques such as secure aggregation, where updates from multiple clients are aggregated and noise is added to the combined result. This reduces the number of communication rounds needed, optimizing the communication process [24, 61], but these techniques have implications for communication efficiency, as they involve additional computations and potentially increased communication overhead. It involves adding carefully calibrated noise or perturbations to the computations performed during the FL process. Figure 5 illustrates a shuffler-based differential privacy approach in FL, where the shuffler significantly enhances privacy by anonymizing client updates before they reach the server. While this architecture can reduce the frequency of communication rounds and minimize direct communication between clients and the server, potential overhead may arise from the need to manage the shuffling process and maintain synchronization across clients. These factors can contribute to communication complexity, particularly in asynchronous settings. Nonetheless, the method remains effective in reducing the total communication volume due to its privacy-preserving design. Algorithm 2 describing the differential privacy technique in PFL.

Algorithm 2 Differential Privacy in PFL

Input(s): Global model M, Privacy parameter ϵ
Output(s): Privatized global model $M_{\text{privatized}}$
Algorithm:

```
Initialization
while not converged do
    Client Selection
    for each selected client do
        Receive M from server
        Perform local model training using client data
        Compute local model update ΔM_local
        Apply noise to ΔM_local with Laplace or Gaussian method ε
    end for
    Aggregate privatized local model updates
    Update global model M using aggregated updates
end while
Privatize final global model M to M_privatized
```

1. Local Model Update is similar to equation 9
2. Differential Privacy with Noise Addition:

$$\Delta M_{\text{local}}^{k,\text{priv}} = \Delta M_{\text{local}}^{k} + \mathcal{N}(\sigma^2) \quad \text{or} \quad \Delta M_{\text{local}}^{k} + \mathcal{L}(\epsilon) \tag{10}$$

3. Aggregation of Privatized Local Updates:

$$\Delta M_{\text{aggregated}} = \frac{1}{|S_t|} \sum_{k \in S_t} \Delta M_{\text{local}}^{k,\text{priv}} \tag{11}$$

4. Global Model Update:

$$M_{t+1} = M_t + \eta \cdot \Delta M_{\text{aggregated}} \tag{12}$$

5. Final Model Privatization:

$$M_{\text{privatized}} = M + \mathcal{N}(\sigma^2) \quad \text{or} \quad M + \mathcal{L}(\epsilon) \tag{13}$$

Where, M: Global model, M_t: Global model at iteration t, $\Delta M^k_{\text{local}}$: Local model update for client k, $\Delta M^{k,\text{priv}}_{\text{local}}$: Privatized local model update for client k, $\Delta M_{\text{aggregated}}$: Aggregated model update across clients, $\mathcal{L}(M; \mathcal{D}_k)$: Loss function for client k on data $\mathcal{D}_k$, λ: Regularization parameter, $\mathcal{S}_t$: Set of selected clients at iteration t, $\mathcal{N}(\sigma^2)$: Gaussian noise with variance σ^2, $\mathcal{L}(\epsilon)$: Laplace noise with privacy parameter ϵ, η: Learning rate, ϵ: Privacy parameter, $M_{\text{privatized}}$: Final privatized global model.

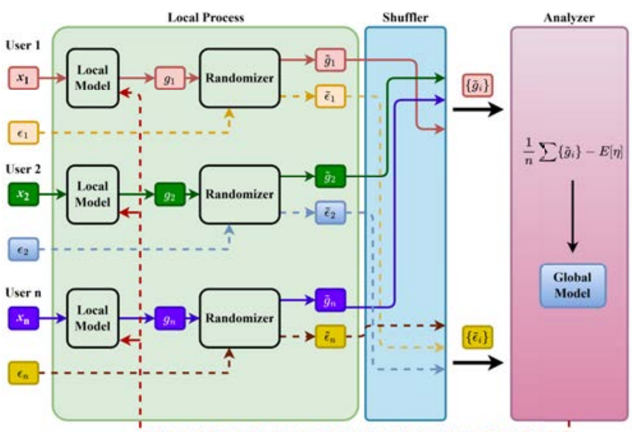

Figure 5: Personalized local differential privacy in Federated Learning [62]

The addition of noise can facilitate model compression techniques (e.g., quantization or sparsification) since noise masks finer details, making it easier to compress updates. This reduces the size of the data that is transmitted [63]. DP enables selective sharing of critical information while adding noise to less important details. This selective approach ensures that only necessary updates are communicated, thus optimizing communication. By incorporating these aspects, DP enhances both privacy and communication efficiency in FL systems. These optimizations are particularly relevant in PFL, where frequent updates and communication costs are significant concerns.

CMP-Fed [59], technique compresses model updates before adding random noise, thereby reducing the data sent during each communication round. This strategy effectively addresses communication overhead while maintaining model accuracy. Compression also leads to agent-level differential privacy, ensuring privacy without sacrificing performance.

In [62], authors proposed privacy Amplification framework for PErsonalized private federated learning with Shuffle model (APES), a comprehensive framework that enhances privacy in FL by leveraging the privacy amplification effect of the shuffle model while respecting personalized local privacy. APES quantifies the contributions of each user to central privacy and introduces neighbor divergence and clip-laplace mechanism to measure their perturbation's ability to generate "echos". The S-APES (APES with post-Sparsification technique) framework incorporates post sparsification techniques to reduce privacy loss in high-dimensional scenarios. These models introduce the concept of generating 'echoes' through data perturbation, enhancing the accuracy of the global model. While the method is geared toward improving model accuracy, it also indirectly optimizes communication by improving the efficiency of model updates.

In [64], the authors introduced DP-FedSAM, a novel FL algorithm that addresses the challenges of client-level differentially private federated learning (DPFL). Existing DPFL methods suffer performance degradation due to a sharp loss landscape and lack of robustness against weight perturbations. DP-FedSAM leverages gradient perturbation and incorporates the Sharpness Aware Minimization (SAM) optimizer to generate stable local models that are resistant to weight perturbations. This improves performance by reducing local update norms and enhancing robustness against DP noise. The paper also introduced DP-FedSAM-top_k, which further enhances the performance by employing local update sparsification. Through gradient perturbation, DP-FedSAM aims to counteract the negative impact of DP on model performance. By focusing on minimizing the trade-offs between privacy and utility, this method ensures communication efficiency by reducing the number of communication rounds needed to achieve high accuracy.

In [65], authors presented a comprehensive review of FL in the healthcare domain. FL has gained popularity in medical settings due to its ability to analyze data from individual devices while ensuring data privacy. However, FL poses unique challenges in healthcare, particularly in safeguarding sensitive patient information and complying with regulations like Health Insurance Portability and Accountability Act(HIPAA) and General Data Protection Regulation (GDPR). The paper discusses the structure of FL techniques, regulatory frameworks, and challenges associated with privacy, performance, implementation, computation, and adversaries.

The authors in [66], addressed privacy concerns in chatbot applications by introducing Fedbot, a proof-of-concept (POC) privacy-preserving chatbot. Traditional methods of training deep learning models on shared data raise privacy concerns, but FL offers a solution. Fedbot combines Deep Bidirectional Transformer models with FL algorithms to protect customer data privacy during collaborative model training. The POC demonstrates the potential of privacy-preserving chatbots in transforming the customer support industry, providing personalized and efficient customer service while complying with data privacy regulations. This method combines deep bidirectional transformers with FL algorithms, ensuring robust privacy protection over time. The integration of transformers improves both communication efficiency and accuracy in customer data protection applications.

In [67], authors presented a novel FL framework that addresses the personalized privacy needs of clients by incorporating local differential privacy (LDP). Existing studies on FL with LDP often overlook the diverse privacy requirements of individual clients. The proposed framework considers both independent and non-independent identically distributed datasets, employing tailored model perturbation methods for each scenario. Additionally, two model aggregation techniques are introduced to mitigate the impact of privacy-conscious clients on the overall federated model. Experimental evaluations on MNIST, Fashion-MNIST, and forest cover-types datasets demonstrate the effectiveness of the proposed aggregation methods in preserving personalized privacy while maintaining model accuracy. LDP focuses on managing privacy budgets for clients and ensures a better balance between privacy preservation and communication performance. It is particularly beneficial in personalized privacy-preserving scenarios, where communication can be optimized according to privacy constraints.

In [69], authors introduced Personalized Privacy-Preserving Federated Learning (PPPFL), a novel framework that addresses challenges related to privacy breaches and non-IID data in FL. PPPFL focuses on cross-silo FL and employs a stabilized variant of the Model-Agnostic Meta-Learning (MAML) algorithm. The framework utilizes synthetic data generated by Differential Private Generative Adversarial Networks (DP-GANs) for collaborative training of a global initialization. Once convergence is achieved, each client adapts the global initialization locally to their private data. Extensive experiments conducted on datasets such as MNIST, Fashion-MNIST, CIFAR-10, and CIFAR-100 demonstrate the superiority of PPPFL over various FL baselines. This technique uses a differentially private generative adversarial network (DP-GAN) to handle the challenges of non-IID data. By stabilizing the model's initialization, PPPFL reduces the communication rounds required for model convergence, enhancing both efficiency and performance.

Communication efficient and Utility aware Adaptive Gaussian Differential Privacy for Personalized Federated Learning (CUAG-PFL) [70] proposed a dynamic layer compression scheme for model gradients with improved communication efficiency and reduced loss of model utility. CUAG-PFL implements dynamic layer compression, optimizing communication by reducing the amount of data sent during each round. This technique is particularly effective at maintaining high model utility while minimizing communication overhead. The authors in [71], introduced a novel approach called blockchain-enabled distributed edge cluster for PFL (BPFL), taking advantage of blockchain and edge computing. Using blockchain technology, client privacy and security can be enhanced through the use of immutable distributed ledgers. BPFL leverages blockchain and edge computing to improve communication latency and real-time performance in PFL. Blockchain integration ensures data privacy while reducing the number of rounds required for communication, enhancing both efficiency and security. Another technique; CE-PFML [72], was proposed to reduce communication overhead by extracting lower-dimensional updates via representation learning. This not only minimizes communication requirements but also improves convergence speed and model accuracy.

Addressing FL's challenges in personalization and communication efficiency, [72] proposed a communication-efficient personalized federated meta-learning algorithm. It employs auto-encoders to reduce communication overhead, while privacy is ensured with differential privacy. In [73], authors introduced the Personalized Differential Privacy Mechanism (PDPM), a novel perturbation algorithm that enhances FL by providing personalized local differential privacy (PLDP). Traditional FL approaches often lack sufficient privacy protection, while secure FL schemes based on local differential privacy overlook personalized privacy requirements for individual clients. PDPM addresses these limitations by allowing clients to adjust privacy parameters based on the sensitivity of their data, ensuring personalized privacy protection. PDPM addresses the challenge of uniform privacy budget allocation by tailoring privacy budgets to individual clients, thus enhancing communication efficiency while maintaining high model quality.

In [84], authors proposed a PFL framework called Personalized Federated Local Differential Privacy (PFed-LDP) that aims to mitigate the accuracy loss caused by privacy-preserving techniques. The framework specifically targets IoT sensing data and incorporates LDP to ensure privacy protection. It introduces a dynamic layer sharing mechanism that separates the local model into global layers and personalized layers. LDP noise is applied to the global layers, which are then transmitted to the FL framework for aggregation. Each local client updates their model by incorporating both the local personalized layers and the aggregated global layers, enabling them to perform IoT tasks effectively. This model separates local and global model layers, reducing the amount of data that needs to be transmitted. This dynamic layer-sharing mechanism ensures that the worst-performing clients still benefit from personalized privacy while reducing overall communication costs. In [74], authors proposed a novel algorithm called Personalized Local Update Federated Learning with Optimal Aggregation (PLU-FedOA), which introduces personalized local differential privacy into the optimization of deep neural networks in horizontal FL. Their algorithm consists of two key components: PLU and FedOA. PLU allows individual clients to upload their local updates while adhering to personally selected privacy levels under differential privacy guarantees. FedOA, on the other hand, facilitates server-side aggregation by optimizing the weighting of local parameters in scenarios involving mixed

Table 2
Summary of contributions in PFL Communication Optimization Techniques using Differential Privacy

Technique(s)	Main Idea/Contribution	Dataset(s) Used	Clients	Communication Rounds	Accuracy (%)
CMP-Fed [59]	Involves compressing the shared model updates before introducing random noise during each communication round within the FL process.	Fashion-MNIST	100	180	77.93
APES, S-APES [62]	The capacity to generate 'echos' through the perturbation of individual users' data is meticulously quantified using the devised methodologies.	QMNIST	-	40	79.67
DP-FedSAM [64]	Leverages gradient perturbation to mitigate the negative impact of DP.	EMNIST CIFAR-100 CIFAR-100	10	30	84.80 57.00 21.24
Fedbot [66]	Integrates Deep Bidirectional Transformer models and FL algorithms to ensure the preservation of customer data privacy.	Twitter dataset	30	10	32
LDP [67]	The primary concept revolves around mitigating the influence of privacy-conscious clients, who opt for relatively constrained privacy budgets.	MNIST Fashion-MNIST Forest cover-types [68]	10	10 10 20	75 72 45
PPPFL [69]	A stabilized variant of the MAML algorithm to collaboratively train a global initialization from synthetic data generated by differential private generative adversarial networks (DP-GAN).	MNIST Fashion-MNIST CIFAR-10	5	30	98.26 86.87 71.33
CUAG-PFL [70]	A dynamic layer compression scheme for model gradients.	CIFAR-10 CIFAR-100	-	-	83
BPFL [71]	Combines the benefits of blockchain and edge computing.	MNIST	-	50	84
CE-PFML [72]	Incorporates representation learning to diminish communication overhead by extracting efficient and condensed local updates with lower dimensionality.	MNIST FEMNIST CIFAR-10	100	1000	96.36 85.43 48.95
PLDP-FL [73]	Addresses the challenge of uniform privacy budget allocation across all clients.	MNIST Fashion-MNIST	100	100	98.2 87.3
PLU-FedOA [74]	Designed two methodologies: PLU enables clients to upload local updates under personalized differential privacy constraints, and FedOA facilitates server-side aggregation of local parameters with optimized weighting.	MNIST	100 1000	600	94 98
PDP [75]	A personalized Differential Privacy (PDP) framework that caters to each client's unique privacy requirements and combines PDP with sampling to minimize noise addition.	MNIST	10	50	-
d-privacy [76]	Incorporates a metric-based obfuscation method to preserve the original data's topological distribution.	MNIST	-	60	85
FLUP [77]	Provides user-level personalized privacy protection while maintaining high data utility.	MNIST CIFAR-10	100	50	- 55.1
UM-PFSSL [78]	A personalized semi-supervised learning paradigm that allows clients with partial or unlabeled data to enhance local data perception with helper agents.	Fashion-MNIST CIFAR-10	100	30	79 51.14
PGC-LDP [79]	Allows users to select privacy levels via federated stochastic gradient descent with local differential privacy.	MNIST CIFAR-10	200	50	- -
FedEmbed [80]	Utilizes sub-populations and personal embeddings for global model personalization.	MNIST	900	-	77
OPFL [81]	An online personalized FL framework for privacy-preserving indoor localization.	UJIndoorLoc [82] Shopping Mall [83]	10	100	-

privacy preservation. PLU enables clients to upload local updates under personalized differential privacy constraints, while FedOA improves server-side aggregation. This dual mechanism improves communication efficiency by reducing the data sent and optimizing server computations.

Table 2 shows the contributions made so far in PFL communication optimization. In [75], a personalized differential privacy (PDP) framework is introduced, catering to each client's unique privacy requirements. A novel approach is proposed, combining PDP with sampling to minimize noise addition while upholding clients' privacy needs. The paper also presents a client selection mechanism integrating a new metric score that simultaneously considers local loss and privacy demands. The PDP framework minimizes noise addition through selective sampling, leading to more efficient communication while preserving privacy. By reducing unnecessary noise, this approach ensures better communication efficiency and utility. In [76], the concept of d-privacy, a variant of local differential privacy, is presented. This technique incorporates a metric-based obfuscation method to preserve the original data's topological distribution. The approach has a dual purpose: safeguarding client data privacy and facilitating personalized model training to enhance fairness and utility within the FL framework. By leveraging the inherent attributes of d-privacy, group privacy assurances are achieved, enabling the creation of personalized models within the FL paradigm, accommodating the diverse characteristics of client communities. This method uses a topological obfuscation technique to preserve data privacy while optimizing communication in PFL. By focusing on maintaining the topological structure of the data, this method reduces communication costs and supports personalized model training.

The authors in [77], proposed a novel FL framework called Federated Learning with User-level Personalization (FLUP) that provides user-level personalized privacy protection while maintaining high data utility. The framework incorporates a personalized DP mechanism that combines a personalized sampling algorithm and Gaussian perturbation to meet each user's unique differential privacy needs. FLUP achieves user-level personalized privacy protection using a combination of personalized sampling and Gaussian perturbation. This combination ensures high utility and communication efficiency by catering to the individual privacy needs of users. UM-PFSSL [78], the semi-supervised learning approach allows clients with limited labeled data to collaborate with helper agents. This method reduces network communication while maintaining high accuracy by ensuring reliable pseudo-labels. In [78], authors introduced a personalized semi-supervised learning paradigm enabling clients with partial or unlabeled data to enhance local data perception through helper agents. An uncertainty-based data-relation metric ensured reliable pseudo-labels from selected helpers. A helper selection protocol mitigated network overload during helper search. Addressing varying privacy needs in FL participants, [79] proposed Personalized Gradient Clipping with Local Differential Privacy (PGC-LDP), a PFL approach. Users could select privacy levels via federated stochastic gradient descent with local differential privacy. A novel client-side computation algorithm and optimized server-side aggregation method were developed. This method enables clients to select their privacy levels during federated stochastic gradient descent, allowing for dynamic adaptation of communication based on privacy needs. As a result, it handles varying privacy requirements while ensuring communication efficiency. In [80], "FedEmbed" was presented as a novel PFL approach using sub-populations and personal embeddings for global model personalization. FedEmbed uses personal embeddings and sub-population strategies to personalize global models. This technique improves performance by 45% compared to baseline PFL methods, optimizing communication by reducing unnecessary exchanges.

In [81], authors introduced OPFL, designed for privacy-preserving indoor localization, OPFL combines artificial noise via DP with a focus on performance preservation, addressing communication efficiency by balancing privacy and utility in an online framework. For PFL with joint differential privacy, [85] focused on user-level privacy constraints for local and global models. They introduced coordination between local and private centralized learning, achieving improved accuracy and privacy trade-off with generically useful results. Generalization guarantees were supported by experiments on real-world and synthetic datasets.

In their research [86], authors extensively evaluated various FL and differential privacy techniques using the MIMIC-III dataset. Their analysis focused on the impact of parameters like data distribution, communication strategies, and federation approaches on model performance. The study also compared two differential privacy methods: stochastic gradient descent-based differential privacy algorithm (DP-SGD), and a sparse vector differential privacy technique (DP-SVT). The results highlighted that extreme data distribution imbalances could affect the performance of the FedAvg strategy, whereas the FedProx strategy with suitable hyperparameter tuning effectively mitigated this issue.

Differential privacy in PFL ensures the privacy of client data during model updates by adding noise to gradients or model parameters. CMP-Fed and APES minimize privacy leakage through compressed updates and noise generation, requiring around 180 and 40 communication rounds, respectively. DP-FedSAM adds gradient perturbation, achieving notable accuracies (84.8% on EMNIST, 57% on CIFAR-100) within only 30 rounds. PLDP-FL and PLU-FedOA provide adaptive privacy budgets and constrained privacy uploads over 100 to 600 rounds, ensuring client-specific adjustments. Techniques like CE-PFML and PGC-LDP employ local differential privacy and representation learning, reducing communication rounds to a maximum of 50 while preserving data utility. Finally, OPFL and FedEmbed offer unique, low-frequency updates in online and personalized applications, respectively, with up to 900 rounds for enhanced localization or embedding-based model personalization. Each approach thus emphasizes communication frequency as a trade-off with privacy and model performance. Applications range from healthcare to chatbots, leveraging strategies like dynamic layer compression, blockchain integration, and personalized privacy settings to improve communication efficiency and model performance.

3.3. Client Selection

In PFL, not all clients participate in each round of model updates. The client selection techniques aim to select a subset of clients for participation in a round based on various criteria, such as heterogeneity of the data, computational resources, and communication capabilities as shown in Figure 6. By carefully selecting clients, communication overhead can be reduced without compromising the quality of the global model. Using client selection techniques that consider data heterogeneity, computational resources, communication constraints, performance tracking, and adaptability, PFL can optimize communication efficiency. The selection of an appropriate subset of clients for each round of communication reduces communication overhead while ensuring the effectiveness and accuracy of the global model. The algorithm 3 shows the simple implementation of client selection in PFL.

Algorithm 3 Client Selection in PFL

Input(s): List of clients C, Client data heterogeneity, Computational resources, Communication constraints
Output(s): Selected client subset S
Algorithm:

```
Initialization
Compute client scores based on data heterogeneity, computational
  resources, and communication constraints
Sort clients in descending order of scores
S ← Empty list
for each client c in C do
    if fits communication constraints and computational resources then
        Append c to S
    end if
    if termination condition met then
        break
    end if
end for
```

Let C represent the complete list of clients, and S denote the subset of selected clients. Each client c in the list C is assigned a score, denoted by score(c), which is calculated based on factors such as data heterogeneity, computational resources, and communication constraints. This score calculation is performed through a function $f(\cdot)$, which combines these factors to yield a single score value for each client. The list of clients is then sorted by these scores in descending order, resulting in C_{sorted}, which prioritizes clients with higher scores. To determine the final selected subset S, each client c in C_{sorted} is evaluated to check if it meets specific communication constraints and has sufficient computational resources to participate in FL. Only clients that satisfy both the communication and computational requirements are added to S.

1. Client Score Calculation:

$$\begin{aligned} \text{score}(c) = f(&\text{data heterogeneity}(c), \\ &\text{computational resources}(c), \\ &\text{communication constraints}(c)) \end{aligned} \tag{14}$$

2. Client Sorting:

$$C_{\text{sorted}} = \text{sort}(C, \text{by descending score}) \tag{15}$$

3. Client Selection with Constraints:

$$\begin{aligned} S = \{c \in C_{\text{sorted}} \mid &\text{ communication constraints}(c) \text{ and} \\ &\text{computational resources}(c) \text{ are satisfied}\} \end{aligned} \tag{16}$$

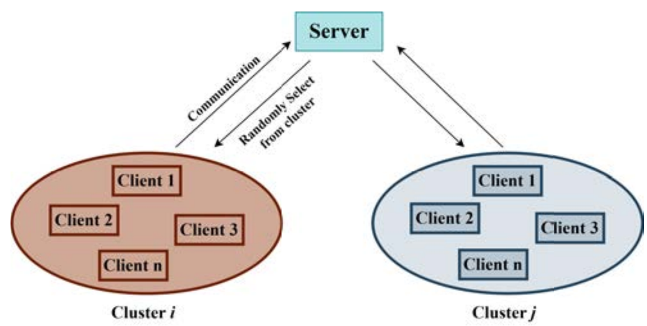

Figure 6: Clients Selection [87].

Table 3 summarizes the contributions made in PFL for communication optimization using clients selection. In [90], Han et al. presented a confidence-based similarity-aware personalized FL algorithm (FedCS), which combines public average confidence (PAC) measure, a client grouping strategy with dynamic sampling (CGDS), and a sequential aggregated weight (SAW) strategy to address the challenges of global model convergence and personalization in FL under non-identically distributed (non-IID) data. These methods collectively improve the convergence of the global model and improve personalization by effectively selecting and aggregating client updates.

A Communication efficient and Fair personalized Federated Sequential Recommendation (CF-FedSR) implements a communication efficient scheme that employs adaptive client selection and clustering based sampling to accelerate the training process. This technique aids clients in making personalized recommendations and enhances recommendation performance through local fine-tuning and model adaptation, thereby optimizing the balance between communication efficiency and model accuracy [87]. Adaptive quantization in device selection strategy (AQUILA) integrates a sophisticated device selection method that prioritizes the quality and usefulness of device updates. This method enables precise device selection by storing the exact global model on devices, reducing model deviation, and limiting the need for hyperparameter adjustments, leading to a more stable and effective training process [93]. A federated dynamic client

Table 3
Summary of contributions in PFL Communication Optimization Techniques using Client Selection

Technique(s)	Main Idea/Contribution	Dataset(s) Used	Clients	Communication Rounds	Accuracy
CF-FedSR [87]	A communication-efficient scheme that employs adaptive client selection and clustering-based sampling to accelerate the training process.	Amazon [88] Wikipedia [89]	10	200	42.23 87.22
FedCS [90]	Introduces PAC measure, client grouping with dynamic sampling, and a sequential aggregated weight strategy.	CIFAR10 OrganS-MNIST [91] COVID [92]	20	300	87.60 86.45 91.19
AQUILA [93]	Integrates a device selection method prioritizing quality and usefulness of updates, with exact global model storage for precision.	CIFAR-10 CIFAR-100 WikiText-2 [94]	100	200 1000 1000	91.3 70.12 -
FedSDR [95]	Clusters clients by computational efficiency, targeting fair selection in edge computing.	MNIST-Fed CIFAR-10-Fed	100	200 300	94.51 65.44
pFedCAS [96]	Control unit adapts model sparsity for privacy protection, with reduced communication costs.	HAM10000	-	400	92.53
CFFR [97]	Fair-aware model aggregation that adjusts for performance and data distribution disparities.	MovieLens 100K Movielens 1M [98]	200	50	-
FedeRiCo [99]	Allows clients to learn from others based on optimal fit for local data.	CIFAR-10 CIFAR-100 Office Home2 [100]	8	150 150 400	78.22 41.41 93.76
FedRec++ [101]	Lossless federated recommendation with denoising client allocation for privacy.	MovieLens	943	100	-
FCFL, MAFL [102]	Full-stack learning system tailored for wearable computers, enhancing communication efficiency and personalization.	FEMNIST Google Speech	3600 2600	350 800	82.33 65.89

selection method based on data representativity (FedSDR) addresses the issue of unfair federated client selection in edge computing by clustering clients into groups based on their local computational efficiency. This approach ensures a fairer distribution of computational tasks and resources, thereby improving overall system fairness and efficiency [95].

Personalized FL framework based on Communication quality and Adaptive Sparsification (pFedCAS) enhances privacy protection and training efficiency by introducing a control unit that adjusts the sparsity of the local model adaptively, and a selection unit that selects suitable clients for parameter updates during global model aggregation. This method achieves a 15% improvement in training accuracy and a 30% reduction in training costs, demonstrating robustness to non-IID data and effective communication optimization [96]. In [97], authors proposed a communication-efficient and fair PFL approach (CFFR). CFFR used adaptive client group selection to personalize models while accelerating the training process. The authors proposed a fair-aware model aggregation algorithm that adaptively captures performance and data imbalance among different clients to address the unfairness problem.

In [99], authors proposed a model; PFL with the Right Collaborators (FedeRiCo), a decentralized framework that allows each client to determine the optimal extent of learning from other clients based on their local data distribution. FedeRiCo employs an expectation-maximization algorithm to estimate the utility of other participants' models on each client's data, enabling the selection of appropriate collaborators for learning. Notably, FedeRiCo is the only approach that consistently surpasses the performance achieved by training with local data alone. In [103], the authors introduced the CF-FedSR (communication efficient and fair personalized federated sequential recommendation) algorithm. This approach incorporates adaptive client selection and clustering-based sampling to enhance the efficiency of communication during the training process. To tackle fairness concerns, a model aggregation algorithm was designed to account for imbalances in data and performance among diverse clients. The personalization module integrates local fine-tuning and model adaptation, aiding clients in generating personalized recommendations and elevating the overall recommendation performance.

In [101], authors proposed a novel lossless federated recommendation method, FedRec++, which assigns denoising clients to eliminate noise in a privacy-preserving manner. The authors analyzed FedRec++ in terms of security, losslessness, and generality compared to existing works. Extensive experiments demonstrate the effectiveness of FedRec++ in providing accurate and privacy-preserving recommendations with minimal additional communication cost. In [102], authors proposed Fair and Communication-efficient Federated Learning (FCFL), a full-stack learning system specifically designed for wearable computers that improves SOTA performance in communication efficiency, fairness, personalization, and user experience. The authors introduced a technique named ThrowRightAway (TRA) to loosen the network capacity constraints, enabling clients with poor networks to participate and improve representation while guaranteeing fairness. They also propose Movement Aware Federated Learning (MAFL) to aggregate only the model updates with top contributions for communication efficiency.

Client selection in PFL involves choosing a subset of clients for model updates based on criteria such as data heterogeneity, computational resources, and communication capabilities, aiming to optimize communication efficiency without compromising model quality. CF-FedSR uses adaptive client clustering to reduce communication rounds, achieving accuracies like 87.22% on Wikipedia with 200 rounds. FedCS introduces PAC measures and dynamic sampling, showing competitive accuracy across datasets with 300 rounds. AQUILA prioritizes high-quality device updates, with exact model storage achieving 91.3% accuracy on CIFAR-10 in 200 rounds. Techniques like FedSDR and CFFR focus on fair client selection, considering computational efficiency and data distribution disparities. pFedCAS adapts model sparsity to optimize privacy while reducing communication to 400 rounds. FedeRiCo and FedRec++ emphasize personalized recommendations through optimal client matching, with FedRec++ achieving high communication efficiency even with a large client base. Lastly, FCFL and MAFL offer full-stack solutions for wearable devices, handling up to 800 rounds for effective personalization. Each approach balances communication frequency, accuracy, and data privacy, illustrating different methods for optimizing client selection in FL. These client selection strategies collectively improve communication efficiency, model accuracy, and fairness in PFL.

3.4. Asynchronous Updates

Asynchronous updates allow clients to transmit their updates independently, reducing the waiting time and enabling clients with faster computation capabilities to proceed without waiting for slower clients. Algorithm 4 shows the implementation of asynchronous updates in PFL which can significantly improve the communication efficiency in PFL.

Algorithm 4 Asynchronous Updates in PFL

Input(s): List of clients C, Communication and computational resources
Output(s): Global model update U_{global}
Algorithm:

```
Initialization
U_global ← Empty update
for each client c in C do
    if fits communication and computational resources then
        Perform local model training on client c
        Compute local model update U_c
        Transmit U_c to server asynchronously
    end if
end for
Receive and aggregate updates from clients asynchronously
```

1. **Global Model Initialization** Define the initial global model as:

$$M_{\text{global}}^{(0)} = \text{initialize model} \tag{17}$$

2. **Local Model Update on Client** c For each selected client c that meets the communication and computational resource constraints, the local update U_c is computed based on local data D_c:

$$U_c = \text{train}(M_{\text{global}}^{(t)}, D_c) \tag{18}$$

3. **Asynchronous Update Transmission** Each client c asynchronously transmits the local update U_c to the server.
4. **Global Model Aggregation** The server asynchronously aggregates updates received from clients. Let $U_{\text{global}}^{(t+1)}$ be the updated global model after aggregation:

$$U_{\text{global}}^{(t+1)} = U_{\text{global}}^{(t)} + \frac{1}{|C|} \sum_{c \in S_t} U_c \tag{19}$$

where S_t is the set of clients that sent updates in iteration t.

5. **Output of Global Model** The final aggregated update for the global model, $U_{\text{global}}^{\text{final}}$, is obtained after the desired number of iterations or convergence criterion is met.

Personalized Moreau Envelopes-based Asynchronous Federated Learning (APFedMe) leverages the Moreau Envelopes technique to address optimization challenges by utilizing asynchronous weight updates to enhance communication efficiency and handle data heterogeneity. The approach creates a personalized learning environment by combining these elements, resulting in improved convergence speed and communication efficiency in FL settings [105]. Resource efficient Federated Recommender System (ReFRS) proposes a lightweight, self-supervised local model based on the variational autoencoder to capture users' temporal preferences from sequences of interacted items. This method excels in both accuracy and scalability, making it a potent solution for personalized recommendation systems where learning from temporal data is crucial [106].

In [108], authors addressed the challenge of conducting FL on demand in heterogeneous edge devices with varying resource constraints. The proposed solution is a cost-adjustable FL framework called AnycostFL, which enables efficient local updates on various edge devices with different efficiency constraints. The framework incorporates model shrinking to support local model training with adjustable computation cost, gradient compression for dynamic parameter transmission with varying communication overhead, and enhanced element-wise parameter aggregation to enhance model performance. An optimization design is proposed to minimize global training loss while satisfying personalized latency and energy constraints. In

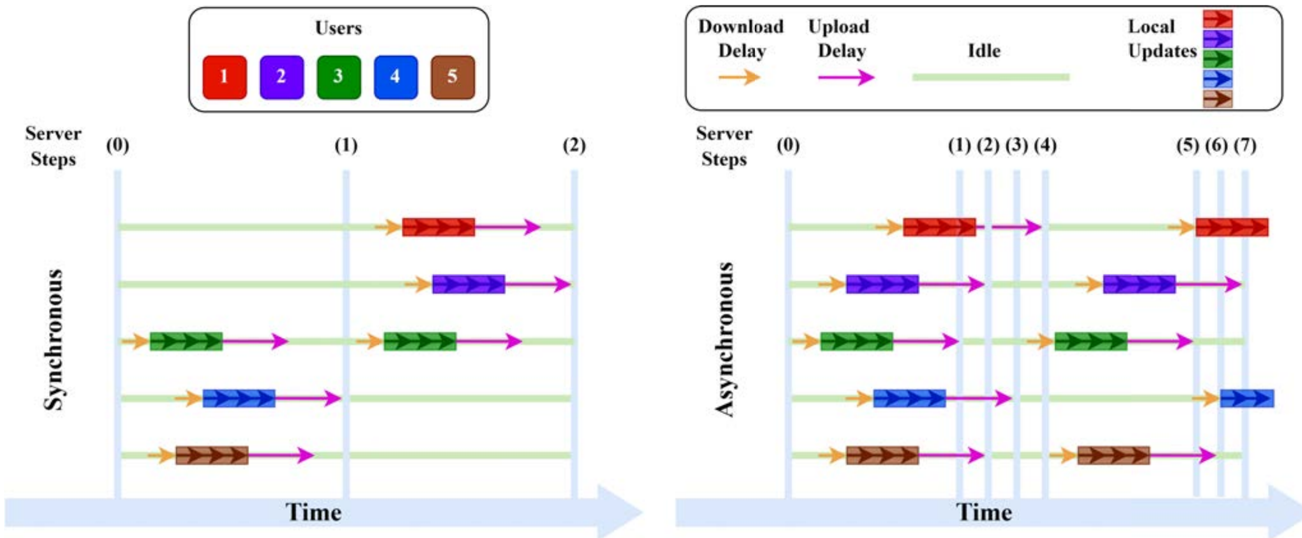

Figure 7: Overview of Asynchronous Updates [104]

[109], authors investigated the problem of PFL with asynchronous updates, where each client aims to achieve a personalized model that outperforms both local and global models. They proposed; personalized asynchronous federated learning (PersA-FL) considering two optimization frameworks, namely Model-Agnostic Meta-Learning (MAML) and Moreau Envelope (ME), for achieving personalization. MAML involves fine-tuning a joint model for each client, while ME utilizes a bi-level optimization problem with regularized losses to enforce personalization. The main focus is on improving the scalability of PFL by relaxing the assumption of synchronous communication. Furthermore, they extend the class of functions considered by removing the requirement of boundedness on the gradient norm.

To address the privacy concerns, in [110] proposed an FL based deep learning model for identifying household characteristics. The proposed hybrid model combines convolutional neural networks (CNNs) and long short-term neural networks (LSTMs) to effectively learn spatial-temporal features from load profiles. The model is implemented in a decentralized manner using the FL framework, which enables collaborative model training without sharing raw data. To improve the efficiency and accuracy of the training process, the study introduces an asynchronous stochastic gradient descent with delay compensation method to update the global model parameters. This approach enhances the training speed by allowing clients to perform local updates asynchronously while compensating for the potential delays in parameter synchronization.

In [104], authors introduced, a novel semi-asynchronous FL framework; FedSEA (Figure 7), tailored for extremely heterogeneous devices. Theoretical analysis reveals that the accuracy drop in semi-asynchronous FL (SAFL) is caused by unbalanced aggregation frequencies. Building on this insight, a training configuration scheduler is designed to balance the aggregation frequency of devices, leading to improved accuracy. To enhance efficiency in realistic scenarios where devices exhibit dynamic on-device resource availability, a scheduler is proposed that predicts the arrival time of local updates from devices and adjusts the synchronization time point accordingly.

In [112], authors introduced An Adaptive Communication Efficient Asynchronous Framework (FedACA), which is an asynchronous FL approach that incorporates feedback loops at two levels. FedACA includes a self-adjusting local training step with active participant selection to expedite the convergence of the global model. It also utilizes an adaptive uploading policy that reduces communication overhead by leveraging model similarity and L2-norm differences between current and previous local gradients. Moreover, contrastive learning is employed to regulate local training and measure model similarity in the uploading policy, thereby addressing data heterogeneity. To address the issue of staleness effects caused by asynchrony, in [113], authors proposed an optimized solution called hyper-mix Async-DFL (Asynchronous Dynamic Federated Learning). This solution incorporates a hyper parameter to mitigate the impact of staleness. Through experimental evaluations, the feasibility of these approaches is demonstrated and it is shown that the hyper-mix solution outperforms the underlying Async-DFL algorithm.

In [114], authors introduced Federated Learning with Concept Drift (FedConD), a novel approach for detecting and handling concept drift on local devices in asynchronous FL while minimizing its impact on model performance. This approach utilizes an adaptive mechanism to detect drift based on historical performance of local models. Adjusts the regularization parameter of the objective function on each device to adapt to the detected drift. Furthermore, a communication strategy is designed on the server side to

Table 4
Summary of contributions in PFL Communication Optimization Techniques using Asynchronous Updates

Technique(s)	Main Idea/Contribution	Dataset(s) Used	Clients	Communication Rounds	Accuracy
APFedMe [105]	Combines Moreau Envelopes for optimization with asynchronous updates, enhancing communication efficiency and handling data heterogeneity.	MNIST	20	500	94.79
ReFRS [106]	Lightweight self-supervised model using variational autoencoder to learn user preferences from interaction sequences.	MovieLens 100K Last.fm 1K [107]	1,682 1090	-	48.51 35.66
AnycostFL [108]	Allows localized updates across diverse edge devices, optimizing for device efficiency.	FMNIST CIFAR-10	60	214 372	90.32 84.91
PersA-FL [109]	Applies controlled staleness, usable with both MAML and Moreau Envelope frameworks.	MNIST CIFAR-10	30	-	86 67
FL-CNN-LSTM [110]	Integrates CNN and LSTM to capture spatial-temporal features from data profiles.	commission for energy regulation (CER) [111]	10	200	75.51
FedSEA [104]	Mixes synchronous and asynchronous updates to optimize efficiency and model accuracy.	MNIST CIFAR-10	-	2000	96.36 54.13
FedACA [112]	Dynamic asynchronous aggregation adapting to client capabilities and data distributions.	CIFAR-10 CIFAR-100 Fashion-MNIST	-	50 60 40	67.48 66.51 89.51
hyper-mix Async-DFL [113]	Combines hyperparameter tuning with asynchronous FL for flexible training efficiency.	MNIST Fashion-MNIST	10	100	88 72
FedConD [114]	Employs conditional updates, reducing communication by sending only significant updates.	Fashion-MNIST CIFAR-10 FitRec [115] Air Quality [116] ExtraSensory [117]	-	1000	0.907 0.953 0.822 0.430 0.721
ASTW_FedAVG [118]	Enhances FedAVG with controlled staleness tolerance for flexible client participation.	human activity recognition (HAR) MNIST	20	1000 200	95.9 98.1

carefully select local updates, thereby accelerating model convergence. In [118], the authors proposed an enhanced FL technique that combines an asynchronous learning strategy on client devices with a temporally weighted aggregation strategy on the server named; Asynchronous Model Update and Temporally Weighted (ASTW). The asynchronous learning strategy updates the parameters of deep layers less frequently compared to the shallow layers, taking into account the computational requirements of different layers.

In [119], authors proposed an asynchronously weight updating FL framework. The framework is designed to be efficient, reliable, and privacy-preserving while meeting the requirements of low latency and low network overhead. The proposed approach allows accurate resource allocation decisions for different 5G users without compromising their privacy or adding additional load to the network. Specifically, the proposed technique achieves a reduction in network overhead while maintaining a consistent and significantly high prediction accuracy, thus validating its advantages in terms of low latency and efficiency. In [120], authors proposed an asynchronous FL system that uses the principles of FL to train models on local data without the need to share it. Each participant trains a personalized ML model using their private data to improve leak identification. These personal models are then merged into a global model, which learns from the collective knowledge of all participants, while preserving data privacy. To optimize the performance of local models, the authors employed personalization techniques tailored to each participant's unique dataset. They also develop merging and benchmarking algorithms to effectively combine the personal models into a robust global model.

Through Table 4 it can be concluded that asynchronous updates in PFL enhance communication efficiency by reducing synchronization overhead, enabling overlapping computation and communication, facilitating scalability, mitigating the impact of stragglers, and supporting adaptive communication strategies. Techniques like APFedMe combine Moreau Envelopes with asynchronous updates, efficiently handling data heterogeneity and achieving 94.79% accuracy on MNIST with 500 communication rounds. ReFRS uses a lightweight self-supervised model with a variational autoencoder to learn user preferences, achieving accuracies of 48.51% on MovieLens 100K and 35.66% on Last.fm. Meanwhile, AnycostFL introduces localized updates to enhance efficiency across heterogeneous edge devices, achieving up to 90.32% accuracy on FMNIST with 214 rounds. PersA-FL applies controlled staleness, suitable for MAML frameworks, to achieve 86% accuracy on MNIST.

Further innovations include FedSEA, which integrates synchronous and asynchronous updates to improve efficiency and achieve a high accuracy of 96.36% on MNIST with 2000 rounds, and FedACA, which dynamically adapts aggregation based on client heterogeneity, resulting in robust accuracies on CIFAR-10 and Fashion-MNIST datasets. hyper-mix Async-DFL leverages hyperparameter tuning for efficient training, achieving 88% accuracy on MNIST in just 100 rounds. FedConD emphasizes conditional updates to reduce communication by transmitting only significant changes, demonstrating high performance across datasets such as CIFAR-10 (95.3%) and ExtraSensory (72.1%). Lastly, ASTW_FedAVG enhances FedAVG with staleness tolerance, allowing flexible client participation and achieving an impressive 98.1% accuracy on MNIST with only 200 rounds.

Each of these methods highlights different approaches to balance communication efficiency and model accuracy while accommodating diverse device capabilities and dataset requirements. Together, they represent significant steps forward in making PFL both scalable and effective across varied environments. These techniques leverage the distributed nature of FL to maximize the use of computational and communication resources while minimizing communication delays.

3.5. Gradient Compression

Instead of sending the full gradients, gradient compression techniques can be employed to transmit compressed versions of the gradients. Techniques such as sparsity and quantization, both of which are critical in reducing communication overhead in FL. Sparsity involves sending only the most important gradient updates and skipping the rest, while quantization reduces the precision of the transmitted gradients. Although these approaches function differently, they share a common goal of minimizing communication costs. Sparsity reduces the number of communicated elements, while quantization lowers the communication bandwidth per element, both contributing to efficient communication in PFL. Algorithm 5 shows the basic implementation of Gradient Compression in PFL.

Algorithm 5 Gradient Compression in PFL

Input(s): Local gradients G_{local}, Compression technique
Output(s): Compressed gradients $G_{\text{compressed}}$
Algorithm:

```
if Compression technique is Quantization then
    Perform quantization on G_local
    G_compressed ← Compressed gradients using quantization
end if
if Compression technique is Sparsification then
    Perform sparsification on G_local
    G_compressed ← Compressed gradients using sparsification
end if
if Compression technique is Randomized Quantization then
    Perform randomized quantization on G_local
    G_compressed ← Compressed gradients using randomized quantization
end if
if Compression technique is Error Feedback then
    Compute the difference between G_local and server gradients
    G_compressed ← Compressed gradients using error feedback
end if
if Compression technique is Differential Compression then
    Compute the difference between consecutive G_local
    G_compressed ← Compressed gradients using differential compression
end if
```

1. **Quantization:** If the compression technique is quantization, the local gradient G_{local} is quantized to produce the compressed gradient $G_{\text{compressed}}$ as:

$$G_{\text{compressed}} = \text{quantize}(G_{\text{local}}) \tag{20}$$

2. **Sparsification:** If the compression technique is sparsification, then G_{local} is sparsified, retaining only significant gradient values:

$$G_{\text{compressed}} = \text{sparsify}(G_{\text{local}}) \tag{21}$$

3. **Randomized Quantization:** For randomized quantization, a random quantization function is applied to G_{local}:

$$G_{\text{compressed}} = \text{random\_quantize}(G_{\text{local}}) \tag{22}$$

4. **Error Feedback:** When using error feedback, the error between the local gradient G_{local} and the server gradient G_{server} is computed and compressed:

$$G_{\text{compressed}} = G_{\text{local}} - G_{\text{server}} \tag{23}$$

5. **Differential Compression:** In differential compression, the difference between consecutive local gradients $G^{(t)}_{\text{local}}$ and $G^{(t-1)}_{\text{local}}$ is computed:

$$G_{\text{compressed}} = G^{(t)}_{\text{local}} - G^{(t-1)}_{\text{local}} \tag{24}$$

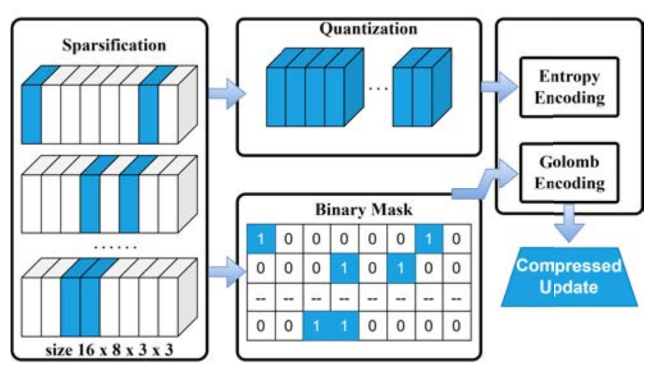

Figure 8: Overview of Gradient Compression [121].

In [122], authors proposed the ClusterGrad algorithm, which compresses gradients and considerably reduces the volume of computations communicated. Their design is based on the observation that only a small fraction of gradients have values that are far away from the origin in each round of interaction in FL. They used the K-means algorithm to identify these crucial gradients that are far away from zero. These gradient values are approximated using a novel clustering-based quantization algorithm. Furthermore, they approximate the rest of the gradients, which lie close to zero, with a single value. Wang et al. [123], tackled the constraints of PFL and introduced an economical approach that centers on weight gradients, the pivotal parameters exchanged in FL. The authors noted that these weight gradients, computed by the clients, exhibit notable sparsity and can be forecasted through straightforward bit-wise operations on a bit-stream, circumventing resource-intensive high-precision computations. Moreover, the uploaded weight gradient from each client displays a distinctive pattern influenced by its localized training data distribution.

The concept of a hub-and-spoke network topology was introduced by Kuo et al. [121], where clients interact with a central server and data compression techniques are employed to alleviate communication overhead, shown in Figure 8. Nevertheless, the challenge of unbalanced data distribution, particularly in scenarios where data on each client are non-Independent and Identically Distributed (non-IID), persists. In response, the authors introduced a novel compression compensation strategy termed Global Momentum Fusion (GMF), aimed at diminishing communication overhead while preserving model accuracy amidst non-IID data conditions. Wu et al. introduced Personalized Federated Learning via Gradient Fusion (pFedGF) in their work [124], a PFL method centered on gradient fusion. In each pFedGF round, clients maintain global gradients for collective insight and local gradients for individual distribution representation. These gradients amalgamate to update the personalized model's direction for each client.

In [125], authors introduced the first analysis of gradient compression methods using without replacement sampling. The authors propose a distributed version of the RR method with gradient compression, called Q-RR, and demonstrate how to reduce compression variance using control iterates. To better suit FL applications, they introduced a variant called Q-NASTYA that includes local computation with different local and global stepsizes. They also show how to reduce compression variance in this setting. In [126], authors introduced a novel communication efficient adaptive FL approach (FedCAMS), that provides theoretical convergence guarantees. Their approach achieves the same convergence rate as non-compressed counterparts in the nonconvex stochastic optimization setting.

In [127], the authors introduced a gradient compression strategy named FedOComp (Two-Timescale Online Gradient Compression for Over-the-Air Federated Learning). This scheme capitalizes on the inherent correlations among stochastic gradients within FL systems, enabling effective compression of high-dimensional gradients during over-the-air aggregation. The devised approach capitalizes on the structural relationships present in the gradients, resulting in a reduced impact on training convergence speed and facilitating direct over-the-air aggregation to conserve communication resources. In [128], authors proposed a novel upstream communication scheme where instead of transmitting the model update, each client generates and transmits a lightweight synthetic dataset that leads to similar performance as the real training data. The server recovers the local model update through the synthetic data and applies standard aggregation. The authors also introduced Gradient Compression via Synthetic Data in Federated Learning (FedSynth), a new algorithm for learning the synthetic data locally.

In their work [129], Jiang et al. introduced adaptive client selection and gradient compression (FedCG), an FL framework that embraces heterogeneity awareness through adaptive client selection and gradient compression. The parameter server (PS) identifies a subset of representative clients utilizing statistical heterogeneity and transmits the global model to them. Post local training, these chosen clients upload compressed model updates tailored to their capacities, subsequently aggregated by the PS. This approach substantially minimizes communication load and alleviates the straggler effect. In another study [130], Cui et al. proposed Federated Learning Method Based on Knowledge Distillation and Deep Gradient Compression (Fed-KDDGC-SGD). Clients train a teacher network, generating soft labels for a student network. During training, compressed gradient vectors are sent from the student network to the central server using a deep gradient compression algorithm. This process transmits only the top R% of the gradient values according to magnitude, reducing the communication bandwidth while maintaining training efficiency.

In their work [131], Nikoloutsopoulos et al. proposed a novel approach to PFL called; Personalized Federated Learning with Exact Gradient based Optimization (PFLEGO), that achieves exact stochastic gradient descent (SGD) minimization. Their method, built upon the FedPer neural network architecture, involves selecting random clients in optimization rounds for client-specific weight updates, ultimately enabling precise and unbiased SGD steps across the entire parameter set in a distributed manner. In a separate study [132], Melas Kyriazi et al. established a link between intrinsic dimension and gradient compressibility. This insight underpins the development of low-bandwidth FL strategies, coined as intrinsic gradient compression algorithms.

Addressing the issue of non-IID and imbalanced data distributions in FL, Yang et al. [133], devised an adaptive gradient compression algorithm. This approach tailors a unique compression rate for each client, enhancing communication efficiency while upholding model accuracy. In the context of stochastic gradient descent, adjacent rounds often exhibit high gradient correlation due to shared model learning. Exploiting this characteristic, Liang et al. [134], proposed a pragmatic gradient compression mechanism for FL. Their approach employs historical gradients for compression, employing Wyner-Ziv coding without necessitating probabilistic assumptions.

Table 5
Summary of contributions in PFL Communication Optimization Techniques using Gradient Compression

Technique(s)	Main Idea/Contribution	Dataset(s) Used	Clients	Communication Rounds	Accuracy
ClusterGrad [122]	Utilizes K-means clustering to identify important gradients, reducing computational volume.	CIFAR-10	100	300	87
BS-pFL [123]	Uses bit-streams for predicting gradient sparsity, optimizing device training cost-effectively.	MNIST CIFAR-10	10	100	97.45 86.23
GMF [121]	Compensation strategy for gradient compression.	CIFAR-10 Shakespeare	20 100	220 80	72.46 41.9
PFLFM/PPGC [135]	Uses feature fusion-based mutual-learning for communication efficiency.	MNIST CIFAR-10	10	500	- 67.7
AnycostFL [136]	Executes localized updates on edge devices, meeting diverse efficiency requirements.	FMNIST CIFAR-10	60	214 372	90.32 84.91
pFedGF [124]	Implements dual-gradient scheme per client round, for global and local insights.	MNIST Fashion-MNIST CIFAR-10.	100	120	94.57 91.21 69.59
SPFL [137]	Uses Softmax Normalized Gradient Similarity (SNGS) for personalized global model distribution.	CIFAR-10 CIFAR-100 MNIST EMNIST	-	-	67.22 24.97 98.33 87.49
Q-RR, Q-NASTYA [125]	Introduces gradient compression with without-replacement sampling.	mushrooms w8a a9a	20	5000	-
FedCAMS [126]	Communication-efficient adaptive FL with convergence guarantees.	CIFAR-10 CIFAR-100	100	500	90.2 82.4
FedOComp [127]	Correlates stochastic gradients for high-dimensional compression.	Fashion-MNIST	-	3000	90
FedSynth [128]	Clients send synthetic datasets instead of model updates; server reconstructs local updates.	FEMNIST MNIST Reddit [138]	1000 60 100	-	-
FedCG [129]	Combines adaptive client selection and gradient compression to reduce load and mitigate stragglers.	MNIST CIFAR-10 CIFAR-100 Tiny-ImageNet	30	-	90 74 54 37
Fed-KDDGC-SGD [130]	Transmitting only top gradients.	MNIST	-	120	95.9
PFLEGO [131]	Personalized FL with exact SGD minimization.	Omniglot [139] CIFAR-10 MNIST Fashion-MNIST EMNIST	-	5000 200 200 200 200	78 87.81 98.43 96.34 98.49
Intrinsic Gradient Compression [132]	Links intrinsic dimension to gradient compressibility for low-bandwidth FL.	CIFAR-10 Stanford Sentiment Treebank-v2 (SST-2) [140]	10000 500	8000 30	75 88
Adaptive Gradient Compression [133]	Assigns client-specific compression rates for non-IID and imbalanced data.	MNIST	-	1000	99.09
Wyner-Ziv [134]	Gradient compression with historical gradients using Wyner-Ziv coding.	CIFAR-10 CIFAR-100	8	100	94.53 76.45
FedGreen [141]	Fine-grained compression for energy-efficient MEC deployment, with device-side reduction and server aggregation.	CIFAR-10	16	300	84
Gaussian Compression [142]	Compresses updates by learning Gaussian distributions for gradient parameters.	CIFAR-10 CIFAR-100	5	50	87 72

In their paper [141], Li et al. introduced FedGreen, an FL framework incorporating fine grained gradient compression to enhance energy-efficient Mobile Edge Computing (MEC) deployment. They utilized device-side gradient reduction and server side element wise aggregation to enable effective compression. They evaluated the impact of compressed gradients on learning accuracy and energy efficiency, determining optimal compression ratios and computing frequencies for each device. In [142], authors proposed a novel method of model-update compression. The method learns multiple Gaussian distributions that best describe the high-dimensional gradient parameters, and in the FL server, these parameters are used to repopulate high dimensional gradients. Since the distribution information parameters make up a small percentage of values compared to the high-dimensional gradients themselves, the method can significantly save uplink bandwidth while preserving model accuracy.

The Table 5 shows the summary of contributions made in gradient compression approach for PFL. Techniques such as ClusterGrad leverage K-means clustering to select key gradients, optimizing computations with an accuracy of 87% on CIFAR-10. BS-pFL utilizes bit-stream predictions for gradient sparsity, achieving 97.45% on MNIST and 86.23% on CIFAR-10 with only 100 clients, while GMF incorporates a compensation strategy for gradient compression, recording accuracies of 72.46% and 41.9% on CIFAR-10 and Shakespeare datasets, respectively.

Other notable methods include PFLFM/PPGC, which employs feature fusion-based mutual learning to enhance communication efficiency, achieving an accuracy of 67.7% on CIFAR-10 with 500 rounds. AnycostFL focuses on localized updates across edge devices, achieving 90.32% accuracy on FMNIST and 84.91% on CIFAR-10. pFedGF introduces a dual-gradient approach, achieving up to 94.57% accuracy on MNIST and significant results on Fashion-MNIST and CIFAR-10.

Techniques like SPFL utilize Softmax Normalized Gradient Similarity (SNGS) for distributing personalized global models, reaching accuracies as high as 98.33% on MNIST and 87.49% on EMNIST. Q-RR and Q-NASTYA explore gradient compression with without-replacement sampling for smaller datasets like mushrooms and a9a. FedCAMS optimizes adaptive aggregation, attaining 90.2% on CIFAR-10, while FedOComp correlates stochastic gradients to enhance high-dimensional data compression, achieving a 90% accuracy on Fashion-MNIST. More advanced methods include FedSynth, where clients send synthetic datasets rather than model updates, and FedCG, which integrates adaptive client selection to mitigate stragglers. PFLEGO personalizes FL with exact SGD, yielding accuracies of up to 98.43% on MNIST and Fashion-MNIST, and Intrinsic Gradient Compression links compressibility to intrinsic dimensions, reporting 88% on SST-2. Techniques like Adaptive Gradient Compression and Wyner-Ziv coding further optimize performance, achieving high accuracy on various datasets, such as 99.09% on MNIST and 76.45% on CIFAR-100. Overall these techniques are used in FL to reduce communication overhead by transmitting compressed gradients instead of full gradients.

3.6. Model Caching

Model caching is a technique used in PFL to improve communication efficiency by reducing the amount of data transmitted between clients and the central server. It involves caching and reusing certain model components on the client side, which eliminates the need to send them repeatedly during each round of communication. In subsequent rounds, instead of transmitting the entire model, only the model differences or updates need to be communicated. This technique minimizes redundant communication and speeds up the convergence process. The Algorithm 6 shows the basic implementation of Model Caching technique for PFL.

Algorithm 6 Model Caching in PFL

Input(s): Global model M_{global}, Personalized model $M_{\text{personalized}}$, Communication round t
Output(s): Local model update U_{local}, Cached model M_{cache}
Algorithm:

```
1: if not first communication round then
2:     Load cached model M_cache from previous round
3: end if
4: if Global model not in cache then
5:     Cache global model M_global
6: end if
7: Update personalized model M_personalized using local data
8: Calculate local model update U_local as the difference between
   M_personalized and M_cache
9: Cache M_personalized as M_cache for future rounds
10: Transmit U_local to the central server
```

1. **Cached Model Check:**
   If this is not the first communication round, load the cached model M_{cache} from the previous round:
   $$M_{\text{cache}} \leftarrow M_{\text{cache}}^{(t-1)} \tag{25}$$
2. **Global Model Caching:**
   If the global model M_{global} is not yet in cache, cache it for the current round:
   $$M_{\text{cache}} = M_{\text{global}} \tag{26}$$
3. **Personalized Model Update:**
   Update the personalized model $M_{\text{personalized}}$ using local data at client i:
   $$M_{\text{personalized}}^{(t)} = \text{update}(M_{\text{personalized}}^{(t-1)}, \text{local data}_i) \tag{27}$$
4. **Local Model Update Calculation:**
   Calculate the local model update U_{local} as the difference between the personalized model $M_{\text{personalized}}^{(t)}$ and the cached model M_{cache}:
   $$U_{\text{local}} = M_{\text{personalized}}^{(t)} - M_{\text{cache}} \tag{28}$$
5. **Update Cache:**
   Cache the updated personalized model for future communication rounds:
   $$M_{\text{cache}} = M_{\text{personalized}}^{(t)} \tag{29}$$

Table 6
Summary of contributions in PFL Communication Optimization Techniques using Model Caching

Technique(s)	Main Idea/Contribution	Dataset(s) Used	Clients	Communication Rounds	Accuracy
FM-DRL/FLCC [143]	Federated Multi-agent Deep Reinforcement Learning for Edge Caching, enhancing edge co-operation for personalized service retrieval.	CIFAR10	100	5000	-
FedFilter [144]	Personalized FL using model decomposition and hierarchical aggregation, caching tailored content based on individual preferences.	ML100K	-	70	-
CREAT [145]	Integrates IoT, edge nodes, and blockchain to streamline caching in distributed environments.	MovieLens	30	100	-
FedCache [146]	Optimizes cache allocation, focusing on minimal communication while ensuring fair resource distribution.	IoT Data	-	1000	-
Edge Caching in IoV [147]	Edge caching framework for Internet of Vehicles (IoV), enhancing scalability and supporting decentralized region-to-region data exchanges.	Vehicle Location Data	-	-	-
Federated-CNN [148]	Proactive cache algorithm leveraging FL and CNN models to predict and cache popular content.	Movielens-1M	10	100	96.3
User-Centric Aggregation Rules [149]	Implements user-centric aggregation at the server, balancing personalization with communication efficiency through tailored aggregation rules.	EMNIST CIFAR-10 BigQuery [150]	20/100 20 35	100 200 150	73.2/76.4 48.8 84
FedQNN [151]	FL framework with ultralow-bitwidth quantization, integrating sparsification for efficient IoT communication.	MNIST Fashion-MNIST CIFAR10	100	5000 10000 20000	99.10 90.69 89.42

6. **Transmit Local Update:**
   Transmit the local model update U_{local} to the central server for aggregation.

Federated Multi-agent Deep Reinforcement Learning (FM-DRL) [143] integrates a cache-enabled FL system (FREC) designed for efficient data retrieval to provide personalized services. This method leverages model caching to minimize learning latency and ensure good convergence of the learning process. By caching frequently accessed data, FM-DRL can quickly retrieve necessary information, reducing the overall time required for training and updates in FL environments. Federated Learning and edge Cache-assisted Cybertwin (FLCC) [143] focuses on edge cooperation and optimization using a FM-DRL algorithm. This method customizes edge computing services to tackle specific challenges, such as data heterogeneity and dynamic network conditions. By coordinating multiple edge nodes, FLCC optimizes resource allocation and data caching, improving the overall efficiency and effectiveness of the FL process.

FedFilter [144] is an edge caching solution based on FL, addressing the challenges of content caching in diverse user environments. It uses a personalized FL approach involving model decomposition and hierarchical aggregation to cache content according to individual user preferences. FedFilter enhances the cache hit rate, reduces backhaul load, and minimizes service latency. Additionally, it detects and mitigates the effects of invalid data on the global model, ensuring the robustness and efficiency of the caching system.

CREAT [145] combines Internet of Things (IoT) devices, edge nodes, remote cloud, and blockchain technology to improve the efficiency of caching in FL systems. This approach improves the cache hit rate and reduces the time required to upload data by leveraging a distributed caching framework supported by blockchain for secure and efficient data management. Chilukuri et al. [146] introduced FedCache, a dynamic Cache allocation technique based on FL designed for edge caches in dynamic, resource-constrained networks. FedCache employs FL to acquire insights into optimal cache allocations with minimal communication overhead. Through local learning, edge nodes adapt to varying network conditions and collaboratively exchange this knowledge, eliminating the need to transmit all data to a central location. This FL-based approach enables nodes to determine resource allocations that maximize fairness or efficiency in terms of cache hit ratio, aligning with the current network state.

The study conducted by Oualil et al. [147] evaluated the application of advanced machine learning (ML) paradigms to enhance the precision of personalized edge caching and replacement decisions. Their focus included maintaining data privacy, accommodating vehicle mobility, accounting for the popularity of content that changes over time, and incorporating location awareness. The research demonstrated that conventional FL-based approaches struggle to maintain satisfactory performance in these conditions. To address this, the authors introduced a scalable-by-design edge caching scheme tailored for the Internet of Vehicles (IoV) context. This scheme harnesses decentralized exchanges among region-to-region Road Side Units (RSUs) to enhance local models within each region, thus optimizing edge caching effectiveness.

Zhu et al. [148] introduced a proactive cache algorithm based on the FL framework. In this approach, distributed virtual Content Delivery Network (vCDN) nodes utilize their respective data to train convolutional neural network (CNN) models, predicting the popularity of upcoming content. Through federated averaging of these node models, a global popularity prediction model is derived. Each vCDN node then employs this global model to forecast incoming requests, enhancing caching efficiency by preemptively predicting content popularity. The Federated-CNN approach ensures localized training instead of centralizing user data, reducing transmission requirements while preserving user privacy. The simulation results highlight the benefits of federated aggregation, with Federated CNN achieving approximately 22% higher accuracy in popularity prediction compared to other caching algorithms.

The study by Mestoukirdi et al. [149] focuses on overcoming the limitations of PFL by introducing the capability for personalization through the utilization of multiple user-centric aggregation rules at the parameter server. This approach offers the potential to generate personalized models for individual users, albeit with added downlink communication overhead. To balance the trade-off between personalization and communication efficiency, the authors propose a broadcast protocol that restricts the number of personalized streams while retaining the crucial benefits of their learning approach.

Excessive communication between the server and the clients can lead to demanding bandwidth prerequisites and increased energy consumption in various IoT systems. Addressing this, in [151], a framework named federated learning of quantized neural network (FedQNN) is introduced. This marks the pioneering incorporation of ultralow-bitwidth quantization into the FL context. This innovation enables clients to execute lightweight fixed-point computations with reduced power consumption.

In summary, Table 6 presents model caching techniques which are used in PFL to enhance communication efficiency by reducing the data transmitted between clients and the central server. Techniques like FM-DRL/FLCC utilize Federated Multi-agent Deep Reinforcement Learning for efficient edge caching, enhancing edge cooperation to retrieve personalized services on the CIFAR10 dataset with 100 clients and 5000 rounds. FedFilter focuses on model decomposition and hierarchical aggregation to personalize cached content based on user preferences, applied on ML100K with 70 communication rounds. CREAT integrates IoT, edge nodes, and blockchain to streamline caching in distributed setups, using MovieLens with 30 clients and 100 rounds.

Other methods include FedCache, which optimizes cache allocation by minimizing communication while maintaining fair resource distribution, applied to IoT data with 1000 rounds. In the Edge Caching in IoV framework, edge caching supports decentralized exchanges of vehicle location data, improving scalability for medium-to-large client groups. Federated-CNN employs a proactive cache algorithm with FL and CNN to predict popular content on Movielens-1M, reaching 96.3% accuracy.

The User-Centric Aggregation Rules approach provides user-specific aggregation at the server, balancing personalization and efficiency. This was applied on datasets like EMNIST and CIFAR-10, achieving accuracies of 73.2%, 76.4%, and 84% across different datasets and communication rounds. Finally, FedQNN integrates ultralow-bitwidth quantization with sparsification for IoT, reaching high accuracy rates such as 99.1% on MNIST and 90.69% on Fashion-MNIST over large-scale communication rounds, highlighting the potential of caching strategies in reducing communication load in PFL scenarios. These techniques leverage model caching to enhance edge caching efficiency, reduce service latency, maintain data privacy, and accommodate dynamic network conditions and user preferences.

4. Challenges and Open Research Directions

This section discusses challenges and potential future research directions in PFL communication optimization techniques. The table 7 presents the challenges and open research directions in PFL communication optimization techniques.

For Model Compression, challenges include high computational demands, particularly for resource-constrained devices, and the difficulty of balancing personalized tasks with communication efficiency. Techniques like FedMPT and QuPed attempt to address these issues, but still struggle with privacy and bandwidth. Future research should focus on enhancing multi-task performance in sensitive fields, exploring adaptive communication protocols, and investigating larger datasets. Methods like FedNILM and FedSUMM show promise in improving accuracy while reducing overhead.

Differential Privacy faces the challenge of balancing robust privacy with communication efficiency, especially in the context of non-IID data. Techniques such as PPPFL and PDPM tackle these issues but encounter scalability and adaptability concerns across clients. Future work could emphasize dynamic privacy mechanisms that optimize communication, adaptive methods, and integrating differential privacy with blockchain for secure FL.

In Client Selection, ensuring model accuracy and fairness while managing data heterogeneity and resource constraints is essential. Approaches like FedCS enhance client selection but still struggle with privacy issues. Future research should refine client selection through real-time performance metrics, improve clustering techniques, and develop decentralized frameworks for personalized recommendations.

Table 7
PFL Communication Optimization Techniques, Challenges and Open Research Directions

Model/Method	Challenges	Open Research Directions
Model compression	Model compression in FL can demand significant computational resources, often unsuitable for resource-constrained devices. Handling multiple personalized tasks simultaneously while maintaining efficient communication remains challenging [57, 55].	Future work could focus on enhancing multi-task performance in privacy-sensitive applications such as healthcare and finance, where balancing data privacy with computational efficiency is crucial. Research should investigate performance on larger datasets and explore adaptive, energy-efficient communication protocols that leverage model compression techniques [37, 41, 45, 49].
Differential Privacy	Achieving both robust privacy and communication efficiency in FL is complex, particularly with non-IID data and diverse client privacy requirements. Techniques such as PPPFL tackle these by using differentially private GANs to reduce communication rounds while maintaining high data utility, especially in cross-silo FL [69, 73, 84, 74, 76, 67].	Adaptive privacy-preserving methods like CUAG-PFL and PGC-LDP offer promising avenues to improve both privacy and communication efficiency, as they adjust to client resource constraints and minimize communication for specific model updates [70, 71, 81, 64], Comprehensive personalized privacy protection, performance and scalability, efficient and secure FL algorithms, extension to other healthcare applications [66].
Client Selection	Accurate device selection based on model quality and efficient client clustering based on computational efficiency are essential, but their effectiveness can vary with different datasets and participants. Challenges include the assumption of cooperative clients and concerns about privacy protection during training. The FCFL approach addresses biases, aggregation inefficiencies, and privacy issues, but it has limitations compared to other methods [93, 95, 96, 102].	Future work could focus on refining client selection by integrating real-time performance metrics and communication capabilities to dynamically adjust client participation [93, 95, 96, 99]. Enhance client grouping strategies to better handle extreme data heterogeneity [90], generalization of FedRec++ [101], Denoising strategy applicability, FedeRiCo for limited resources [99].
Asynchronous Updates	Asynchronous updates in Personalized Federated Learning (PFL) bring diverse approaches to handling client heterogeneity and communication efficiency, with each technique presenting unique challenges and opportunities for future research [105, 106]. Design for heterogeneous edge devices, employs cost-adjustable local updates, but refining model shrinking techniques tailored to device-specific constraints remains an open direction [136, 109, 104].	Further advancements in asynchronous PFL include FedACA, which introduces a two-level feedback system to reduce communication costs amidst data heterogeneity; context-aware adaptive policies could advance its resilience [112, 113, 114, 118].
Gradient Compression	Many existing techniques struggle to maintain model accuracy when clients' data are highly heterogeneous [121, 133]. The efficiency of various methods can fluctuate based on specific client conditions and data distributions [125, 123]. Moreover, while adaptive compression strategies that account for client capacities have shown promise [129, 127], highlighting the need for strategies that balance communication efficiency with effective model training.	Ensuring consistent model performance across diverse client environments [126, 128]. There is also a need for more comprehensive studies on reducing compression variance to enhance the reliability of gradient updates [125]. Exploring adaptive compression methods that consider client-specific conditions could yield improved communication efficiency while maintaining training effectiveness [133, 129, 132, 131].
Model Caching	Despite the potential benefits of model caching in federated learning (FL), several challenges persist. One significant issue is the management of data heterogeneity, as client devices may have varying data distributions that complicate effective caching strategies [143, 144]. Additionally, dynamic network conditions can affect the performance of caching algorithms, leading to increased latency and reduced service quality [143].	Exploring decentralized caching strategies that leverage the collaborative nature of FL can enhance efficiency while preserving data privacy [146]. Additionally, integrating advanced machine learning paradigms with caching solutions could improve the accuracy of content prediction and allocation decisions, especially in rapidly changing environments [147].

Asynchronous Updates present challenges in managing client heterogeneity and ensuring efficient communication. Techniques like APFedMe and ReFRS show potential but require further real-time adaptability and scalability improvements. Innovations such as FedACA and hyper-mix Async-DFL could enhance performance by balancing feedback and communication costs, while also addressing concept drift and staleness.

The challenges of Gradient Compression include managing non-IID data distributions and compression variance without sacrificing model accuracy. There is a significant trade-off between the degree of compression and convergence speed. Future research should focus on developing robust algorithms that effectively manage data variability, reduce compression variance, and explore the interactions between gradient compression and other optimization techniques like model caching.

Lastly, Model Caching faces challenges in managing data heterogeneity and dynamic network conditions while ensuring data privacy. The trade-off between personalization and communication efficiency remains a critical issue. Future research should develop adaptive caching algorithms and decentralized strategies that leverage the collaborative nature of federated learning. Integrating advanced machine learning paradigms with caching solutions and balancing personalization with communication efficiency are vital areas for exploration.

5. Conclusion

In conclusion, this research survey delved into a comprehensive exploration of various communication optimization techniques within the realm of personalized federated learning (PFL). Throughout the paper, we examined the strategies including model compression, differential privacy, client selection, asynchronous updates, gradient compression, and model caching. These techniques collectively address the multifaceted challenges inherent in FL, with a specific focus on enhancing communication efficiency without compromising the quality of personalized model training. **Model compression** techniques, by reducing the model's size, minimize communication overhead while maintaining performance. **Differential privacy** mechanisms safeguard sensitive data during the aggregation process, protecting privacy without limiting collaborative learning. The **selection of clients** optimizes the subset of participants, promoting efficiency in the training process. **Asynchronous updates** accommodate varying client availability, allowing for flexible and efficient model updates. **Gradient compression** methods reduce communication load by transmitting compressed gradients, optimizing bandwidth usage. Lastly, **model caching** leverages prior training to enhance local updates, minimizing the need for extensive communication during subsequent rounds.

Incorporating these communication optimization techniques collectively enriches the landscape of personalized FL, enabling efficient, secure, and precise model training across distributed devices while mitigating the challenges posed by communication constraints. As the field of PFL continues to evolve, these insights into communication optimization strategies will undoubtedly play a pivotal role in shaping the future of collaborative and privacy-aware machine learning paradigms. The conclusions of our study clearly identify a huge potential for future research in communication optimization approaches for PFL.

Acknowledgment

This work is sponsored by the National Key R&D Program of China under Grant number (2022YFB3103100). This work is sponsored by the R& D Program of Beijing Municipal Education Commission (KM202210005028). This work is also supported by National Natural Science Foundation of China (62302020) and Major Research Plan of National Natural Science Foundation of China (92167102). This work is supported by Importation and Development of High-Caliber Talents Project of Beijing Municipal Institutions (CIT&TCD 20190308) and the "Engineering Research Center of Intelligent Perception and Autonomous Control, Ministry of Education"

References

[1] Kairouz Peter, McMahan H. Brendan, and Avent Brendan et al. Advances and open problems in federated learning. *Foundations and trends in machine learning*, 14(1–2):1–210, 2021.

[2] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Trans. Intell. Syst. Technol.*, 10(2):1–19, jan 2019.

[3] Jakub Konecnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint*, 8, 2016.

[4] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60, 2020.

[5] Pushpa Singh, Murari Kumar Singh, Rajnesh Singh, and Narendra Singh. Federated learning: Challenges, methods, and future directions. In *Federated Learning for IoT Applications*, pages 199–214. Springer, Cham, 2022.

[6] Kavita Bhosle and Vijaya Musande. Evaluation of deep learning cnn model for recognition of devanagari digit. *Artificial Intelligence and Applications*, 1(2):114–118, Feb. 2023.

[7] Qianyang Sun, Jiaming Chen, Li Zhou, Shifeng Ding, and Sen Han. A study on ice resistance prediction based on deep learning data generation method. *Ocean Engineering*, 301:117467, 2024.

[8] Yingjie Song, Lihuan Han, Bin Zhang, and Wu Deng. A dual-time dual-population multi-objective evolutionary algorithm with application to the portfolio optimization problem. *Engineering Applications of Artificial Intelligence*, 133:108638, 2024.

[9] Timileyin Opeyemi Akande, Oluwaseyi O. Alabi, and Sunday A. Ajagbe. A deep learning-based cae approach for simulating 3d vehicle wheels under real-world conditions. *Artificial Intelligence and Applications*, Jan. 2024.

[10] Zheng-yi Chai, Chuan-dong Yang, and Ya-lun Li. Communication efficiency optimization in federated learning based on multi-objective evolutionary algorithm. *Evolutionary Intelligence*, 16(3):1033–1044, 2023.

[11] Yuwei Fan, Wei Xi, Hengyi Zhu, and Jizhong Zhao. Minipfl: Mini federations for hierarchical personalized federated learning. *Future Generation Computer Systems*, 157:41–50, 2024.

[12] Jingke Tu, Jiaming Huang, Lei Yang, and Wanyu Lin. Personalized federated learning with layer-wise feature transformation via meta-learning. *ACM Trans. Knowl. Discov. Data*, 18(4), February 2024.

[13] Zhengrong Song, Chuan Ma, Ming Ding, Howard H. Yang, Yuwen Qian, and Xiangwei Zhou. Personalized federated deep reinforcement learning-based trajectory optimization for multi-uav assisted edge computing. In *2023 IEEE/CIC International Conference on Communications in China (ICCC)*, pages 1–6, Aug 2023.

[14] Jiagao Wu, Yu Wang, Zhangchi Shen, and Linfeng Liu. Adaptive client and communication optimizations in federated learning. *Information Systems*, 116:102226, 2023.

[15] Chuhan Wu, Fangzhao Wu, Lingjuan Lyu, Yongfeng Huang, and Xing Xie. Communication-efficient federated learning via knowledge distillation. *Nature communications*, 13(1):2032, 2022.

[16] Peiyan Yuan, Ling Shi, Xiaoyan Zhao, and Junna Zhang. A lightweight and personalized edge federated learning model. *Complex & Intelligent Systems*, pages 1–16, 2024.

[17] Shaoshuai Fan, Jie Ni, and Hui Tian. Fast personalized federated learning in wireless networks with heterogeneous data and limited communication resources. *IEEE Internet of Things Journal*, 11(17):28555–28565, Sep. 2024.

[18] Zhihan Wang, Xiangxue Ma, Haixia Zhang, and Dongfeng Yuan. Communication-efficient personalized federated learning for digital twin in heterogeneous industrial iot. In *2023 IEEE International Conference on Communications Workshops (ICC Workshops)*, pages 237–241, May 2023.

[19] Tian Wang, Yan Liu, Xi Zheng, Hong-Ning Dai, Weijia Jia, and Mande Xie. Edge-based communication optimization for distributed federated learning. *IEEE Transactions on Network Science and Engineering*, 9(4):2015–2024, July 2022.

[20] Jun Wu, Wenxuan Bao, Elizabeth Ainsworth, and Jingrui He. Personalized federated learning with parameter propagation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 2594–2605, New York, NY, USA, 2023. Association for Computing Machinery.

[21] Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet Talwalkar. Federated multi-task learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4424–4434, Long Beach, California, USA, 2017. Curran Associates Inc.

[22] Jiajun Wu, Fan Dong, Henry Leung, Zhuangdi Zhu, Jiayu Zhou, and Steve Drew. Topology-aware federated learning in edge computing: A comprehensive survey. *ACM Comput. Surv.*, apr 2024.

[23] Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12):9587–9603, 2023.

[24] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, page 1175–1191, Dallas, Texas, USA, 2017. Association for Computing Machinery.

[25] Jie Wen, Zhixia Zhang, Yang Lan, Zhihua Cui, Jianghui Cai, and Wensheng Zhang. A survey on federated learning: challenges and applications. *International Journal of Machine Learning and Cybernetics*, 14(2):513 – 535, 2023.

[26] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. Scaffold: stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20, page 5132–5143, Vienna, Austria, 2020. JMLR.org.

[27] Fei Chen, Mi Luo, Zhenhua Dong, Zhenguo Li, and Xiuqiang He. Federated meta-learning with fast convergence and efficient communication. *arXiv preprint*, 2018.

[28] Taki Hasan Rafi, Faiza Anan Noor, Tahmid Hussain, and Dong-Kyu Chae. Fairness and privacy preserving in federated learning: A survey. *Information Fusion*, 105:102198, 2024.

[29] Zihao Zhao, Yuzhu Mao, Zhenpeng Shi, Yang Liu, Tian Lan, Wenbo Ding, and Xiao-Ping Zhang. Aquila: Communication efficient federated learning with adaptive quantization in device selection strategy. *IEEE Transactions on Mobile Computing*, 23(6):7363–7376, 2024.

[30] Thuy Dung Nguyen, Tuan Nguyen, Phi Le Nguyen, Hieu H. Pham, Khoa D. Doan, and Kok-Seng Wong. Backdoor attacks and defenses in federated learning: Survey, challenges and future research directions. *Engineering Applications of Artificial Intelligence*, 127:107166, 2024.

[31] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. *arXiv preprint*, 2019.

[32] Hao Wang, Zakhary Kaplan, Di Niu, and Baochun Li. Optimizing federated learning on non-iid data with reinforcement learning. In *IEEE INFOCOM 2020-IEEE conference on computer communications*, pages 1698–1707, Toronto, ON, Canada, 2020. IEEE.

[33] Fahad Sabah, Yuwen Chen, Zhen Yang, Muhammad Azam, Nadeem Ahmad, and Raheem Sarwar. Model optimization techniques in personalized federated learning: A survey. *Expert Systems with Applications*, 243:122874, 2024.

[34] Viraaji Mothukuri, Reza M Parizi, Seyedamin Pouriyeh, Yan Huang, Ali Dehghantanha, and Gautam Srivastava. A survey on security and privacy of federated learning. *Future Generation Computer Systems*, 115:619–640, 2021.

[35] Lingjuan Lyu, Han Yu, Xingjun Ma, Chen Chen, Lichao Sun, Jun Zhao, Qiang Yang, and Philip S. Yu. Privacy and robustness in federated learning: Attacks and defenses. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–21, 2022.

[36] Peichun Li, Guoliang Cheng, Xumin Huang, Jiawen Kang, Rong Yu, Yuan Wu, Miao Pan, and Dusit Niyato. Snowball: Energy efficient and accurate federated learning with coarse-to-fine compression over heterogeneous wireless edge devices. *IEEE Transactions on Wireless Communications*, 22(10):6778–6792, 2023.

[37] Longfei Zheng, Yingting Liu, Xiaolong Xu, Chaochao Chen, Yuzhou Tang, Lei Wang, and Xiaolong Hu. Fedpse: Personalized sparsification with element-wise aggregation for federated learning. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 3514–3523, New York, NY, USA, 2023. Association for Computing Machinery.

[38] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

[39] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[40] Ji Liu and Stephen J. Wright. Asynchronous stochastic coordinate descent: Parallelism and convergence properties. *SIAM Journal on Optimization*, 25(1):351–376, 2015.

[41] Xinglin Zhang, Zhaojing Ou, and Zheng Yang. Fedmpt: Federated learning for multiple personalized tasks over mobile computing. *IEEE Transactions on Network Science and Engineering*, 10(4):2358 – 2371, 2023.

[42] Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. Dynamic routing between capsules. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 3859–3869, Red Hook, NY, USA, 2017. Curran Associates Inc.

[43] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015.

[44] Kaan Ozkara, Navjot Singh, Deepesh Data, and Suhas Diggavi. Quped: Quantized personalization via distillation with applications to federated learning. In *Advances in Neural Information Processing Systems*, volume 5, page 3622 – 3634, virtual, 2021.

[45] Yu Zhang, Guoming Tang, Qianyi Huang, Yi Wang, Kui Wu, Keping Yu, and Xun Shao. Fednilm: Applying federated learning to nilm applications at the edge. *IEEE Transactions on Green Communications and Networking*, 7(2):857 – 868, 2023.

[46] David Murray, Lina Stankovic, and Vladimir Stankovic. An electrical load measurements dataset of united kingdom households from a two-year longitudinal study. *Scientific data*, 4(1):1–12, 2017.

[47] Jack Kelly and William Knottenbelt. The uk-dale dataset, domestic appliance-level electricity demand and whole-house demand from five uk homes. *Scientific data*, 2(1):1–14, 2015.

[48] J Zico Kolter and Matthew J Johnson. Redd: A public data set for energy disaggregation research. In *Workshop on data mining applications in sustainability (SIGKDD), San Diego, CA*, volume 25, pages 59–62. Citeseer, 2011.

[49] Rong Pan, Jianzong Wang, Lingwei Kong, Zhangcheng Huang, and Jing Xiao. Personalized federated learning via gradient modulation for heterogeneous text summarization. *arXiv preprint*, abs/2304.11524, 2023.

[50] Yudong Li, Yuqing Zhang, Zhe Zhao, Linlin Shen, Weijie Liu, Weiquan Mao, and Hui Zhang. Csl: A large-scale chinese scientific literature dataset. *arXiv preprint*, 2022.

[51] Xiaojun Liu, Chuang Zhang, Xiaojun Chen, Yanan Cao, and Jinpeng Li. Clts: a new chinese long text summarization dataset. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 531–542. Springer, 2020.

[52] Baotian Hu, Qingcai Chen, and Fangze Zhu. Lcsts: A large scale chinese short text summarization dataset. *arXiv preprint*, 2015.

[53] Maosong Sun, Jingyang Li, Zhipeng Guo, Zhao Yu, Yabin Zheng, Xiance Si, and Zhiyuan Liu. Thuctc: an efficient chinese text classifier. *GitHub Repository*, 2016.

[54] El Houcine Bergou, Konstantin Burlachenko, Aritra Dutta, and Peter Richt'arik. Personalized federated learning with communication compression. *arXiv preprint*, abs/2209.05148, 2022.

[55] Zhize Li, Haoyu Zhao, Boyue Li, and Yuejie Chi. Soteriafl: A unified framework for private federated learning with communication compression. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 4285–4300, New Orleans, LA, USA, 2022. Curran Associates, Inc.

[56] Rong Dai, Li Shen, Fengxiang He, Xinmei Tian, and Dacheng Tao. DisPFL: Towards communication-efficient personalized federated learning via decentralized sparse training. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 4587–4604, Baltimore, Maryland, USA, 17–23 Jul 2022. PMLR.

[57] Filip Hanzely, Boxin Zhao, and Mladen Kolar. Personalized federated learning: A unified framework and universal optimization techniques. *arXiv preprint*, 2021.

[58] Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. *CoRR*, abs/1812.01718, 2018.

[59] Yuanxiong Guo, Rui Hu, and Yanmin Gong. Agent-level differentially private federated learning via compressed model perturbation. In *2022 IEEE Conference on Communications and Network Security (CNS)*, pages 127–135, Austin, TX, USA, 2022.

[60] Leming Wu, Yaochu Jin, and Kuangrong Hao. Optimized compressed sensing for communication efficient federated learning. *Knowledge-Based Systems*, 278:110805, 2023.

[61] Irem Ergün, Hasin Us Sami, and Başak Güler. Communication-efficient secure aggregation for federated learning. In *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, pages 3881–3886, 2022.

[62] Yixuan Liu, Suyun Zhao, Li Xiong, Yuhan Liu, and Hong Chen. Echo of neighbors: Privacy amplification for personalized private federated learning with shuffle model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11865–11872, Washington, DC USA, 2023.

[63] Kang Wei, Jun Li, Chuan Ma, Ming Ding, Wen Chen, Jun Wu, Meixia Tao, and H. Vincent Poor. Personalized federated learning with differential privacy and convergence guarantee. *IEEE Transactions on Information Forensics and Security*, 18:4488 – 4503, 2023.

[64] Yi Shi, Kang Wei, Li Shen, Yingqi Liu, Xueqian Wang, Bo Yuan, and Dacheng Tao. Towards the flatter landscape and better generalization in federated learning under client-level differential privacy. *arXiv preprint*, abs/2305.00873, 2023.

[65] David Odera. Federated learning and differential privacy in clinical health: Extensive survey. *World Journal of Advanced Engineering Technology and Sciences*, 8(2):305–329, 2023.

[66] Addi Ait-Mlouk, Sadi Alawadi, Salman Zubair Toor, and Andreas Hellander. Fedbot: Enhancing privacy in chatbots with federated learning. *arXiv preprint*, abs/2304.03228, 2023.

[67] Xia Wu, Lei Xu, and Liehuang Zhu. Local differential privacy-based federated learning under personalized settings. *Applied Sciences*, 13(7):1–17, 2023.

[68] Jock Blackard. Covertype. UCI Machine Learning Repository, 1998. DOI: https://doi.org/10.24432/C50K5N.

[69] Van-Tuan Tran, Huy-Hieu Pham, and Kok-Seng Wong. Personalized privacy-preserving framework for cross-silo federated learning. *IEEE Transactions on Emerging Topics in Computing*, page 1–12, 2024.

[70] Min Li, Di Xiao, and Lü-Jun Chen. Communication-efficient and utility-aware adaptive gaussian differential privacy for personalized federated learning. *Jisuanji Xuebao/Chinese Journal of Computers*, 47(4):924 – 946, 2024.

[71] Muhammad Firdaus, Siwan Noh, Zhuohao Qian, Harashta Tatimma Larasati, and Kyung-Hyune Rhee. Personalized federated learning for heterogeneous data: A distributed edge clustering approach. *Mathematical Biosciences and Engineering*, 20(6):10725–10740, 2023.

[72] Feng Yu, Hui Lin, Xiaoding Wang, Sahil Garg, Georges Kaddoum, Satinderbir Singh, and Mohammad Mehedi Hassan. Communication-efficient personalized federated meta-learning in edge networks. *IEEE Transactions on Network and Service Management*, 20:1558–1571, 2023.

[73] Xiaoying Shen, Hang Jiang, Yange Chen, Baocang Wang, and Le Gao. Pldp-fl: Federated learning with personalized local differential privacy. *Entropy*, 25(3):1–20, 2023.

[74] Ge Yang, Shaowei Wang, and Haijie Wang. Federated learning with personalized local differential privacy. In *2021 IEEE 6th International Conference on Computer and Communication Systems (ICCCS)*, pages 484–489, Chengdu, China, 2021.

[75] Yunting Xie and Lan Zhang. Federated learning with personalized differential privacy combining client selection. In *2022 8th International Conference on Big Data Computing and Communications (BigCom)*, pages 79–87, Xiamen, China, 2022.

[76] Filippo Galli, Sayan Biswas, Kangsoo Jung, Catuscia Palamidessi, and Tommaso Cucinotta. Group privacy for personalized federated learning. In *International Conference on Information Systems Security and Privacy*, pages 1–15, Lisbon, Portugal, 2022.

[77] Jinhao Zhou, Zhou Su, Jianbing Ni, Yuntao Wang, Yanghe Pan, and Rui Xing. Personalized privacy-preserving federated learning: Optimized trade-off between utility and privacy. In *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, pages 4872–4877, Rio de Janeiro, Brazil, 2022.

[78] Yanhang Shi, Siguang Chen, and Haijun Zhang. Uncertainty minimization for personalized federated semi-supervised learning. *IEEE Transactions on Network Science and Engineering*, 10:1060–1073, 2022.

[79] Zhenyu Li. A personalized privacy-preserving scheme for federated learning. In *2022 IEEE International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA)*, pages 1352–1356, Changchun, China, 2022.

[80] Andrew Silva, Katherine Metcalf, Nicholas Apostoloff, and Barry-John Theobald. Fedembed: Personalized private federated learning. *arXiv preprint*, abs/2202.09472, 2022.

[81] Zheshun Wu, Xiaoping Wu, Xiaoli Long, and Yunliang Long. A privacy-preserved online personalized federated learning framework for indoor localization. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2834–2839, Melbourne, Australia, 2021.

[82] Joaquín Torres-Sospedra, Raúl Montoliu, Adolfo Martínez-Usó, Joan P. Avariento, Tomás J. Arnau, Mauri Benedito-Bordonau, and Joaquín Huerta. Ujiindoorloc: A new multi-building and multi-floor database for wlan fingerprint-based indoor localization problems. In *2014 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, pages 261–270, 2014.

[83] Shaojian Chen, Qiongqiong Zhu, Zihao Li, and Yunliang Long. Deep neural network based on feature fusion for indoor wireless localization. In *2018 International Conference on Microwave and Millimeter Wave Technology (ICMMT)*, pages 1–3, 2018.

[84] Jiechao Gao, Mingyue Tang, Tianhao Wang, and Bradford Campbell. Pfed-ldp: A personalized federated local differential privacy framework for iot sensing data. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, SenSys '22, page 835–836, New York, NY, USA, 2023. Association for Computing Machinery.

[85] Alberto Bietti, Chen-Yu Wei, Miroslav Dudik, John Langford, and Steven Wu. Personalization improves privacy-accuracy tradeoffs in federated learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 1945–1962, Baltimore, Maryland, USA,

17–23 Jul 2022. PMLR.

[86] Aron N. Horvath, Matteo Berchier, Farhad Nooralahzadeh, Ahmed Allam, and M. Krauthammer. Exploratory analysis of federated learning methods with differential privacy on mimic-iii. *arXiv preprint*, abs/2302.04208, 2023.

[87] Sichun Luo, Yuanzhang Xiao, Yang Liu, Congduan Li, and Linqi Song. Towards communication efficient and fair federated personalized sequential recommendation. In *2022 5th International Conference on Information Communication and Signal Processing, ICICSP 2022*, page 448 – 453, Shenzhen, China, 2022.

[88] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52, 2015.

[89] Jure Leskovec and Rok Sosič. Snap: A general-purpose network analysis and graph-mining library. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(1):1–20, 2016.

[90] Xuming Han, Qiaohong Zhang, Zaobo He, and Zhipeng Cai. Confidence-based similarity-aware personalized federated learning for autonomous iot. *IEEE Internet of Things Journal*, 11(7):13070 – 13081, 2024.

[91] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

[92] Unais Sait, K Lal, S Prajapati, Rahul Bhaumik, Tarun Kumar, S Sanjana, and Kriti Bhalla. Curated dataset for covid-19 posterior-anterior chest radiography images (x-rays). *Mendeley Data*, 1(J), 2020.

[93] Zihao Zhao, Yuzhu Mao, Zhenpeng Shi, Yang Liu, Tian Lan, Wenbo Ding, and Xiao-Ping Zhang. Aquila: Communication efficient federated learning with adaptive quantization in device selection strategy. *IEEE Transactions on Mobile Computing*, 23(6):7363–7376, 2024.

[94] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint*, 2016.

[95] Ying-Chi Mao, Li-Juan Shen, Jun Wu, Ping Ping, and Jie Wu. Federated dynamic client selection for fairness guarantee in heterogeneous edge computing. *J. Comput. Sci. Technol*, pages 139–158, 2024.

[96] Ziqi Chen, Jun Du, Xiangwang Hou, Keping Yu, Jintao Wang, and Zhu Han. Channel adaptive and sparsity personalized federated learning for privacy protection in smart healthcare systems. *IEEE Journal of Biomedical and Health Informatics*, page 1–9, 2024.

[97] Lingtao Wei. Communication efficient federated personalized recommendation. *Frontiers in Computing and Intelligent Systems*, 2(3):63–67, 2023.

[98] F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4), December 2015.

[99] Yi Sui, Junfeng Wen, Yenson Lau, Brendan Leigh Ross, and Jesse C Cresswell. Find your friends: Personalized federated learning with the right collaborators. In *Workshop on Federated Learning: Recent Advances and New Challenges (in Conjunction with NeurIPS 2022)*, New Orleans, LA, USA, 2022.

[100] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, pages 213–226. Springer, 2010.

[101] Guanyu Lin, Feng Liang, Weike Pan, and Zhong Ming. Fedrec: Federated recommendation with explicit feedback. *IEEE Intelligent Systems*, 36(5):21–30, 2021.

[102] Pengyuan Zhou, Hengwei Xu, Lik Hang Lee, Pei Fang, and Pan Hui. Are you left out? an efficient and fair federated learning for personalized profiles on wearable devices of inferior networking conditions. In *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, pages 1–25, New York, NY, USA, 2022. Association for Computing Machinery.

[103] Sichun Luo, Yuanzhang Xiao, Yang Liu, Congduan Li, and Linqi Song. Towards communication efficient and fair federated personalized sequential recommendation. In *2022 5th International Conference on Information Communication and Signal Processing (ICICSP)*, pages 1–6, Shenzhen, China, 2022.

[104] Jingwei Sun, Ang Li, Lin Duan, Samiul Alam, Xuliang Deng, Xin Guo, Haiming Wang, Maria Gorlatova, Mi Zhang, Hai Li, and Yiran Chen. Fedsea: A semi-asynchronous federated learning framework for extremely heterogeneous devices. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, SenSys '22, page 106–119, New York, NY, USA, 2023. Association for Computing Machinery.

[105] Anwar Asad, Mostafa M. Fouda, Zubair Md Fadlullah, Mohamed I. Ibrahem, and Nidal Nasser. Moreau envelopes-based personalized asynchronous federated learning: Improving practicality in network edge intelligence. In *Proceedings - IEEE Global Communications Conference, GLOBECOM*, page 2033 – 2038, Kuala Lumpur, Malaysia, 2023.

[106] Mubashir Imran, Hongzhi Yin, Tong Chen, Quoc Viet Hung Nguyen, Alexander Zhou, and Kai Zheng. Refrs: Resource-efficient federated recommender system for dynamic and diversified user preferences. *ACM Transactions on Information Systems*, 41(3), 2023.

[107] Òscar Celma Herrada et al. *Music recommendation and discovery in the long tail*. Universitat Pompeu Fabra, 2009.

[108] Peichun Li, Guoliang Cheng, Xumin Huang, Jiawen Kang, Rong Yu, Yuan Wu, and Miao Pan. Anycostfl: Efficient on-demand federated learning over heterogeneous edge devices. *arXiv preprint*, abs/2301.03062, 2023.

[109] Mohammad Taha Toghani, Soomin Lee, and César A Uribe. Persafl: personalized asynchronous federated learning. *Optimization Methods and Software*, pages 1–38, 2023.

[110] Jun Lin, Jin Ma, and Jianguo Zhu. Privacy-preserving household characteristic identification with federated learning method. *IEEE Transactions on Smart Grid*, 13:1088–1099, 2022.

[111] Cathy Mannion. Smart metering project commission for energy regulation (cer) ireland. In *IET Seminar on Smart Metering 2010: Delivering a Smart UK*, pages 1–12, 2010.

[112] Shuang Zhou, Yuankai Huo, Shunxing Bao, Bennett A. Landman, and Aniruddha S. Gokhale. Fedaca: An adaptive communication-efficient asynchronous framework for federated learning. In *2022 IEEE International Conference on Autonomic Computing and Self-Organizing Systems (ACSOS)*, pages 71–80, CA, USA, 2022.

[113] Zhikun Chen, Jiaqi Pan, and Sihai Zhang. Asynchronous federated learning in decentralized topology based on dynamic average consensus. In *ICC 2022 - IEEE International Conference on Communications*, pages 2822–2827, Seoul, Korea, 2022.

[114] Yujing Chen, Zheng Chai, Yue Cheng, and Huzefa Rangwala. Asynchronous federated learning for sensor data with concept drift. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 4822–4831, Orlando, FL, USA, 2021.

[115] Jianmo Ni, Larry Muhlstein, and Julian McAuley. Modeling heart rate and activity data for personalized fitness recommendation. In *The World Wide Web Conference*, pages 1343–1353, 2019.

[116] Kdd cup of fresh air.

[117] Yonatan Vaizman, Katherine Ellis, and Gert Lanckriet. The extrasensory dataset: Recognizing detailed human context in-the-wild from smartphones and smartwatches, 2016. Available at: [http://extrasensory.ucsd.edu/].

[118] Y. Chen, Xiao yan Sun, and Yaochu Jin. Communication-efficient federated deep learning with layerwise asynchronous model update and temporally weighted aggregation. *IEEE Transactions on Neural Networks and Learning Systems*, 31:4229–4238, 2019.

[119] Khaled Ben Letaief, Zubair Md. Fadlullah, and Mostafa M. Fouda. Efficient wireless network slicing in 5g networks: An asynchronous federated learning approach. In *2022 IEEE International Conference*

on Internet of Things and Intelligence Systems (IoTaIS), pages 285–289, Bali, Indonesia, 2022.
[120] Sabrina Kall and Slim Trabelsi. An asynchronous federated learning approach for a security source code scanner. In *7th International Conference on Information Systems Security and Privacy (ICISSP 2021)*, pages 572–579, 2021.
[121] Chun-Chih Kuo, Ted T. Kuo, and Chia-Yu Lin. Improving federated learning communication efficiency with global momentum fusion for gradient compression schemes. *arXiv preprint*, abs/2211.09320, 2022.
[122] Laizhong Cui, Xiaoxin Su, Yipeng Zhou, and Lei Zhang. Clustergrad: Adaptive gradient compression by clustering in federated learning. In *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, pages 1–7, Taipei, Taiwan, 2020.
[123] Lening Wang, Manojna Sistla, Mingsong Chen, and Xin Fu. Bspfl: Enabling low-cost personalized federated learning by exploring weight gradient sparsity. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, Padua, Italy, 2022.
[124] Xinghao Wu, Jianwei Niu, Xuefeng Liu, Tao Ren, Zhangmin Huang, and Zhetao Li. pfedgf: Enabling personalized federated learning via gradient fusion. In *2022 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 639–649, Lyon, France, 2022.
[125] Abdurakhmon Sadiev, Grigory Malinovsky, Eduard Gorbunov, Igor Sokolov, Ahmed Khaled, Konstantin Pavlovich Burlachenko, and Peter Richtárik. Federated optimization algorithms with random reshuffling and gradient compression. In *40th International Conference on Machine Learning*, Honolulu, Hawaii, USA, 2024.
[126] Yujia Wang, Lu Lin, and Jinghui Chen. Communication-efficient adaptive federated learning. *arXiv preprint*, abs/2205.02719, 2022.
[127] Ye Xue, Liqun Su, and Vincent K. N. Lau. Fedocomp: Two-timescale online gradient compression for over-the-air federated learning. *IEEE Internet of Things Journal*, 9:19330–19345, 2022.
[128] Shengyuan Hu, Jack Goetz, Kshitiz Malik, Hongyuan Zhan, Zhe Liu, and Yue Liu. Fedsynth: Gradient compression via synthetic data in federated learning. *arXiv preprint*, abs/2204.01273, 2022.
[129] Zhida Jiang, Yang Xu, Hong-Ze Xu, Zhiyuan Wang, and Chen Qian. Adaptive control of client selection and gradient compression for efficient federated learning. *arXiv preprint*, abs/2212.09483, 2022.
[130] Haiyan Cui, Junping Du, Yang Jiang, Yue Wang, and Runyu Yu. Federated learning method based on knowledge distillation and deep gradient compression. In *2021 IEEE 7th International Conference on Cloud Computing and Intelligent Systems (CCIS)*, pages 423–427, Xi'an, China, 2021.
[131] Sotirios Nikoloutsopoulos, Iordanis Koutsopoulos, and Michalis K. Titsias. Personalized federated learning with exact stochastic gradient descent. *arXiv preprint*, abs/2202.09848, 2022.
[132] Luke Melas-Kyriazi and Franklyn Wang. Intrinsic gradient compression for scalable and efficient federated learning. In *Proceedings of the First Workshop on Federated Learning for Natural Language Processing (FL4NLP 2022)*, page 27–41, Dublin, Ireland, 2022.
[133] Wei Yang, Yuan Yang, Xiaobin Dang, Hao Jiang, Yizhe Zhang, and Wei Xiang. A novel adaptive gradient compression approach for communication-efficient federated learning. In *2021 China Automation Congress (CAC)*, pages 674–678, Beijing, China, 2021.
[134] Kai Liang, Huiru Zhong, Haoning Chen, and Youlong Wu. Wyner-ziv gradient compression for federated learning. *arXiv preprint*, abs/2111.08277, 2021.
[135] Qian Wang, Siguang Chen, and Meng Wu. Communication-efficient personalized federated learning with privacy-preserving. *IEEE Transactions on Network and Service Management*, 21(2):2374 – 2388, 2024.
[136] Peichun Li, Guoliang Cheng, Xumin Huang, Jiawen Kang, Rong Yu, Yuan Wu, and Miao Pan. Anycostfl: Efficient on-demand federated learning over heterogeneous edge devices. In *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*, pages 1–10, New York City, NY, USA, 2023. IEEE.
[137] Jing Xie, Xiang Yin, Xiyi Zhang, Juan Chen, and Quan Wen. Personalized federated learning with gradient similarity. In *2021 18th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages 268–271, Chengdu, China, 2021.
[138] Sebastian Caldas, Peter Wu, Tian Li, Jakub Konečný, H. Brendan McMahan, Virginia Smith, and Ameet Talwalkar. LEAF: A benchmark for federated settings. *CoRR*, abs/1812.01097, 2018.
[139] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.
[140] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.
[141] Peichun Li, Xumin Huang, Miao Pan, and Rong Yu. Fedgreen: Federated learning with fine-grained gradient compression for green mobile edge computing. In *2021 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6, Madrid, Spain, 2021.
[142] Birendra Kathariya, Zhu Li, Jianle Chen, and Geert Van der Auwera. Gradient compression with a variational coding scheme for federated learning. In *2021 International Conference on Visual Communications and Image Processing (VCIP)*, pages 1–5, Munich, Germany, 2021.
[143] Sahaya Beni Prathiba, Gunasekaran Raja, Sudha Anbalagan, Sugeerthi Gurumoorthy, Neeraj Kumar, and Mohsen Guizani. Cybertwin-driven federated learning based personalized service provision for 6g-v2x. *IEEE Transactions on Vehicular Technology*, 71(5):4632 – 4641, 2022.
[144] Pengfei Wang, Zhaohong Yan, Mohammad S. Obaidat, Zhiwei Yuan, Leyou Yang, Junxiang Zhang, Zongzheng Wei, and Qiang Zhang. Edge caching with federated unlearning for low-latency v2x communications. *IEEE Communications Magazine*, page 1–7, 2023.
[145] Laizhong Cui, Xiaoxin Su, Zhongxing Ming, Ziteng Chen, Shu Yang, Yipeng Zhou, and Wei Xiao. Creat: Blockchain-assisted compression algorithm of federated learning for content caching in edge computing. *IEEE Internet of Things Journal*, 9:14151–14161, 2022.
[146] Shanti Chilukuri and Dirk Pesch. Achieving optimal cache utility in constrained wireless networks through federated learning. In *2020 IEEE 21st International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM)*, pages 254–263, Cork, Ireland, 2020.
[147] Soufiane Oualil, Rachid Oucheikh, Mohamed El-Kamili, and Ismail Berrada. A personalized learning scheme for internet of vehicles caching. In *2021 IEEE Global Communications Conference (GLOBECOM)*, pages 01–06, Madrid, Spain, 2021.
[148] Wenlan Zhu, Jia Chen, Long You, Jing Chen, Xin Cheng, Kuo Guo, Chenxi Liao, and Xu Huang. A federated-cnn based proactive caching algorithm for vcdn system. In *2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML)*, pages 50–55, Hangzhou, China, 2022.
[149] Mohamad Mestoukirdi, Matteo Zecchin, David Gesbert, and Qianrui Li. User-centric federated learning: Trading off wireless resources for personalization. *IEEE Transactions on Machine Learning in Communications and Networking*, 1:346–359, 2023.
[150] Stack overflow dataset, 2017. Available: `https://storage.googleapis.com/download.tensorflow.org/data/stack_overflow_16k.tar.gz`.
[151] Yu Ji and Lan Chen. Fedqnn: A computation–communication-efficient federated learning framework for iot with low-bitwidth neural network quantization. *IEEE Internet of Things Journal*, 10(3):2494–2507, 2023.