Federated Long-Tail Learning: A Survey

www.surveyx.cn

Abstract

Federated Long-Tail Learning (FLTL) is an emerging paradigm that synergizes federated learning's privacy-preserving capabilities with strategies like classifier fine-tuning and distribution alignment to tackle the challenges of long-tail distributions. These distributions, characterized by a few classes with abundant samples and many with scarce representation, pose significant hurdles in model training, particularly in federated environments with heterogeneous client data. This survey elucidates the foundational principles of FLTL, highlighting its role in addressing data privacy and distribution imbalances. Techniques such as local embedding and Wasserstein barycenters facilitate a unified feature space across clients, enhancing model generalization. Additionally, methods like FedPipe and FedHyb optimize model performance by integrating hybrid learning strategies and edge server-based personalization. The survey further explores the application of FLTL in real-world scenarios, such as medical relation extraction, demonstrating its efficacy in managing diverse datasets. Future directions emphasize enhancing communication and computational efficiency, integrating advanced learning techniques, and expanding applications across domains. These advancements underscore FLTL's potential to revolutionize federated learning by ensuring robust, personalized, and equitable model outcomes across varied data environments.

1 Introduction

1.1 Concept of Federated Long-Tail Learning

Federated Long-Tail Learning (FLTL) addresses the intertwined challenges of data privacy and distribution imbalances prevalent in machine learning tasks. This paradigm builds upon Federated Learning (FL), a decentralized approach that enables model training across multiple clients without exchanging raw data, thus preserving user privacy and accommodating diverse data qualities [1].

The long-tail distribution problem, characterized by a few classes having numerous samples while many others are underrepresented, adversely affects model training and performance [2]. In federated settings, client data heterogeneity exacerbates this issue, leading to skewed label distributions that can bias the global model and diminish its accuracy [3]. FLTL aims to tailor the global model to individual clients' unique data distributions, enhancing overall model performance [4].

A key aspect of FLTL is establishing a unified feature space across clients with heterogeneous data characteristics, which is crucial for the global model's generalization across diverse datasets. Techniques such as classifier fine-tuning and distribution alignment are employed to address imbalanced data challenges, thereby improving model personalization and efficacy [2].

By combining the privacy-preserving benefits of FL with innovative strategies like hybrid knowledge distillation and dual-decoupling, FLTL enhances the model's generalization capability across varied datasets while minimizing collaboration bias and improving convergence stability. This approach fosters a more equitable learning environment for clients with diverse data characteristics and significantly boosts the efficiency and accuracy of the learning process, particularly in non-IID scenarios exhibiting long-tailed characteristics [5, 6, 4].

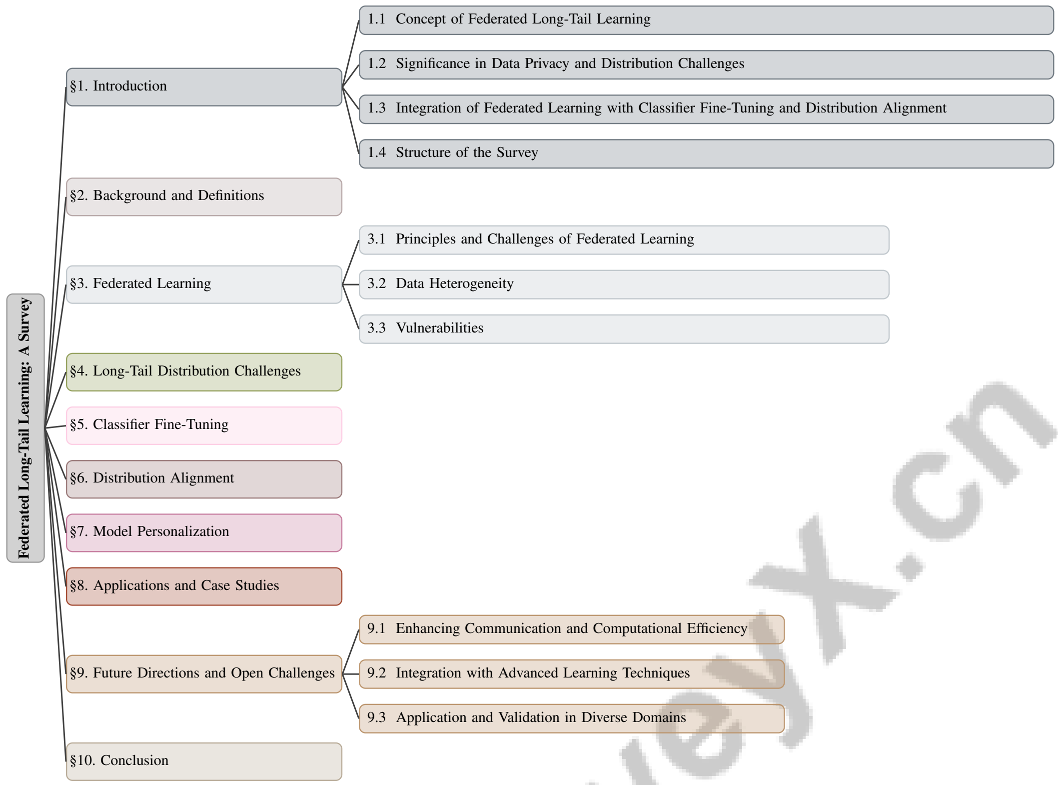


Figure 1: chapter structure

1.2 Significance in Data Privacy and Distribution Challenges

The significance of Federated Long-Tail Learning (FLTL) lies in its innovative methodology for tackling data privacy and distribution imbalances, which are pivotal in contemporary machine learning applications. A primary challenge in FL is the non-identical class distributions and imbalanced client sizes, complicating the optimization for both generic and personalized performance. The divergence of data distributions among clients further intensifies these challenges, hindering existing FL methods from effectively optimizing model performance across diverse datasets [3].

In FL contexts, the non-IID nature of data presents substantial hurdles, as label distributions can vary markedly among clients, leading to discrepancies in feature representation and model generalization [4]. This heterogeneity complicates model training and poses challenges to maintaining data privacy, as traditional methods often struggle to handle such diverse environments [4]. The long-tail distribution issue, where certain classes have very few training examples, skews model performance towards head classes, neglecting minority classes [2].

These challenges have practical implications across various domains, particularly in visual classification tasks, where performance degradation due to non-identical class distributions and imbalanced client sizes is a significant concern [1]. FLTL addresses these issues by promoting a learning process that respects data privacy while effectively managing distributional imbalances. Through strategies like classifier fine-tuning and distribution alignment, FLTL enhances model personalization and performance, ensuring that individual clients' unique data distributions are adequately considered [2]. This approach mitigates collaboration bias and stabilizes convergence, thereby improving the overall efficacy of federated learning models in handling long-tail distributions [4].

1.3 Integration of Federated Learning with Classifier Fine-Tuning and Distribution Alignment

Integrating federated learning with classifier fine-tuning and distribution alignment is essential for addressing data heterogeneity and long-tail distribution challenges in federated environments. This integration not only enhances model performance but also tailors models to individual clients' unique

AI-generated, for reference only.

data distributions. Techniques such as the FLIC framework, which utilizes local embedding functions and Wasserstein barycenters, are pivotal in aligning distributions, thereby improving federated learning outcomes [7].

Personalized federated learning algorithms leveraging a client-edge-cloud structure exemplify this integration, optimizing the learning process for each client [8]. The ReGL method effectively addresses label distribution skew by locally recovering the global data distribution, thus enhancing model performance [9].

FedCMC introduces an innovative approach by utilizing major classifier vectors from all clients to guide local training, thereby mitigating the risk of overfitting to local label distributions [10]. Frameworks like FedPipe enhance the fine-tuning of large language models by identifying critical weights and configuring low-rank adapters specific to each edge server's resources [11].

MuPFL incorporates components such as Biased Activation Value Dropout, Adaptive Cluster-based Model Update, and Prior Knowledge-assisted Classifier Fine-tuning to effectively personalize model training [6]. Additionally, FedDDC improves model accuracy by separately calibrating the feature extractor and classifier [5].

Moreover, Adversarial Federated Consensus Learning (AFedCL) facilitates personalized federated learning by dynamically adjusting the model training process according to each client's unique data distribution [12]. The unified distribution alignment strategy proposed by Zhang et al. calibrates classifier outputs to match a reference distribution, thereby enhancing performance across long-tail visual recognition tasks [2].

The FedHyb approach combines instance-based sampling for feature extraction with class-balanced sampling for classifier fine-tuning, enhancing generalization and performance on long-tailed data [4]. Collectively, these strategies underscore the critical role of integrating federated learning with classifier fine-tuning and distribution alignment, ensuring robust, personalized models capable of navigating complex real-world data distributions.

1.4 Structure of the Survey

This survey is systematically structured to provide a comprehensive overview of Federated Long-Tail Learning (FLTL), beginning with an introduction to its foundational concepts and significance in addressing data privacy and distribution challenges. Following the introduction, a detailed background elucidates core definitions and interrelations of key concepts such as federated learning, long-tail distribution, classifier fine-tuning, distribution alignment, and model personalization.

Subsequent sections explore the principles and challenges of federated learning, emphasizing data heterogeneity, non-IID challenges, and label distribution skew. The intricacies of long-tail distribution challenges and their impact on model training and performance are examined, focusing on model aggregation and the utilization of unlabeled data.

The discussion on classifier fine-tuning emphasizes innovative methods for adapting classifiers to effectively manage imbalanced data distributions, such as the Flexible Distribution Alignment (FlexDA) approach, which dynamically aligns predictions with the actual distribution of unlabeled data, and the FedCMC method, which utilizes major classifier vectors to mitigate overfitting in federated learning scenarios with heterogeneous label distributions. These strategies enhance model calibration and performance across various benchmarks while addressing the challenges posed by skewed class distributions in both semi-supervised and federated learning contexts [10, 11, 13]. An analysis of distribution alignment techniques, including local embedding and Wasserstein barycenters, as well as synthetic data generation for distribution recovery and personalization, complements this discussion.

The examination of model personalization strategies highlights various techniques for adapting a global model to individual clients' unique data distributions, emphasizing challenges posed by heterogeneous feature spaces. The FLIC framework addresses these challenges by mapping diverse data representations onto a common feature space through local embedding functions trained via distribution alignment in a federated manner. The implementation of personalized federated learning within a client-edge-cloud architecture demonstrates how edge servers can enhance model personalization by effectively mixing local and global models, utilizing learnable parameters to optimize performance, and employing a similarity aggregation method to improve collaboration among edge

AI-generated, for reference only.

servers. This approach accelerates model convergence and enhances the accuracy of personalized models in non-IID scenarios [8, 7]. The survey also reviews specific algorithms such as ESPerHFL and AFedCL, which are pivotal in personalized federated learning.

The application section provides real-world examples and case studies, illustrating successful FLTL implementations across diverse datasets and environments, while addressing necessary optimizations for heterogeneous settings.

Finally, the survey identifies future research directions and open challenges, focusing on enhancing communication and computational efficiency, integrating advanced learning techniques, and expanding applications across various domains. The conclusion synthesizes critical insights, emphasizing the pivotal role of Federated Learning (FL) in modern machine learning applications, particularly in addressing privacy concerns and optimizing the fine-tuning of large language models through innovative techniques such as automated federated pipelines and robust aggregation methods [14, 10, 15, 11].The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Background and Definitions

Federated learning (FL) is a distributed framework that facilitates deep learning across multiple clients, especially within mobile edge computing environments, while preserving data privacy by avoiding raw data exchange [9, 5]. This approach is particularly adept at addressing the challenges of non-IID data, which often result in performance degradation due to variability in client data distributions. Recent advancements, such as federated dual-decoupling and generative models, enhance model accuracy by mitigating feature extraction biases and addressing class imbalances [9, 5].

A notable challenge in FL is managing long-tail distributions, where certain classes are overrepresented while others remain underrepresented [12]. This imbalance can skew the global model towards majority classes, necessitating specialized strategies to ensure comprehensive class generalization in the global model. Classifier fine-tuning is a pivotal technique for adapting models to better handle imbalanced data, enhancing representation of all classes, including those in the tail [6]. This often involves distribution alignment to balance class representation across clients, utilizing methods like local embedding and Wasserstein barycenters [8].

Model personalization is integral to federated long-tail learning, focusing on tailoring the global model to the distinctive data distributions of individual clients. Frameworks such as MuPFL optimize model training for each client, effectively addressing non-IID and long-tailed data challenges [6]. This personalization enhances model efficacy across diverse scenarios.

In federated long-tail learning, the integration of FL with techniques like classifier fine-tuning, distribution alignment, and model personalization forms a comprehensive strategy to tackle data privacy and distribution imbalances. Methods such as Biased Activation Value Dropout, Adaptive Cluster-based Model Update, and hybrid knowledge distillation are employed to reduce overfitting, ensure coherent global aggregation, and enhance feature representation and model accuracy [11, 6, 7, 4, 3]. This approach not only improves model performance across diverse client data but also promotes equitable treatment of varied data distributions, facilitating effective deployment in practical applications.

In the realm of machine learning, Federated Learning has emerged as a pivotal approach that addresses various challenges associated with data privacy and distributed data environments. As depicted in Figure 2, this figure illustrates the hierarchical structure of Federated Learning, highlighting its principles, challenges, and solutions. It categorizes the main aspects into three primary areas: Principles and Challenges of Federated Learning, Data Heterogeneity and Non-IID Challenges, and Vulnerabilities and Security Concerns. Each category is further broken down into specific challenges and solutions, emphasizing the complexity and critical considerations in Federated Learning systems. This structured representation not only aids in understanding the multifaceted nature of Federated Learning but also serves as a guide for researchers aiming to navigate its intricacies effectively.

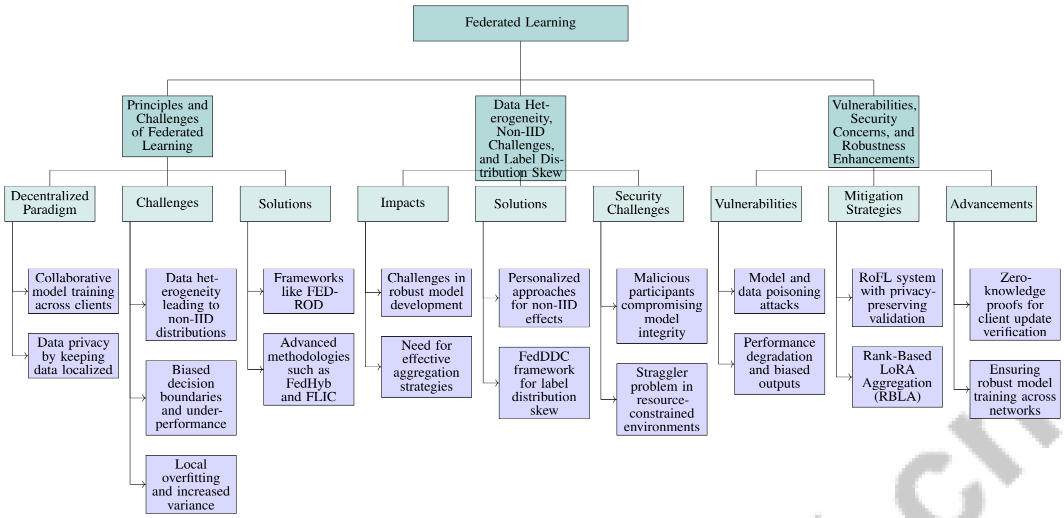


Figure 2: This figure illustrates the hierarchical structure of Federated Learning, highlighting its principles, challenges, and solutions. It categorizes the main aspects into three primary areas: Principles and Challenges of Federated Learning, Data Heterogeneity and Non-IID Challenges, and Vulnerabilities and Security Concerns. Each category is further broken down into specific challenges and solutions, emphasizing the complexity and critical considerations in Federated Learning systems.

3 Federated Learning

3.1 Principles and Challenges of Federated Learning

Federated learning (FL) is a decentralized paradigm enabling collaborative model training across multiple clients while preserving data privacy by keeping data localized and sharing only model updates [3]. This approach is particularly advantageous in privacy-sensitive environments, such as mobile edge computing [15]. A primary challenge in FL is managing data heterogeneity, which leads to non-IID data distributions across clients, causing biased decision boundaries and underperformance on minority classes [2]. This variability complicates the generalization of the global model across diverse datasets, often resulting in local overfitting and increased variance [4].

As illustrated in Figure 3, the figure highlights the primary challenges and solutions in federated learning, focusing on data heterogeneity, security threats, and performance enhancement methods. FL systems are vulnerable to attacks like model and data poisoning, which threaten model integrity and robustness [15]. To counter these, frameworks such as FED-ROD have been developed to decouple learning objectives, optimizing both generic and personalized tasks while leveraging shared knowledge [3]. Addressing local overfitting and variance is crucial for enhancing FL's real-world applicability [4]. Advanced methodologies, including FedHyb and FLIC, aim to improve model performance by managing distributed data complexities while maintaining privacy [10, 15, 9, 7, 4].

3.2 Data Heterogeneity, Non-IID Challenges, and Label Distribution Skew

Data heterogeneity and non-IID challenges are significant in FL, impacting the development of robust models. Diverse data across clients necessitates effective aggregation strategies to handle non-IID distributions [7]. Personalized approaches are crucial for mitigating non-IID effects, tailoring models to client-specific data characteristics [8]. Label distribution skew, where certain classes dominate, biases decision boundaries. The FedDDC framework addresses this by decoupling model components for targeted calibration, ensuring minority class representation [5].

The open nature of FL systems introduces security challenges, as malicious participants can compromise model integrity [15]. Additionally, the straggler problem, particularly in resource-constrained environments, poses significant challenges by delaying model updates, necessitating strategies to mitigate these effects [11].

AI-generated, for reference only.

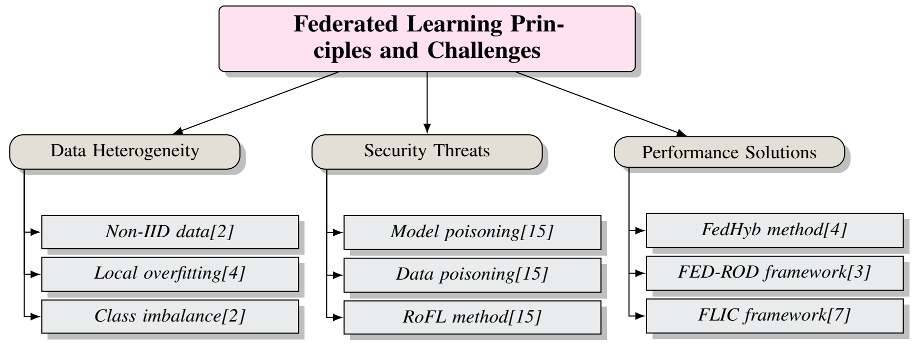


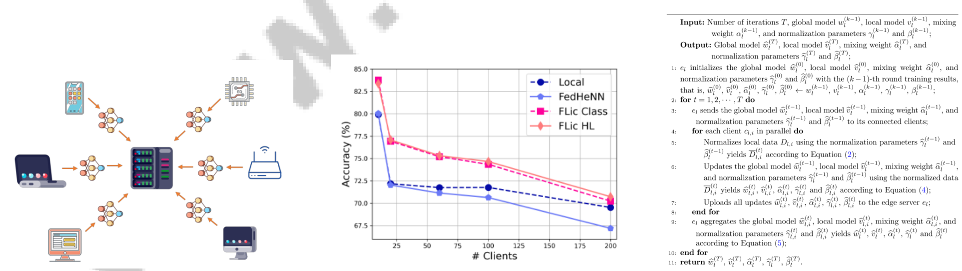
Figure 3: This figure illustrates the primary challenges and solutions in federated learning, focusing on data heterogeneity, security threats, and performance enhancement methods.

3.3 Vulnerabilities, Security Concerns, and Robustness Enhancements

FL systems, while offering privacy advantages, face vulnerabilities like model and data poisoning attacks that exploit client-generated updates. Robust mitigation strategies are essential, such as the RoFL system, which uses privacy-preserving validation and norm bounds on model updates to protect training integrity [9, 14, 15, 11]. These attacks can degrade performance or bias outputs, necessitating innovative methods to enhance robustness and security.

Rank-Based LoRA Aggregation (RBLA) improves global model convergence by aggregating models of differing ranks without structural sparsity [14]. RoFL further enhances security by verifying client updates with zero-knowledge proofs, ensuring only legitimate contributions are integrated [15]. These advancements address security vulnerabilities, ensuring robust model training across diverse and adversarial networks [14, 10, 11, 15, 9].

As illustrated in Figure 5, this figure illustrates the key vulnerabilities and enhancements in Federated Learning (FL) systems, highlighting model and data poisoning as primary vulnerabilities, and presenting RoFL and RBLA as significant security and robustness strategies. The visualizations include a networked system, comparative accuracy analysis across methods and client sizes, and a flowchart of a distributed training algorithm, emphasizing the complexity and critical security considerations in FL systems [14, 7, 8].



(a) A Networked System[14]

(b) Comparison of Accuracy Across Different Federated Learning Methods and Client Sizes[7]

(c) The image shows a flowchart for a distributed training algorithm.[8]

Figure 4: Examples of Vulnerabilities, Security Concerns, and Robustness Enhancements

4 Long-Tail Distribution Challenges

4.1 Long-Tail Distribution and Its Impact

Long-tail distributions significantly impact model training in federated learning by causing class imbalances due to uneven data distribution among clients. This imbalance, characterized by a few

AI-generated, for reference only.

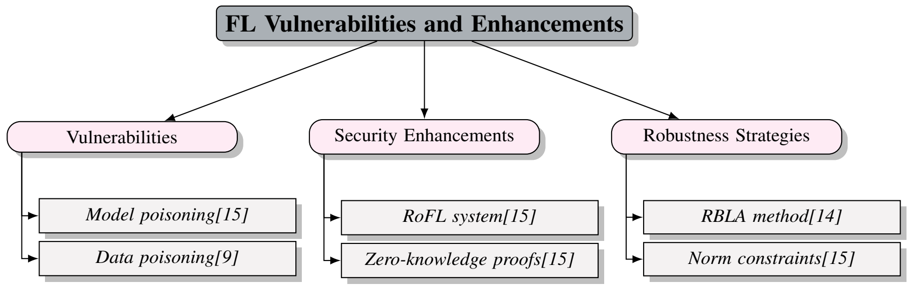


Figure 5: This figure illustrates the key vulnerabilities and enhancements in Federated Learning (FL) systems, highlighting model and data poisoning as primary vulnerabilities, and presenting RoFL and RBLA as significant security and robustness strategies.

classes with many samples and many classes with few samples, complicates classifier training and introduces biases, particularly in semi-supervised learning. Strategies like Flexible Distribution Alignment (FlexDA) and Class-Specific Distribution Alignment (CSDA) dynamically align predictions with the actual distribution of unlabeled data, ensuring balanced class representation throughout training [9, 16, 13]. These methods counteract the tendency of deep learning models to favor majority classes, thereby enhancing performance and generalization for minority classes.

Du et al. propose a method using major classifier vectors, selecting optimal class vectors from clients instead of weighted averages, enhancing robustness against skewed distributions and improving model calibration [10]. Aimar et al. show that FlexDA significantly enhances model performance in long-tail semi-supervised learning by addressing label distribution shifts and class imbalances, leading to better model calibration [13]. Similarly, Zhang et al.'s DisAlign framework uses class priors and data inputs to align model predictions with balanced distributions, mitigating the adverse effects of long-tail distributions [2].

Fang et al. demonstrate the effectiveness of FedPipe in fine-tuning large language models within federated learning, reducing training time and improving accuracy while addressing long-tail distribution challenges [11]. Additionally, FedHyb employs a hybrid knowledge distillation approach that coordinates various sampling strategies within a two-stage learning framework, enhancing convergence speed and model accuracy [4].

4.2 Challenges in Model Aggregation and Utilizing Unlabeled Data

Model aggregation in federated learning, especially with long-tail distributions, presents challenges such as the introduction of non-informative values through zero-padding, impairing feature representation and generalization across diverse client datasets [14]. Effectively utilizing unlabeled data, abundant in real-world scenarios, is crucial for enhancing model performance. Aligning the distribution of unlabeled data with labeled data is essential to avoid biased pseudo-labels and neglecting underrepresented classes, which leads to poorly calibrated classifiers. The FlexDA framework addresses these issues by dynamically aligning predictions with the true distribution of unlabeled data, incorporating a distillation-based consistency loss to ensure fair data usage across classes, thus improving performance in imbalanced scenarios [2, 13].

Incorporating unlabeled data must be managed to prevent overfitting to majority classes. Semi-supervised and active learning techniques can strategically integrate unlabeled data to enhance performance, particularly through FlexDA and CSDA. FlexDA improves classifier calibration across diverse benchmarks by dynamically aligning predictions with the actual distribution of unlabeled data. CSDA captures class-dependent marginal predictions, ensuring a balanced selection of unlabeled samples, leading to improved outcomes in semi-supervised tasks involving imbalanced datasets [16, 13]. These methods emphasize the most informative examples, helping balance class representation and enhancing the model's generalization capabilities.

To illustrate these concepts, Figure 6 depicts the primary challenges in model aggregation and the utilization of unlabeled data within federated learning, focusing on non-informative values, data distribution alignment, and improving feature extraction. Addressing these challenges is crucial for improving federated learning systems in long-tail distribution contexts, particularly concerning collaboration bias, unstable convergence, and the impact of heterogeneous client data, which can degrade performance. Advanced techniques like hybrid knowledge distillation and dual-decoupling approaches enhance feature extraction and classifier fine-tuning, resulting in more robust and accurate global models without compromising client privacy or increasing vulnerability to attacks [5, 4]. By developing robust aggregation strategies and effectively incorporating unlabeled data, models can achieve better generalization and equitable performance across all classes.

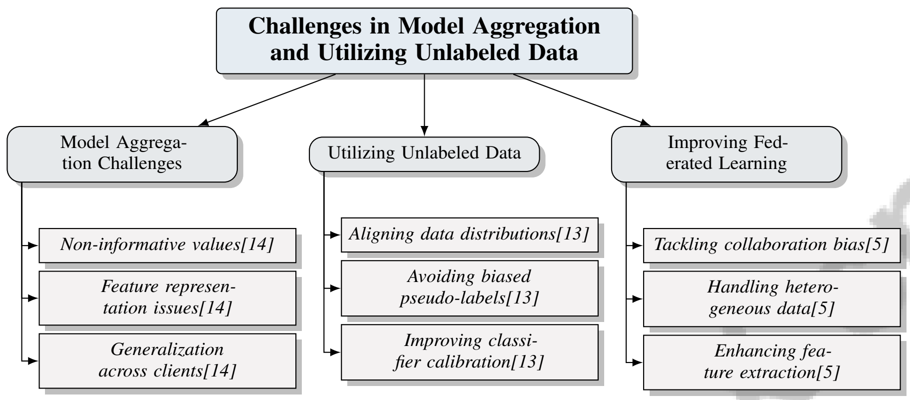


Figure 6: This figure illustrates the primary challenges in model aggregation and the utilization of unlabeled data within federated learning, focusing on non-informative values, data distribution alignment, and improving feature extraction.

5 Classifier Fine-Tuning

5.1 Classifier Fine-Tuning and Distribution Alignment

Method Name	Model Calibration	Data Distribution	Personalization Techniques
MuPFL[6]	Model Updates	Distribution Shifts	Personalized Federated Learning
FedDDC[5]	Model Calibration	Distribution Shifts	Confidence Re-weighting
FlexDA[13]	Improving Model Calibration	Align Model Predictions	-
CSDA[16]	Change OF Basis	Distribution Alignment	-
ReGL[9]	Adaptively Fine-tuning	Global Data Distribution	Local Personalization Tasks
FedHyb[4]	Ensemble Distillation	Class-balanced Sampling	Personalized Federated Learning

Table 1: Comparison of Federated Learning Methods: This table presents a detailed analysis of various federated learning methods, focusing on their model calibration strategies, data distribution handling, and personalization techniques. The comparison highlights the diverse approaches to address data imbalances and enhance model performance in non-IID scenarios.

In federated learning, classifier fine-tuning and distribution alignment are essential for addressing data imbalances and non-IID challenges, enhancing global model performance by integrating diverse client data. Techniques such as client confidence re-weighting and logit calibration significantly improve model accuracy in skewed data scenarios [11, 13, 5, 9, 7]. Fine-tuning adjusts model parameters to better capture tail class complexities, as demonstrated by Zhang et al.'s multi-level personalized federated learning approach, which mitigates overfitting and refines model updates for improved tail class accuracy [6]. Wang et al. highlight the importance of calibrating feature extractors and classifiers to enhance global model accuracy in non-IID and long-tailed contexts [5].

Distribution alignment ensures that predictions reflect true data distributions for both labeled and unlabeled data. FlexDA, introduced by Aimar et al., uses an adaptive logit-adjusted loss framework to align predictions with the actual distribution of unlabeled data, effectively addressing label distribution shifts [13]. Huang et al.'s CSDA further refines accuracy by aligning predictions in a class-specific manner [16]. Synthetic data generation techniques, such as ReGL, employ generative models to create synthetic images fine-tuned with local data, enhancing class representation and model generalization [9].

FedHyb's two-stage learning process, combining client-level self-distillation and server-level ensemble distillation, enhances model calibration and performance, ensuring robustness against skewed distributions [4]. Table 1 provides a comprehensive comparison of different federated learning methods, illustrating their approaches to classifier fine-tuning and distribution alignment in the context of non-IID data challenges.

5.2 Client Confidence Re-weighting and Classifier Re-balancing

Client confidence re-weighting and classifier re-balancing are crucial for improving model accuracy in federated learning, particularly with imbalanced and non-IID data. These methods mitigate biases from imbalanced class distributions, enhancing global model generalization across client datasets. Techniques like local data synthesis for minority and missing classes, alongside adaptive calibration for balanced classification scores, ensure model robustness while maintaining data privacy. Their effectiveness in overcoming label imbalance is evidenced in applications such as image classification and medical relation extraction [9, 10, 2].

The FedDDC approach exemplifies these techniques by decoupling the feature extractor and classifier, enabling targeted calibration that improves representation learning and decision boundaries [5]. This decoupling addresses imbalances from skewed label distributions, ensuring adequate minority class representation in the global model.

FlexDA enhances these strategies by dynamically adjusting predictions to align with unlabeled data distribution, offering a flexible solution to distribution alignment challenges without assuming uniform distributions or requiring prior knowledge [13]. By aligning predictions with the true distribution, it boosts model calibration and accuracy, particularly in long-tail semi-supervised learning scenarios.

Incorporating client confidence re-weighting and classifier re-balancing is vital for improving federated learning systems' performance, especially in non-IID and long-tailed data contexts. These strategies optimize client contribution weighting, enhancing calibration and reducing bias in the global model, leading to robust and equitable outcomes across diverse client data scenarios [15, 5, 9, 7, 4]. By effectively navigating real-world data complexities, these techniques contribute to improved accuracy and generalization across all classes.

6 Distribution Alignment

6.1 Local Embedding, Wasserstein Barycenters, and Change of Basis

Local embedding functions and Wasserstein barycenters are pivotal in federated learning for managing distribution alignment, particularly within heterogeneous client environments. These methods enable the transformation of diverse client data into a unified feature space, enhancing collaborative learning and mitigating issues like label imbalance and overfitting [9, 10, 2, 7]. The FLIC framework exemplifies this by using local embeddings to integrate updates from diverse clients into a common feature space. Wasserstein barycenters offer a robust mathematical basis for aligning client data distributions with global model expectations.

Secure aggregation mechanisms, such as RoFL, further reinforce federated learning systems by employing zero-knowledge proofs to ensure client updates adhere to norm constraints, thereby protecting against targeted attacks and enhancing learning integrity [15]. This integration of secure aggregation with advanced distribution alignment techniques is crucial for maintaining model reliability, especially in adversarial environments.

Combining local embedding, Wasserstein barycenters, and secure aggregation offers a comprehensive strategy for tackling distribution alignment complexities in federated learning. Techniques like Rank-Based LoRA Aggregation (RBLA) enable efficient model fine-tuning in heterogeneous contexts, while FedCMC uses major classifier vectors to counteract overfitting in local label distributions. Methods such as ReGL allow clients to synthesize data that addresses label imbalances, ensuring equitable model performance across diverse client scenarios [9, 14, 10].

6.2 Synthetic Data Generation for Distribution Recovery and Personalization

Synthetic data generation is crucial in federated learning for distribution recovery and model personalization. This technique produces synthetic data that mirrors the statistical traits of actual datasets, enhancing model generalization across diverse client datasets, particularly under data privacy con-

AI-generated, for reference only.

straints and heterogeneous resource conditions [14, 10, 11, 4]. It is instrumental in recovering balanced distributions and facilitating personalization, especially amidst non-IID and long-tail data distributions.

The ReGL method demonstrates the efficacy of synthetic data generation by using foundational generative models to create class-specific synthetic images, which are fine-tuned with local data to align with global distributions [9]. This process balances class representation and improves the model's generalization across diverse distributions, enhancing performance and fairness.

Additionally, synthetic data generation supports model personalization by creating tailored datasets that reflect individual client data characteristics. This ensures the global model adapts to each client's specific needs, addressing non-IID data and long-tail distribution challenges. By generating synthetic data that aligns with local distributions, federated learning systems can tackle label imbalances, enhance accuracy, and promote equitable outcomes across diverse client populations. This involves using generative models to fill gaps in minority and missing classes, ensuring balanced global data representation while preserving privacy [1, 11, 9, 7, 4].

Synthetic data generation is thus essential in federated learning, providing a robust mechanism for recovering balanced distributions and facilitating personalization. By accurately reflecting global class distributions, federated models can address label imbalance and data heterogeneity challenges, enhancing performance, fairness, and privacy. Comprehensive experiments show significant improvements in accuracy and stability, leading to more equitable outcomes in real-world applications [9, 1, 6, 7].

7 Model Personalization

7.1 Model Personalization and Adaptation Techniques

In federated learning, personalization and adaptation are essential to address the challenges of heterogeneous and non-IID data. Variations in client data necessitate personalized strategies, such as the Federated Robust Decoupling framework, which balances generic and personalized model performance, and the Multi-level Personalized Federated Learning framework, which addresses class imbalances and biases. Techniques like local embedding functions and major classifier vectors further enhance adaptability, ensuring resilience and efficiency across applications [10, 11, 6, 7, 3]. These strategies tailor the global model to individual client data distributions, enhancing performance and personalization.

The FLIC framework exemplifies robust personalization by embedding client features into a common latent space, facilitating collaboration despite data heterogeneity [7]. This shared space enables the global model to accurately interpret diverse client updates, improving personalization and collaboration.

Hierarchical model architectures are crucial for personalizing federated learning models, customizing components at various levels to represent each client's unique data distribution. Strategies like Biased Activation Value Dropout (BAVD) reduce overfitting, Adaptive Cluster-based Model Update (ACMU) ensures coherent global model aggregation, and Prior Knowledge-assisted Classifier Fine-tuning (PKCF) adapts models to local data characteristics. These methods significantly enhance accuracy and efficiency, achieving up to 7.39

Personalization and adaptation techniques are vital for robust and equitable federated learning performance. By leveraging sophisticated techniques like feature embedding and hierarchical architectures, federated learning systems effectively tackle challenges posed by heterogeneous and long-tailed data distributions. Frameworks like FLIC and MuPFL implement strategies like distribution alignment and biased activation value dropout to mitigate overfitting and optimize training efficiency, leading to performance improvements with accuracy gains of up to 7.39

7.2 Heterogeneous Feature Space Handling and Edge Server-Based Personalization

Handling heterogeneous feature spaces and utilizing edge server-based personalization are crucial for model adaptation in federated learning. Client data variability often results in heterogeneous feature spaces, complicating unified global model creation. Many federated learning approaches assume a common data schema, which is rare in real-world applications with diverse data representations.

Recent methodologies, such as personalized federated learning frameworks, map client data into a shared feature space through local embedding functions, facilitating model training despite data heterogeneity. Techniques like hybrid knowledge distillation and generative model-based approaches enhance performance and convergence by addressing label imbalance and optimizing local updates, ensuring the global model's effectiveness across varied client data distributions [10, 9, 7, 4, 3].

Adaptive model architectures that dynamically adjust to each client's data features are essential for managing heterogeneous feature spaces. This adaptability ensures effective integration and generalization across diverse datasets, improving model performance and personalization. Techniques like local embedding functions and hierarchical model architectures facilitate alignment of heterogeneous feature spaces, enabling accurate interpretation and incorporation of client-specific data characteristics [7].

Edge server-based personalization enhances adaptation by enabling localized processing and customization of model components. Edge servers act as intermediaries between the central server and client devices, efficiently handling data and model updates. This setup reduces latency and computational load on client devices while allowing for personalized model updates tailored to each client's unique data distributions. By offloading computational tasks to edge servers, federated learning systems achieve more efficient and scalable personalization, improving the learning process's overall efficacy [11].

The combination of adaptive model architectures and edge server-based personalization represents a comprehensive approach to managing heterogeneous feature spaces in federated learning. These strategies enhance robustness and accuracy while enabling personalized predictions across varied client environments. By addressing complexities in real-world data distributions, such as heterogeneous feature spaces and long-tailed class distributions, these approaches significantly improve federated learning performance. Frameworks like FLIC utilize local embedding functions to align diverse data representations, while MuPFL employs advanced techniques to mitigate overfitting and refine model updates. Methods like FedHyb leverage two-stage learning paradigms to enhance feature generalization and classifier fine-tuning, ensuring stability and robustness against collaboration bias and data heterogeneity. Collectively, these innovations contribute to more effective and efficient federated learning outcomes [6, 4, 10, 7].

7.3 ESPerHFL Algorithm and AFedCL in Personalized Federated Learning

The ESPerHFL algorithm advances personalized federated learning by focusing on model creation at the edge server level. This approach leverages edge servers' capabilities to facilitate personalized model training, efficiently handling diverse client data distributions. By implementing personalization at the edge, ESPerHFL enhances federated learning systems' scalability and adaptability, ensuring models are tailored to each client's specific needs [8].

The Adversarial Federated Consensus Learning (AFedCL) framework offers a robust mechanism for personalized federated learning by dynamically adjusting the model training process according to each client's unique data distribution. AFedCL employs advanced adversarial learning techniques to foster consensus among clients, effectively addressing challenges posed by data heterogeneity. This approach ensures that the global model accommodates diverse data characteristics while enhancing personalization by aligning local models with the global framework. By utilizing a consensus-aware aggregation mechanism and adaptive feature fusion, AFedCL balances global knowledge with local data nuances, significantly improving accuracy and generalization capabilities across various applications, such as industrial surface defect classification [9, 6, 12, 7]. This method enhances accuracy and mitigates risks associated with non-IID data and long-tail distributions, leading to improved performance and fairness.

The ESPerHFL and AFedCL algorithms illustrate the transformative potential of advanced methodologies in personalized federated learning, particularly in addressing challenges posed by heterogeneous feature spaces and non-IID data distributions. By employing innovative techniques like local embedding functions and hierarchical model architectures, these algorithms enhance accuracy and training efficiency, demonstrating significant improvements even under challenging conditions like long-tailed data distributions [6, 7]. By focusing on edge server-based model creation and adversarial consensus strategies, these approaches effectively address real-world data distribution complexities, ensuring federated learning systems deliver accurate, personalized, and equitable outcomes across diverse client datasets.

8 Applications and Case Studies

8.1 Real-World Dataset Applications

Federated Long-Tail Learning (FLTL) demonstrates significant efficacy in real-world applications, especially in the medical domain for tasks like relation extraction. Evaluations on datasets such as the 2010 i2b2/VA challenge, CPR, and PGR highlight FLTL's ability to manage heterogeneous and imbalanced data, typical in medical contexts [10]. FLTL outperforms baseline federated learning algorithms by effectively handling long-tail distributions and non-IID data, crucial for medical applications where some conditions are underrepresented. Techniques like major classifier vectors ensure robust generalization across diverse medical datasets [10].

Federated Learning (FL) technology showcases its broader potential in navigating complex real-world data environments, where data privacy is crucial. FL allows model training on sensitive data without direct access, minimizing privacy risks. Innovations such as the RoFL system enhance FL's resilience against attacks by constraining model updates, while automated pipelines like FedPipe streamline the fine-tuning of large language models across edge servers, addressing computational and communication challenges [15, 11]. The success of FLTL in medical relation extraction underscores its promise for accurate and equitable outcomes, paving the way for adoption in other domains with similar data characteristics.

8.2 Optimizations and Adaptations for Heterogeneous Environments

Implementing federated learning in heterogeneous environments requires optimizations and adaptations to manage diverse, non-uniform data distributions across clients. A key challenge is efficient model aggregation and communication, given client resource variability and network conditions. Techniques like FedPipe address these issues by reducing training time and enhancing model accuracy through strategic fine-tuning of large language models [11]. This pipeline architecture adapts well to the constraints of heterogeneous environments.

Hybrid knowledge distillation methods, such as FedHyb, further optimize federated learning by coordinating sampling strategies within a two-stage learning framework. This approach improves convergence speed and model accuracy by integrating self-distillation at the client side and ensemble distillation at the server side [4]. These hybrid strategies effectively manage heterogeneous data distributions, ensuring models generalize well across diverse client datasets.

Edge server-based personalization is crucial for optimizing federated learning systems in heterogeneous settings. Offloading computational tasks to edge servers reduces latency and computational load on client devices, facilitating efficient and scalable model personalization [11]. This configuration enhances federated learning systems' adaptability, ensuring models are tailored to each client's specific needs, addressing challenges related to non-IID data and long-tail distributions.

Effective federated learning implementation in heterogeneous environments necessitates a multi-faceted approach, incorporating strategic optimizations such as local embedding functions to map diverse client data into a unified feature space, and adaptations like hybrid knowledge distillation techniques to improve model accuracy and convergence speed while mitigating collaboration bias and privacy concerns [4, 11, 7]. By leveraging advanced methodologies, including pipeline architectures, hybrid knowledge distillation, and edge server-based personalization, federated learning systems can achieve robust and equitable performance across diverse client environments.

9 Future Directions and Open Challenges

The advancement of federated learning (FL) relies on enhancing communication and computational efficiency, crucial for its scalability and practical deployment. This section delves into innovative strategies to address existing challenges and improve system performance.

AI-generated, for reference only.

9.1 Enhancing Communication and Computational Efficiency

Improving communication and computational efficiency is vital for scaling federated learning (FL) systems. The Rank-Based LoRA Aggregation (RBLA) offers a promising avenue for enhancing performance in heterogeneous environments by integrating with other federated methods [14]. The FlexDA framework addresses label distribution shifts and can be extended to complex visual tasks by incorporating unknown classes absent in labeled datasets, thus increasing real-world applicability [13]. Personalized federated learning must focus on minimizing communication costs, with future research directed at algorithms that reduce data exchange between clients and servers to enhance learning efficiency [8]. The FedCMC approach holds potential for improving communication efficiency across domains with heterogeneous data distributions [10].

Integrating split learning with frameworks like FedPipe could further improve federated fine-tuning scalability, particularly for large language models, by better distributing computational loads [11]. Optimizing dynamic consensus construction in Adversarial Federated Consensus Learning (AFedCL) could enhance model robustness in heterogeneous environments by refining consensus mechanisms among diverse client datasets [12].

Generative model adaptation advancements, such as ReGL, are crucial for improving adaptability across domains and datasets, thereby enhancing utility in FL contexts [9]. Refining alignment mechanisms in frameworks like FLIC and exploring alternative embedding strategies can significantly boost communication and computational efficiency, contributing to more robust FL systems [7].

Future research should investigate enhancements in the distillation process and the application of FedHyb in complex real-world scenarios to improve communication efficiency and scalability [4]. Exploring norm constraints adaptability in RoFL could enhance defenses against adversarial attacks, maintaining the learning process's integrity and efficiency [15]. These directions emphasize optimizing communication and computational efficiency to advance federated learning technologies.

9.2 Integration with Advanced Learning Techniques

Integrating advanced learning techniques with federated learning (FL) can significantly enhance model performance and adaptability in diverse data environments. Adversarial learning, as demonstrated by the Adversarial Federated Consensus Learning (AFedCL) framework, employs adversarial strategies to ensure client consensus, allowing the global model to incorporate diverse data characteristics while maintaining high personalization and accuracy [12].

Generative models, as seen in the ReGL method, leverage foundation generative models to create synthetic images for each class, facilitating distribution recovery and enhancing model generalization in non-IID and long-tail data scenarios [9]. Further exploration of generative model adaptability across various domains could expand their utility in FL.

Knowledge distillation techniques, exemplified by FedHyb, present significant potential for improving model calibration and performance. This approach employs a two-stage learning process, incorporating self-distillation at the client side and ensemble distillation at the server side, enhancing convergence speed and accuracy, particularly in heterogeneous data distributions [4].

Integrating secure aggregation mechanisms like RoFL, which utilizes zero-knowledge proofs to ensure client update authenticity, can bolster the robustness and security of federated learning systems [15]. This integration is essential for maintaining the learning process's integrity, especially in adversarial settings.

The integration of advanced techniques such as adversarial learning, generative models for data synthesis, hybrid knowledge distillation, and secure aggregation mechanisms holds significant promise for enhancing the efficacy and scalability of FL systems. These innovations address challenges like label imbalance and data heterogeneity, enabling effective adaptation to complex data environments while preserving user privacy and improving convergence speed. For example, generative models can synthesize missing classes to mitigate label imbalance, while hybrid knowledge distillation enhances feature learning and classifier fine-tuning in heterogeneous data scenarios [9, 1, 6, 4].

9.3 Application and Validation in Diverse Domains

Federated learning (FL) offers transformative potential for application and validation across diverse domains, particularly where data privacy and distribution challenges are critical. Future research should enhance algorithms to manage complex distribution shifts, expanding FL's applicability in sectors such as healthcare, finance, and autonomous systems, where data heterogeneity and privacy concerns are paramount [1].

In healthcare, integrating clinical knowledge into semi-supervised learning frameworks can improve prediction interpretability, enhancing decision-making processes and leading to better patient outcomes [16]. FL's adaptability to extreme imbalances, as demonstrated in visual recognition tasks, indicates its potential in fields like natural language processing, where linguistic feature distributions can vary significantly [2].

The ongoing application and validation of federated learning across various domains necessitate extensive research to tackle distinct challenges, such as heterogeneous label distributions in medical data and managing the complexities of fine-tuning large language models while preserving data privacy [10, 11, 9, 7, 4]. By enhancing algorithmic adaptability and integrating domain-specific knowledge, FL can achieve broader applicability and deliver significant benefits across various sectors.

10 Conclusion

The exploration of Federated Long-Tail Learning (FLTL) within this survey underscores its critical role in addressing the dual challenges of data privacy and distribution imbalances that are pervasive in current machine learning paradigms. By integrating federated learning with sophisticated techniques such as classifier fine-tuning and distribution alignment, FLTL mitigates biases introduced by non-IID data and long-tail distributions, thereby enhancing model personalization and performance. Techniques like DisAlign have demonstrated superior results in visual recognition tasks by effectively managing long-tail distribution issues, outperforming previous methodologies.

The robustness of federated learning frameworks is further bolstered by innovative security enhancements, such as the application of zero-knowledge proofs in RoFL, which fortifies model integrity against adversarial threats while maintaining performance levels. Furthermore, the AFedCL technique showcases notable accuracy improvements, highlighting the potential of adversarial learning strategies in the realm of personalized federated learning.

Moreover, the Rank-Based LoRA Aggregation (RBLA) method successfully navigates the complexities of aggregating diverse models in Federated Learning as a Service (FLaaS) environments, resulting in improved convergence rates and overall performance. These developments collectively represent significant strides in enhancing the efficiency, scalability, and security of federated learning systems, reflecting the dynamic and evolving nature of research in this field.

References

[1] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Federated visual classification with real-world data distribution. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 76–92. Springer, 2020.

[2] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition, 2021.

[3] Hong-You Chen and Wei-Lun Chao. On bridging generic and personalized federated learning for image classification. *arXiv preprint arXiv:2107.00778*, 2021.

[4] Senbin Liu, Yuanting Zhang, Kunhua Zhang, and Yi Wang. Federated learning with hybrid knowledge distillations on long-tailed heterogeneous client data. In *ECAI 2024*, pages 2878–2885. IOS Press, 2024.

[5] Zhaohui Wang, Hongjiao Li, Jinguo Li, Renhao Hu, and Baojin Wang. Federated learning on non-iid and long-tailed data via dual-decoupling. *Frontiers of Information Technology & Electronic Engineering*, 25(5):728–741, 2024.

[6] Rongyu Zhang, Yun Chen, Chenrui Wu, Fangxin Wang, and Bo Li. Multi-level personalized federated learning on heterogeneous and long-tailed data, 2024.

[7] Alain Rakotomamonjy, Maxime Vono, Hamlet Jesse Medina Ruiz, and Liva Ralaivola. Personalised federated learning on heterogeneous feature spaces, 2023.

[8] Chunmei Ma, Xiangqian Li, Baogui Huang, Guangshun Li, and Fengyin Li. Personalized client-edge-cloud hierarchical federated learning in mobile edge computing. *Journal of Cloud Computing*, 13(1):161, 2024.

[9] Ziyu Yao. Recovering global data distribution locally in federated learning. *arXiv preprint arXiv:2409.14063*, 2024.

[10] Chunhui Du, Hao He, and Yaohui Jin. Contrast with major classifier vectors for federated medical relation extraction with heterogeneous label distribution. *Applied Intelligence*, 53(23):28895–28909, 2023.

[11] Zihan Fang, Zheng Lin, Zhe Chen, Xianhao Chen, Yue Gao, and Yuguang Fang. Automated federated pipeline for parameter-efficient fine-tuning of large language models. *arXiv preprint arXiv:2404.06448*, 2024.

[12] Jixuan Cui, Jun Li, Zhen Mei, Yiyang Ni, Wen Chen, and Zengxiang Li. Adversarial federated consensus learning for surface defect classification under data heterogeneity in iiot, 2024.

[13] Emanuel Sanchez Aimar, Nathaniel Helgesen, Yonghao Xu, Marco Kuhlmann, and Michael Felsberg. Flexible distribution alignment: Towards long-tailed semi-supervised learning with proper calibration, 2024.

[14] Shuaijun Chen, Omid Tavallaie, Niousha Nazemi, and Albert Y Zomaya. Rbla: Rank-based-lora-aggregation for fine-tuning heterogeneous models in flaas. In *International Conference on Web Services*, pages 47–62. Springer, 2024.

[15] Hidde Lycklama, Lukas Burkhalter, Alexander Viand, Nicolas Küchler, and Anwar Hithnawi. Rofl: Robustness of secure federated learning. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 453–476. IEEE, 2023.

[16] Zhongzheng Huang, Jiawei Wu, Tao Wang, Zuoyong Li, and Anastasia Ioannou. Class-specific distribution alignment for semi-supervised medical image classification, 2023.