Modular Federated Learning for Non-IID Data

By

Samuel Hailemariam Seifu

Submitted to the School of Information Technology and Engineering

In partial fulfillment of the requirements for the degree of

Master of Science in Artificial Intelligence

Supervised by: Dr. Beakal Gizachew

Addis Ababa Institute of Technology

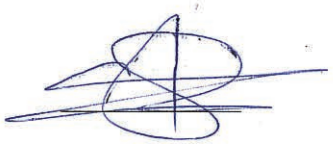
Addis Ababa University

Addis Ababa, Ethiopia

June , 2025

APPROVAL

This is to certify that the thesis titled **Modular Federated Learning for Non-IID Data** is prepared by Samuel Hailemariam Seifu and submitted in partial fulfillment of thesis-option requirements for the Degree of Masters of Science in Artificial Intelligence at School of Information Technology & Engineering, College of Technology and Built Environment, Addis Ababa University.

Name	Signature	Date
_____	_____	_____
(Advisor)		
Dr. MESFIN ABEBE		29/06/25
(External Examiner)		
_____	_____	_____
(Internal Examiner)		
_____	_____	_____
(Chairperson)		

ABSTRACT

Federated Learning (FL) promises privacy-preserving collaboration across distributed clients but is hampered by three key challenges: severe accuracy degradation under non-IID data, high communication and computational demands on edge devices, and a lack of built-in explainability for debugging, user trust, and regulatory compliance. To bridge this gap, we propose two modular FL pipelines—SPATL-XL and SPATL-XLC—that integrate SHAP-driven pruning with, in the latter, dynamic client clustering. SPATL-XL applies SHAP-based pruning to the largest layers, removing low-impact parameters to both reduce model size and sharpen interpretability, whereas SPATL-XLC further groups clients via lightweight clustering to reduce communication overhead and smooth convergence in low-bandwidth, high-client settings. In experiments on CIFAR-10 and Fashion-MNIT over 200 communication rounds under IID and Dirichlet non-IID splits, our pipelines lower per-round communication to 13.26 MB, speed up end-to-end training by 1.13×, raise explanation fidelity from 30–50% to 89%, match or closely approach SCAFFOLD's 70.64% top-1 accuracy (SPATL-XL: 70.36%), and maintain stable clustering quality (Silhouette, CHI, DBI) even when only 40–70% of clients participate. These results demonstrate that combining explainability-driven pruning with adaptive clustering yields practical, communication-efficient, and regulation-ready FL pipelines that simultaneously address non-IID bias, resource constraints, and transparency requirements.

Key words: Federated Learning, Non-IID Data, Explainable AI (XAI), SHAP, Client Clustering, Model Pruning, Communication Efficiency, SPATL-XL, SPATL-XLC.

Acknowledgements

First and foremost, I would like to thank **God** for His guidance and strength throughout every moment of completing this study.

I sincerely thank my **advisor, Dr. Beakal Gizachew**, for his invaluable guidance, feedback, and unwavering support. His extensive support was instrumental in the completion of this dissertation.

I would like to express my deepest gratitude to my **family** for their endless love, patience and encouragement. Your unwavering belief in my dreams has been a driving force and source of comfort.

Finally, I extend my gratitude to all who contributed directly or indirectly to this thesis by providing support, resources, and thoughtful input.

Dedication

To my family.

. . .

To my academic advisor.

. . .

To my self.

. . .

Table of contents

List of Abbreviations and Terms

AI	Artificial Intelligence
CIFAR-10	Canadian Institute for Advanced Research – 10 classes
ECE	Expected Calibration Error
FedAvg	Federated Averaging
FedProx	Federated Proximal
FL	Federated Learning
GDPR	General Data Protection Regulation
GPU	Graphical Processing Unit
Grad-CAM	Gradient-weighted Class Activation Mapping
IID	Independent and Identically Distributed
KPMG	Klynveld Peat Marwick Goerdeler, a global professional services network
LIME	Local Interpretable Model-Agnostic Explanations
non-IID	Non-Independent and Identically Distributed
RL	Reinforcement Learning
RTX	Ray Tracking
SCAFFOLD	Variance-Reduced FL algorithm (control variates)
SHAP	SHapley Additive exPlanations
SPATL	Salient Parameter Aggregation and Transfer Learning
SPATL-X	SPATL with explainable learning
SPATL-XL	SPATL with SHAP pruning the largest layers
SPATL-XLC	SPATL with SHAP pruning and client clustering
VRAM	Video Random Access Memory
XAI	Explainable Artificial Intelligence

List of Figures

List of Tables

Chapter One

1. Introduction

1.1 Background

Federated Learning (FL) is a privacy-preserving paradigm in which multiple clients train local models and share only parameter updates. By eliminating any transfer of raw data, FL mitigates the risk of breaches and supports compliance with regulations such as the General Data Protection Regulation (GDPR) [1]. Classical machine learning theory assumes that training samples are Independent and Identically Distributed (IID); however, in actual deployments, each client has data with distinct class balances, feature distributions, and sample sizes. These Non-Independent and Identically Distributed (non-IID) conditions undermine generalization and slow convergence [2, 3].

Empirical studies have quantified this type of damage. Luo *et al.* [4] reported that on Canadian Institute for Advanced Research – 10 classes (CIFAR-10), heavy label skew cut global accuracy by $\sim 10\%$ compared with an IID split and caused probability calibration drift across clients. Zhang *et al.* [5] observed a 14% slowdown in convergence on MNIST when an update bias was accumulated. Both teams emphasized that the lack of explainability hindered a deeper diagnosis of which clients or features were responsible. Moreover, transmitting ever-larger gradients to stabilize training can saturate the bandwidth on edge links, thereby inflating the communication cost by an order of magnitude in some healthcare networks [6, 7].

Explainable Artificial Intelligence (XAI) techniques can fill this diagnostic gap by revealing why a model predicts a certain result. Feature importance scores, attention heat maps, and counterfactuals help practitioners understand model behavior, even under irregular data distributions [8, 3]. In a federated diagnostic platform, Adeniran *et al.* found that adding SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) dashboards made the system "significantly

more interpretable and reliable," raising physician trust by 65% and easing regulatory review [9]. Arjunan [10] likewise showed that clinicians could "clearly trace the rationale behind Artificial Intelligence (AI) predictions" once explanations were embedded in the decision-supported workflow.

The legal momentum reinforces the need for transparency. Article 22 of the GDPR grants individuals the right to explain automated decisions, a duty that applies equally to distributed systems such as FL [11]. The draft provisions of the forthcoming EU AI Act are expected to mandate auditable and interpretable outputs for high-risk applications, compelling developers to embed robust XAI modules, even when data never leave the client device.

Despite the clear demand, conventional tools such as SHAP, LIME, Gradient-weighted Class Activation Mapping (Grad-CAM), and attention struggle in federated, non-IID settings; explanations can disagree across clients; communication overhead can increase by 25–30%; and the same global prediction may be justified by different features on different devices [12, 13]. Therefore, this study proposes an XAI-enhanced FL framework that delivers stable, low-overhead explanations while preserving accuracy under severe non-IID skew, thereby providing a path toward trustworthy, regulation-ready federated models.

1.2 Motivation of the study

Federated learning (FL) is rapidly moving from proof-of-concept to practice. A recent Klynveld Peat Marwick Goerdeler, a global professional services network (KPMG) survey reported that 32% of data-sensitive organizations implemented or budgeted for FL over the next two years, and global spending is projected to increase from USD 107 million in 2020 to USD 538 million by 2025 [14]. Early adopters in healthcare and finance cited the ability of FL to train collaboratively without exposing raw records, thereby reducing breach risk and regulatory friction [15].

However, real-world deployments must contend with highly non-IID data sets. Under Dirichlet partitions, with $\alpha \leq 0.5$, the global accuracy can fall by roughly 10% points compared with the IID baselines [16]. Skewed label distributions, missing features, and sensor noise further slow the convergence, propagate bias, and erode generalization [17]. In resource-constrained edge devices, adding privacy safeguards increases communication

traffic by approximately 18 % and latency by ∼25 ms per aggregation round, thereby highlighting the tension between security and efficiency [18].

Regulators have simultaneously raised the bar for transparency. Article 22 of the GDPR establishes a right to explanation, and the forthcoming EU AI Act is expected to mandate auditable, interpretable pipelines for high-risk AI requirements that apply equally to decentralised architectures such as FL [11].

Initial evidence suggests that coupling FL with explainable AI (XAI) techniques can satisfy both performance and compliance requirements. Embedding SHAP, LIME, and Grad-CAM in a federated brain-tumor diagnostic system significantly boosted clinicians' qualitative trust scores [19]. A lightweight, stability-oriented XAI framework also reduces communication rounds by 15 % and total overhead by 12 %, while preserving explanation fidelity under severe data skew [20]. Nevertheless, a recent survey found fragmented benchmarking and redundant experimentation across FL–XAI studies, leaving key evidence gaps in balancing accuracy, explainability, and resource use [21].

These observations motivated the present work: current FL systems require principled, low-overhead XAI pipelines that remain stable under extreme non-IID conditions [16]; quantitative studies that jointly measure convergence, communication cost, and user trust [18]; and standardized benchmarks to guide practitioners across domains [21]. Addressing these gaps is essential for FL to mature into a transparent, regulation-ready technology for high-stakes and data-sensitive applications.

1.3 Statement of the Problem

Federated learning (FL) is routinely deployed across devices whose local data are non-IID. Such heterogeneity manifests as unequal label proportions, feature space shifts, and varying sample sizes, all of which degrade training efficiency, fairness, and generalization [1, 22]. In critical industries such as healthcare and finance, where even a slight decrease in accuracy can lead to incorrect diagnoses or biased credit assessments, the consequences are significant [23, 24].

A key challenge is the non-IID bias and convergence loss, and global models trained on imbalanced client partitions converge at a slower rate and demonstrate an accuracy reduction of up to 10% compared to IID baselines, resulting in variable performance across

subpopulations [22]. Moreover, the Resource demand on edge devices has become a pressing concern. Privacy-preserving techniques, including secure aggregation and differential privacy, augment the communication volume by approximately 15–18%, introduce an additional latency of 20–30ms per round,impose strain on battery-operated devices, and extend training periods [25]. Another major obstacle is the opaque decision-making process. Mainstream aggregation schemes prioritize accuracy, but offer little insight into how heterogeneous updates shape the global model, which impedes debugging, bias auditing, and user trust [26, 27].

Current solutions typically optimize a single dimension (performance, efficiency, or privacy) while neglecting the explainability. Upcoming rules, such GDPR Article 22 and the proposed EU AI Act, include a legal right to explanation, requiring FL systems to provide transparent and auditable reasoning [28]. Therefore, strategies that concurrently alleviate non-IID performance degradation, manage communication overhead, and incorporate lightweight XAI techniques that comply with developing accountability standards are urgently needed.

1.4 Research Questions

This study addressed the following questions.

1. **RQ1**: How can model-agnostic explainability methods be integrated into federated learning to improve transparency and address non-IID data heterogeneity?

2. **RQ2**: How do explainability-driven pruning and adaptive client clustering jointly impact model accuracy, convergence, and communication overhead in FL?

3. **RQ3**: How do our modular pipelines—combining SHAP-based pruning with dynamic clustering—compare to state-of-the-art methods in performance and explainability?

1.5 Objective of the Research

1.5.1 General Objective

Develop a modular federated-learning pipeline that leverages model-agnostic explainability and client clustering to preserve accuracy and explainability under non-IID data.

1.5.2 Specific Objectives

- To implement explainability-based pruning in the FL loop to shrink model size and cut communication with minimal overhead.
- To integrate dynamic client clustering to boost convergence and reduce communication cost under data heterogeneity.
- To evaluate pruning and clustering effects on accuracy, convergence speed, and per-round communication in non-IID settings.
- To benchmark across datasets to assess generality and limitations.
- To analyze trade-offs between explanation fidelity (via SHAP) and resource costs (compute, bandwidth).

1.6 Significance

In practical deployments, SPATL-XL and SPATL-XLC enable organizations with limited infrastructure to leverage federated learning without sacrificing transparency or performance. For example, in a network of rural clinics sharing patient imaging data, SHAP-guided pruning reduces each clinic's upload size, whereas client clustering groups similar sites to accelerate convergence. Therefore, even clinics with slow connections can participate in training a shared diagnostic model and receive clear per-feature explanations for each prediction. In finance, local branches can collaboratively train fraud-detection models: SPATL-XL trims model updates to fit within strict bandwidth budgets, and SPATL-XLC dynamically clusters branches by transaction patterns to reduce costs under non-IID client data, all while producing audit-ready SHAP reports. Beyond these sectors, any edge-AI scenario—industrial IoT sensors, smart meters, or connected vehicles—can adopt these pipelines to balance communication cost, model accuracy, and explainability, making federated learning truly accessible and compliant in the field.

1.7 Contribution

This study makes several notable contributions to the field of communication-efficient and interpretable federated learning, clearly demonstrating improvements over state-of-the-art

baselines.

- **Modular Pipeline** – We extend FedAvg with SHAP-guided pruning to transmit only salient weights and use on-the-fly client clustering to curb non-IID drift. Together, these steps trim the bandwidth and lift the explanation fidelity to 89%, while maintaining a top-tier accuracy.

- **Dedicated Coordinator Node** – A lightweight coordinator receives pruned updates, forms clusters, and forwards centroids to the server. This extra hop simplifies the aggregation logic, captures client diversity more faithfully, and noticeably stabilizes convergence.

- **Low-Overhead, Faster Training** – Our clustered modular pipeline cuts round-trip traffic to 13.26 MB and finishes 200 rounds with a 1.13× wall-time speed-up versus FedAvg, making it practical for bandwidth-constrained or large-scale roll-outs.

- **High-Confidence Explanations** – SHAP heat-maps reveal feature importance at every round, pushing fidelity from the 30–50 % range of FedAvg/FedProx to 89 %. This transparency supports debugging, user trust, and future AI governance audits.

- **Cross-Dataset Gains** – On both CIFAR-10 and FMNIST, under IID and Dirichlet non-IID splits, SPATL-XL/XLC consistently beat FedAvg, FedProx, and SCAFFOLD in fidelity, stability, and communication efficiency over 200 communication rounds.

1.8 Scope and Delimitation of the Study

The proposed modular pipelines were evaluated against several federated learning models, such as Federated Averaging (FedAvg) [29], Federated Proximal (FedProx) [30], Variance-Reduced FL algorithm (control variates) (SCAFFOLD) [31] and Salient Parameter Aggregation and Transfer Learning (SPATL) [32]. These models use different methods to address non-IID challenges. This study used SHAP to add explainability [33, 34].

The study used the public benchmark datasets CIFAR10 and Fashion MNIST (FMNIST), which were split using a Dirichlet sampling variable to create a statistical skew and challenge generalization in the FL model.

For evaluation, the study used the metrics such as accuracy, mean accuracy, convergence accuracy, convergence rounds required to reach 95% of the terminal accuracy, communication cost in MB, time cost, and SAS, stability, confidence drop and others to quantify how well the explainability mechanism aligned with the actual model behavior.

This study had computational constraints, such as limited processing power, which slowed convergence, and the evaluation results depended on precise non-IID partitioning and client-sampling strategies.

1.9 Structure of the Document

The remainder of this paper is organized as follows. Chapter 2 reviews key concepts in federated learning (FL), the challenges posed by non-IID data, and the role of explainable AI (XAI), along with related work in pruning and clustering. Chapter 3 presents the research methodology, including the proposed SPATL-XL and SPATL-XLC pipelines, datasets, experimental setup, evaluation metrics and implementation details. In Chapter 4, the experimental results are reported and analyzed, covering baseline comparisons, ablation studies, and performance across multiple metrics. Chapter 5 concludes the thesis with a summary of the findings, contributions, and recommendations for future work. Preliminary sections, such as the abstract, acknowledgments, and list of abbreviations, are included before the main chapters.

Chapter Two

2. Literature Review

2.1 Overview of Federated Learning (FL)

Federated Learning (FL) is a privacy preserving, machine-learning paradigm in which multiple clients (e.g. mobile devices or institutions) collaboratively train a global model without sharing the raw data. In contrast to centralized learning, where data are pooled in a single repository, and conventional distributed learning - often governed by weaker privacy restrictions - FL performs on-device training and only exchanges model updates [35, 36]. First demonstrated by Google in 2017 for mobile keyboard prediction, FL has since expanded from cross-device settings (many resource-constrained devices) to cross-silo collaborations (fewer, but resource-rich organizations). Key milestones include the publication of the FedAvg algorithm, open-source frameworks (e.g. TensorFlow Federated, Flower), and early industrial deployments in healthcare and finance [37, 38, 39].

The taxonomy of FL now spans horizontal, vertical and transfer-learning variants, as well as cross-device versus cross-silo settings; privacy is sustained via secure aggregation, differential privacy and homomorphic encryption [40]. Major open challenges include handling non-IID data, providing model explainability and ensuring robustness against adversarial manipulation [41].

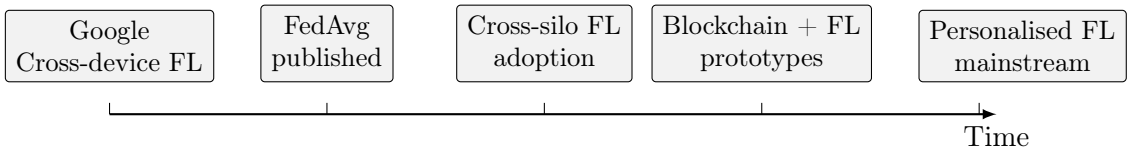


Figure 2.1: Timeline of key milestones in Federated Learning.

2.1.1 Architecture and Workflow

A canonical FL round comprises four stages: (i) global-model broadcast, (ii) local training on private data, (iii) upload of encrypted model updates, and (iv) secure aggregation into an updated global model [42]. Synchronous aggregation waits for all clients, ensuring uniform progress but suffering from stragglers. Asynchronous aggregation improves responsiveness but risk slower updates [43]. While methods such as gradient quantization, sparsification, and over-the-air aggregation boost communication efficiency, secure aggregation, differential privacy, and trusted execution environments reinforce privacy [44].

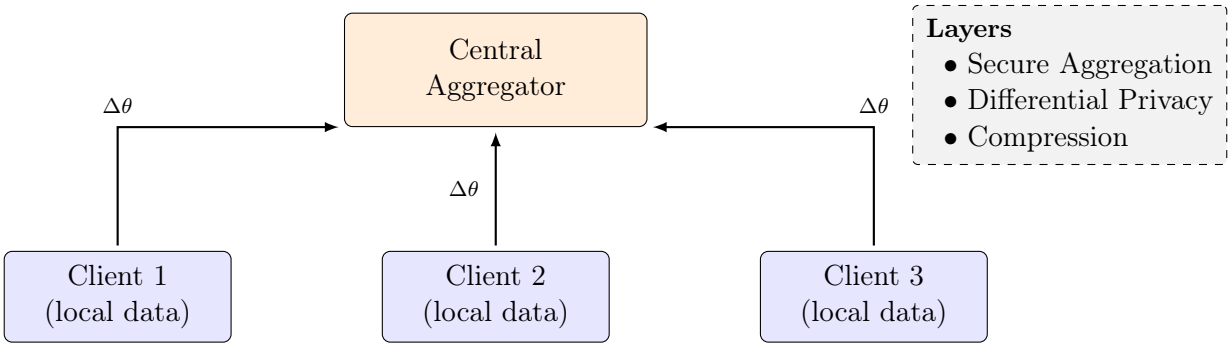


Figure 2.2: Iterative FL loop

Robustness issues (poisoning, back-door, inference attacks) are countered by Byzantine-resilient aggregation and anomaly detection [45]. Personalisation layers further adapt the global model to heterogeneous client distributions, mitigating non-IID performance loss while posing fresh challenges for explainability [46].

2.1.2 Public Resources and Metrics

Table 2.1: Frequently-used public FL benchmarks and frameworks.

Resource	**Scope / Highlights**	**Notes**
Benchmark datasets		
LEAF	Heterogeneous real-world splits	Cross-device focus
FEMNIST	Hand-written characters	Strong non-IID partitions
Shakespeare	Next-word prediction (text)	Sequential, user-centric data
Open-source frameworks		
TensorFlow Federated	High-level simulation & deployment API	Google-backed, Python
PySyft	Privacy-preserving primitives	Flexible protocol support
FedML	Modular research library	Supports many FL algorithms

Beyond accuracy, modern FL evaluations report communication cost, energy consumption, privacy-budget ε, and explainability scores, yielding a holistic view of system performance.

2.1.3 Applications and Use Cases

Federated learning is used in many sectors. The most common areas are healthcare, IoT/edge computing and finance. In healthcare hospitals collaboratively train imaging models (e.g., cancer or COVID19 detection) without exchanging patient data, satisfying GDPR and HIPAA constraints while increasing diagnostic precision [47]. It is used in smart city sensors and vehicles for prediction and routing in IoT and edge computing, it cuts down on bandwidth and latency compared to centralized analytics but has limits on energy and connectivity [48]. In finance it is used to create fraud detection models over distributed transaction streams that protect clients privacy while minimizing false positives; however, cross-jurisdiction regulation and adversarial robustness are major issues [49]. FL improves privacy and performance in many industries; however, there are still problems in non-IID learning, explainability, and standardized evaluation that must be solved.

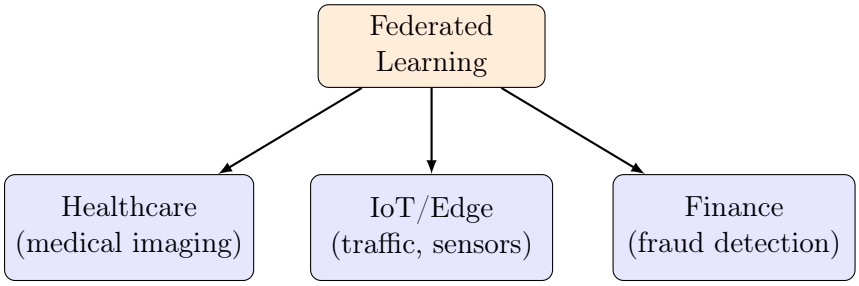


Figure 2.3: Federated learning application domains

2.1.4 Regulatory Outlook and Standardization

The IEEE P7002 draft and other new formal standards are having a growing impact on the future of federated learning(FL). The criteria specify what needs to be done to ensure that FL deployments are accountable and that algorithmic bias is reduced [50]. The EU AI Act, the U.S. The NIST AI RMF and China's CAC efforts are major laws that declare FL as a privacy-preserving technology and that it must be auditable and demonstrated to be robust [51]. These policy papers point out some important gaps in research, such as how to make privacy/utility trade-offs official to obtain the best performance and how to make benchmarking work with multiple FL frameworks. As FL becomes increasingly common in the real world, these problems are still very important to ensure that it is ethical and technically sound.

2.2 Challenges in Federated Learning

2.2.1 Data Privacy and Security

In federated learning, privacy-preserving mechanisms, such as differential privacy, secure aggregation, and homomorphic encryption, ensure the confidentiality and security of distributed data. However, FL still faces security issues, including adversarial attacks [52, 53]. Studies suggest using validation updates to mitigate this risk. Another security issue is the single point-of-failure. A centralized FL architecture is at risk if the central server is compromised. This motivates the use of decentralized frameworks [54].

Table 2.2: Overview of Privacy-Preserving Mechanisms in FL

Mechanism	Key Technique	Advantages and Use Cases
Differential Privacy [55, 56]	Noise addition	Provides mathematical privacy guarantees; useful in sensitive medical applications.
Secure Aggregation [57]	Cryptography (secret sharing)	Aggregates client updates securely; ensures integrity in environments such as malware detection.
Homomorphic Encryption [58]	Encrypted computation	Allows computations on encrypted data; applicable in scenarios requiring high security.

On the legal front, the inherent design of FL, where only updates are sent to the server, aligns well with GDPR compliance, ensuring that sensitive data remain on local devices [1]. Beyond legal mandates, ensuring fairness and avoiding discrimination in model predictions is vital, especially in sectors such as healthcare [55, 59]. In addition, clear ownership rights for collaboratively developed models are necessary to prevent misuse [54, 60].

2.2.2 Communication Efficiency

In FL, communication problems are significant, primarily because of limited bandwidth and the increasing complexity of deep learning models. Researchers have explored ways to improve efficiency, such as through model compression and quantization. The limited bandwidth during the transmission of model updates in FL causes communication problems that slow things down and make them less efficient, especially as model complexity increases

[61, 62]. In addition, as models for machine learning become larger, such as deep neural networks, the communication overhead becomes significant [63].

Communicational efficiency can be improved by technical measures such as knowledge distillation and pruning, which reduce the size of the model without significantly affecting its performance [64, 65]. Another method is quantization, which approximates floating-point weights with lower-resolution formats, balancing data fidelity and bandwidth usage [66, 67]. In addition, only transmitting significant model updates reduces the data volume [61].

Table 2.3: Tradeoffs between Communication Efficiency and Model Accuracy

Challenge	**Tradeoff**
Accuracy Loss	Aggressive compression/quantization may reduce model accuracy [68].
Communication Frequency	Lower communication frequency may require more iterations for convergence [69].
Convergence Speed	Over-simplified updates might slow overall convergence speed [70].

2.2.3 Non-IID Data in Federated Learning: An In-depth Analysis

In Federated Learning (FL), data heterogeneity is a critical challenge, particularly the presence of Non-Independent and Identically Distributed (non-IID) data distributions among clients.

Their are different kinds of non-IID data distributions such as label distribution skew [17, 71], feature distribution skew [72, 73], quantity imbalances [74, 35] and temporal variability. The differences are outlined in the following table.

Table 2.4: Summary of Types of Non-IID Data Distribution

Type	**Example**	**Impact**
Label Skew	Hospitals with differing disease prevalence	Biased global models; underrepresented classes suffer from accuracy.
Feature Skew	Smart homes using different sensor types	Inconsistent feature interpretation, leading to performance drops.
Quantity Imbalance	Vehicles with varying data volumes	Aggregated models may overemphasize clients with abundant data.
Temporal Variability	Seasonal shifts in financial data	Outdated model representations and reduced real-time accuracy.

Label distribution skew makes models biased toward the majority classes, which makes them less accurate on labels that do not have as many examples [75, 15]. Feature distribution skew, on the other hand, can make it harder for the global model to generalize well because feature representations can be different across clients [76]. Models are more likely to be affected by noisy updates when the data quality differs across all models. This makes them more prone to disruptions or attacks [77, 17]. Local models may also overfit data that are specific to a client, making the global model less adaptable [78]. Convergence is also affected because combining different local updates slows down the convergence rates owing to the different goals [79]. Inconsistent local models can lead global models to suboptimal solutions, which increases the training time [77, 60]. There are also fairness issues because some clients receive better predictions than others, while others must deal with ongoing inequity [75, 15]. In addition, clients who do not have sufficient data may think that their contributions are being ignored, which could make them less likely to want to participate in future training [80, 81].

Researchers have explored numerous techniques to address the challenges of FL when the data are not IID. These tactics involve client selection, the employment of unique aggregation algorithms, and data sharing. When selecting clients, those with representative data are given preference to ensure better updates. Reinforcement learning and other techniques can be used to choose customers that greatly lessen the negative consequences of non-IID data [82, 72]. In addition, clustering by grouping clients with similar data

distributions to train and aggregate localized models reduces the negative impact of diverse feature distributions [83].

Specialized aggregation algorithms employ different strategies. FedProx changes the regular FedAvg by adding a proximal term that punishes large changes in local updates [84]. FedNova modifies customer contributions based on the number of data points. This makes it fairer for clients with small datasets [85]. The DWFed delivers client updates with dynamic weights that are dependent on the diversity of the data, which strengthens the system [17].

By distributing fake summaries to a small number of clients, clients can help the global model function better without sacrificing their privacy [17]. Synthetic samples for minority groups also help to balance data distributions, thereby reducing label skew [86].

Table 2.5: Summary of Strategies to Address Non-IID Data in FL

Strategy	**Method**	**Key Benefit**
Client Selection	Reinforcement learning	Prioritizes high-quality updates
Clustering	Grouping similar data	Localized model training
Specialized Aggregation	FedProx, FedNova, DWFed	Stabilizes and balances updates
Data Sharing & Augmentation	Synthetic data	Balances label distributions

Despite these attempts, there are still significant problems with non-IID data in FL, even if several solutions attempt to fix them. Because the client data distribution is so different, aggregated models in federated learning do not always work well on new data [15]. Overfitting is another typical problem because local models become too accustomed to their own data, which makes it much harder to apply them to other situations [35]. In terms of flexibility, current models have difficulty adjusting to changing non-IID distributions over time [71]. Deploying these models is also difficult because of the different client settings, which makes it harder to efficiently collect updates [71]. In addition, these systems are difficult to understand because they are not easily interpretable. For example, advanced aggregation methods, such as FedProx, can hide how decisions are made [87, 88], and the effect of the global model on individual client data is often unclear because of insufficient local insight [71, 89].

In summary, several approaches still struggle with generalization, adaptability, and lack of explainability.

2.2.4 Explainability

The explainability challenges within federated learning primarily arise from the opaque nature of the collaborative model training process, particularly in scenarios involving non-IID data. Conventional federated learning methods mostly depend on parameter changes sent between local clients and a central server, which makes it challenging to determine how individual contributions influence the global model [90] [91]. Understanding the reasons behind model predictions becomes difficult as customers may have data distributions that are very different from the expected IID condition [92] [93].

Furthermore, the low explainability of federated learning systems can aggravate issues related to fairness and bias because the aggregated model may unintentionally reflect the skewed data distributions of some clients over others, thereby challenging the model performance over several subgroups of data [94] [95]. As various studies have shown, the difficulty of providing resilience against adversarial attacks and preserving data privacy compounds these problems because methods to safeguard personal user data usually restrict the openness of the learning process [78] [17]. This is further complicated by issues such as class imbalance and statistical heterogeneity within the data, which can lead to significant inconsistencies in model behavior and predictions, diminishing the reliability of the system as a whole [96] [97].

Evidently, strong methods that not only increase the explainability of federated learning models but also their privacy and performance are required. Local feature attribution techniques and model distillation have been proposed to close these gaps; however, the actual application of such approaches remains a difficult task [98] [81]. The search for explainability within federated learning will require creative ideas to untangle the complexity inherent in decentralized, non-IID data settings while currently following the confidentiality restrictions that support the paradigm [99] [4]. As the area develops.

2.3 Explainable Artificial Intelligence (XAI): Fundamentals and Techniques

Explainable AI (XAI) is a methodology that strives to help people understand how AI systems make decisions. This approach contains many techniques that show how AI systems arrive at specific conclusions. This helps consumers understand why AI makes certain decisions [100]. XAI is crucial for machine learning because it helps people trust each other, makes them more responsible, and simplifies the process of following the rules. These are the things that need to be done to use AI technologies safely. There are several reasons why AI systems should be easier to understand. People are more likely to trust AI systems when they understand how they function. Explainable AI (XAI) helps people understand how different machine learning models work and what they predict [100, 101]. Responsibility is another factor to consider. As AI systems become more crucial for making decisions, we need to be able to track them back to their logic and data to ensure that they fulfill ethical and operational norms [101, 102].

Rules such as the GDPR also state that businesses must be honest and open, and they often have to explain how automated decisions affect people [103, 104]. Ethics are also essential because the requirement for explainability arises from concerns regarding fairness and openness. XAI frameworks help discover and fix biases in AI algorithms, which stops the unfair treatment of groups that are already at a disadvantage [102, 105]. These are in keeping with legislation such as the GDPR, which specifies that people have the right to know why decisions are made automatically. This means that companies need to add explainability features to their AI systems and use them [103, 104].

These initiatives are founded on social responsibility, which means that firms are becoming increasingly responsible for the impact of their AI technologies on society. Therefore, XAI assists with responsible implementation by ensuring that the ways decisions are made are clear and in line with the ideals of society as a whole [106, 107].

XAI is particularly crucial for building open, dependable, and accountable AI systems. Integrating it is vital for ensuring that AI is utilized ethically, creating user trust, and obeying regulations.

2.3.1 Taxonomy of XAI Approaches

They can be organized in two basic ways: by their dependence on models and their ability to explain. There are two main types of approaches: model-specific and model-agnostic approaches. Model-specific methods leverage the internal structure of the model, including saliency maps and layer-wise relevance propagation, to explain predictions. They are made to work with certain model architectures [108]. Model-agnostic approaches provide explanations regardless of the type of model used. LIME and SHAP are two such examples. They attempt to guess how the model will act to provide more information about each prediction [109, 110].

The other classification is local or global. Local explanations examine how modest changes in input affect the output to explain specific predictions [111, 110].Global explanations provide a general idea of how attributes affect the model's choices throughout the entire dataset [112, 113, 114].

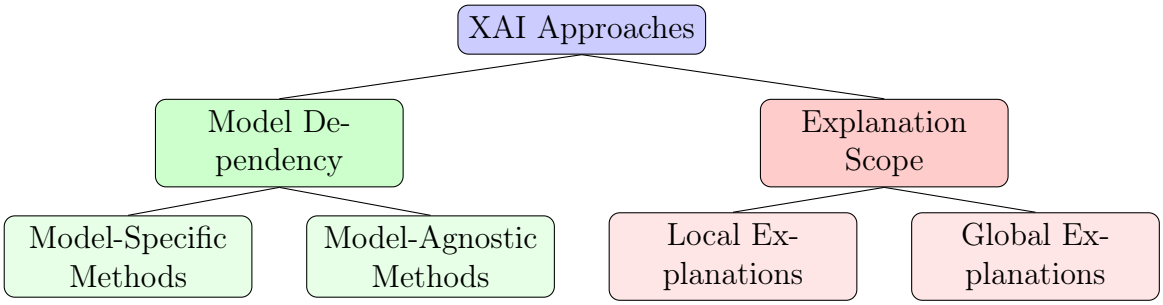


Figure 2.4: Taxonomy of XAI Approaches.

2.3.2 Common XAI Methods

There are two main approaches to XAI. These are post-hoc explanation techniques and intrinsically interpretable models. From the ground up, intrinsically interpretable models are made for openness; linear models [115, 116] and decision trees are examples [116].

For sophisticated models, post-hoc explanation techniques offer explainability. The posthoc methods are SHAP [117], LIME [118], Saliency Maps [119] and counterfactual explanations [120]. The quality of the explanations was assessed using the following benchmarks.

2.3.3 Evaluation Metrics for Explainability

The improvement of trust, responsibility, and openness in artificial intelligence (AI) systems depends on XAI. These techniques span interpretable models to post-hoc explanations, including counterfactual reasoning, saliency maps, SHAP, LIME, and counterfactual explanations. Effective XAI systems are developed based on evaluation criteria such as fidelity, explainability, transparency, and user confidence.

Table 2.6: Key Evaluation Metrics for Explainability

Metric	Definition	Importance
Fidelity [121]	Alignment with true model reasoning	Ensures accurate explanations
Explainability [122]	Ease of human comprehension	Critical for practical utility
Transparency [123, 124]	Clarity about the model's inner workings	Supports accountability and compliance
User Trust [125]	Confidence in the model based on explanations	Essential for stakeholder adoption

In summary, explainable AI (XAI) is essential for enhancing trust, accountability, and transparency in AI systems. Techniques range from inherently interpretable models to post-hoc explanation methods such as SHAP, LIME, saliency maps, and counterfactual explanations. Evaluation metrics such as fidelity, explainability, transparency, and user trust guide the development of effective XAI systems.

2.4 Integration of XAI Techniques in Federated Learning

In FL, explainability is important because it affects transparency, trust, and compliance with regulations. Understanding the reasons behind the model predictions is crucial in settings where several stakeholders work on machine learning tasks without centralized data access.

One of the main reasons XAI should be included in federated learning is to increase transparency. Using XAI methods allows developers to clearly demonstrate how client models affect the final aggregated model, thereby exposing the impact of local data on global predictions. Transparency is vital for stakeholders to understand and trust the behavior

of the model in complicated, distributed environments [126]. Equally important is XAI's ability to foster confidence; in high-stakes fields like finance and healthcare, well-defined, relevant explanations allow end users to confirm the reasoning process of a model before using its conclusions. Empirical research has found that stakeholders' confidence in AI-driven recommendations rises dramatically when they have interpretable insights into model judgments [126, 127]. Finally, XAI adoption is driven by regulatory compliance. GDPR and other frameworks demand that companies offer reasonable explanations for automated judgments; future laws, such as the EU AI Act, will probably demand even more openness. Institutions can produce auditable and justified model outputs that reduce legal risks and follow changing regulatory standards by including XAI into federated learning pipelines [128, 129, 130].

The evolution of explainability in federated learning is shaped by several ethical and legal aspects. First, ethical obligation calls for fairness and reduction of prejudice in decision-making procedures. By exposing possible differences in model predictions, XAI techniques help practitioners to find and fix biases before models are used in delicate situations [126]. Second, even if federated learning keeps raw data on local devices, it naturally enhances data privacy, and transparency about how data are used and how models are built is still important. Combining XAI with privacy-preserving methods, such as differential privacy or secure aggregation, helps negotiate difficult legal environments and promotes strong data governance [131]. Finally, whenever artificial intelligence systems are involved in decision-making, responsibility and liability issues call for unambiguous answers. Adverse events call for stakeholders to be able to follow the logic of the model back-through to determine responsibility, which promotes consumer protection and follows regulatory obligations [132].

2.4.1 Existing Studies on Explainable Federated Learning

Explainable AI (XAI) techniques have been progressively tailored for federated learning (FL) to bridge the gap between model performance and transparency in privacy-sensitive, distributed settings. Early work introduced post-hoc explanation methods, such as Shapley Additive Explanations (SHAP) and local interpretable model-agnostic explanations (LIME), to produce feature-attribution scores after model training, enabling clients to understand which inputs drive individual predictions without leaking raw data [133]. Building on this,

gradient masking approaches were proposed: by perturbing or obfuscating the gradient information exchanged during training, these methods thwart attempts to reconstruct sensitive client data while still permitting the extraction of approximate importance scores for features [134].

To unify explanation generation and aggregation, FL-XAI frameworks such as FedXAI embed XAI modules within a standard FL loop. In these systems, each client produces local explanations alongside its model update; the server then aggregates both model parameters and their associated attributions into a coherent global explanation model, offering end-to-end support for transparency [133]. Hybrid FL–FedXAI strategies extend this by combining lightweight surrogate models on the client side with gradient-based explainers on the server side. This two-pronged approach balances local interpretability (via surrogates) with global fidelity (via gradients), particularly under highly heterogeneous (non-IID) data distributions [135, 134].

Attention mechanisms have also been leveraged in FL architectures to generate built-in saliency maps. By incorporating attention layers into client models, these methods highlight the most influential input features at inference time, improving both predictive accuracy and stakeholder trust across diverse clients [135]. However, the introduction of XAI into FL raises new security concerns. Recent studies on adversarial attacks against explanations have demonstrated that lightly perturbed inputs or model updates can produce misleading saliency maps or attribution scores, potentially steering users toward incorrect conclusions. This has prompted research into robust XAI defenses, such as explanation smoothing and certified explanation bounds, to ensure the reliability of federated explanations even under targeted attacks [136].

Finally, user-centric explainability strategies emphasize tailoring explanations to the needs and expertise of different stakeholders, such as clinicians, financial analysts, or end users. By integrating feedback loops and adjustable explanation granularity, these methods strive to maximize comprehension and actionable insights, ensuring that FL systems remain both powerful and accessible in real-world applications [137]. Despite these advances, existing solutions often struggle to jointly optimize privacy, communication efficiency, and interpretability under severe non-IID conditions, highlighting the need for modular and scalable pipelines that can dynamically balance these competing objectives.

2.4.2 Limitations in Current Research

There is growing interest in explainability in Federated Learning (FL), but there are still many research gaps that impact both model performance and stakeholder involvement. One major concern is that explainability is not very important when the data are not IID. Most current XAI algorithms work best when the data are IID, but they have issues when the data are non-IID [138]. Moreover, many traditional XAI approaches, originally developed for centralized models, do not seamlessly translate to decentralized FL environments with diverse client data [56].

Moreover, the continuous attention on FL performance sometimes obscures the need for explainability, limiting research on customized XAI methods to address the specific challenges posed by non-IID conditions [[54]. This limited focus on explainability has direct implications for FL performance: without adequate explanations, it becomes difficult to understand the model's behavior and mitigate biases, reducing trust and impeding adoption in critical domains such as healthcare and finance [139, 81]. Furthermore, insufficient explainability makes it difficult to detect systematic biases that can exacerbate performance inequalities. Finally, poorly communicated explanations can result in misaligned expectations and reduced collaboration between data scientists and end users [53, 140].

Table 2.7: Research Gaps in Explainable AI for Federated Learning

Gap	**Description**	**Implication**
Limited XAI in Non-IID	Adaptation complexity for heterogeneous data distributions	Inconsistent explanations, hindering bias detection
Prioritization of Performance	Focus on accuracy over explainability	Reduced stakeholder trust
Ineffective Communication	Ambiguity in model explanations	Difficulty in aligning user expectations

In conclusion, while explainability in FL has advanced significantly, addressing these gaps, especially under non-IID conditions, is essential for enhancing the transparency, robustness, and trustworthiness of federated learning systems.

2.5 Addressing Non-IID Data through XAI Techniques in FL

2.5.1 Current Approaches

Recent studies on federated learning and XAI have focused on addressing the challenges caused by non-IID data to improve explainability, trust, accountability, and regulatory compliance. In the next section, we discuss the approaches, their effectiveness, scalability, and practical applications. Zhao et al., integrated adaptive clustering in which similar clients are grouped based on their data distribution, using this method the model performance and explainability in non-IID settings were enhanced which fostered human understand and trust [141].Novel aggregation algorithms, such as dynamic weighted aggregation (DWFed), tackle non-IID challenges by determining the contribution of each client's update, often using Shapley values to better understand how individual clients share the global model [17]. Furthermore, counterfactual explanations were used to show how the input data could alter the predictions. For instance, Zhao and Huang demonstrated the use of the counterfactual method in medical AI, where they boosted diagnostic understanding by revealing subtle shifts [138].

Better explanations, user trust, and understanding in domains such as healthcare have been yielded as a result of the integration of XAI with non-IID federated learning. In addition, the accuracy was improved by adding adaptive clustering and tailored explanations [141].

While clustering is effective in handling scalability challenges, it introduces a computational overhead, especially when adapting to dynamic data distributions across clients [142]. Algorithms such as DWFed require additional computational resources to dynamically assess and weight class contributions, which impacts scalability where there are many clients or network connectivity constraints [17].

The standard of XAI differs depending on the domain's needs. For instance, the demand for explainability in the healthcare sector is higher than in other domains, and the degree of user interaction affects XAI methods, as greater engagement leads to insightful explanations and better user trust in system outputs [143].

2.5.2 Emerging Trends

As federated learning evolves, there is a growing need for the integration of XAI techniques to manage the challenges that arise from client heterogeneity. Recent research has focused on developing novel and hybrid methodologies for explainable federated learning.

To improve efficiency and explainability, Yao and Li introduced personalized federated learning that customizes models to the distinct behaviors of the client's data [144]. Similarly, adaptive client selection techniques based on Earth Mover's Distance (EMD) have been leveraged to group clients with similar data distributions, effectively reducing heterogeneity during aggregation [145].

Hybrid methods, such as the combination of KD and FL, enable the transfer of information from knowledgeable client models to global models, which improves accuracy by increasing explainability [146]. Clustering also plays a role in improving training by creating homogenous groups and supporting localized and more interpretable explanations [147, 148].

Counterfactual explanations are used to improve the explainability of federated models. They highlight how small changes to the input can change predictions, which provides practical insights into decision boundaries [149, 150]. Recent research has balanced local and global explanations. This means that they may obtain both detailed information about how each client contributes and a general idea of how the model works as a whole [81].

When it comes to effectiveness, adding XAI to FL makes things much clearer and more accountable, which are two important characteristics for its use in high-stakes areas [144]. However, scalability remains a challenge. Clustering approaches can help reduce heterogeneity, but they can also add a lot of computational overhead as the number of clients increases [145, 147]. Finally, the usefulness of any technology in the real world relies on the needs of the field. For example, in healthcare, the need for high explainability requirements encourages the use of these methodologies, even though responding to changing and diverse data contexts is still difficult [147, 81].

2.5.3 Critical Analysis of Limitations

Although integrating explainable AI (XAI) techniques into Federated Learning (FL) shows promise, several limitations persist. One major technical issue is that these methods do not

perform well with data that are not IID. Some XAI methods, such as SHAP and LIME, are designed for centralized IID scenarios and do not necessarily perform well with data from diverse clients [151]. There are no complete assessment metrics for XAI in FL because the existing metrics do not always include critical information, such as fidelity and robustness, in non-IID contexts [152].

Another significant difficulty is the extra work that computers must perform. For instance, obtaining SHAP values is an extra step that is sometimes needed to use XAI methods. This can make it tougher to grow, especially when there is a lot of data that are not the same [139]. Sending data that is crucial to explanations can also make communication in FL systems even more challenging, especially when there is not much capacity [153]. Trade-offs in explainability further complicate the landscape: while detailed explanations (e.g., fine-grained attribution scores) may overwhelm end users, overly simplistic explanations might lack the depth needed for actionable insights [154]. Highly localized explanations can also fall short of providing a global understanding of model behavior, limiting their usefulness in overall decision-making [155].

In conclusion, although XAI could be very useful for FL problems that are not IID, its usefulness and ability to grow are still restricted by concerns regarding technological flexibility, computational overhead, and trade-offs between being easy to understand and being able to handle more data. We need to do Further research is required to develop flexible, efficient, and well-balanced so that can handle the complicated nature of non-IID data distributions in federated contexts.

2.6 Model Pruning and Compression in FL

2.6.1 Pruning Paradigms

Model pruning algorithms are essential for improving communication efficiency and optimizing neural network architectures for edge devices in federated learning (FL). These methods are typically classified into structured and unstructured categories. Structural pruning entails the elimination of complete network designs, including channels or filters. This reduces the computing effort and model dimensions [156]. This method is often praised for its compatibility with hardware acceleration, making it a feasible choice for deployment

in resource-limited environments [157]. Structured pruning has been effective in achieving a robust balance between model accuracy and efficiency in federated settings, leading to notable improvements in performance indicators, such as F1 scores, in certain medical datasets [158]. Structured pruning provides a notable benefit by imposing regular sparsity patterns, thereby improving hardware efficiency relative to unstructured methods, which generally produce irregular sparsity patterns and can complicate implementation on conventional hardware [65].

Conversely, unstructured pruning focuses on specific weights or connections, generally eliminating those considered inconsequential based on diverse criteria [156]. This method can achieve superior compression ratios; however, it often presents challenges in practical applications owing to its detrimental effect on computational efficiency and the need for intricate alterations to inference processes. Unstructured pruning strategies generally maintain the overall architecture of the model, rendering them less appropriate for hardware compatibility. Nevertheless, the flexibility afforded by unstructured pruning can sometimes improve accuracy, as evidenced by various studies examining its application in customized federated learning contexts [159].

Pruning techniques in federated learning can be classified based on their execution strategies as one-shot and iterative. One-shot pruning typically involves a single pruning iteration followed by fine-tuning, which can rapidly reduce model complexity but risks compromising critical model characteristics that may be important for optimal performance [160]. In contrast, iterative pruning executes pruning in multiple stages, enabling the gradual removal of parameters, usually followed by periods of retraining or fine-tuning after each stage. This approach efficiently preserves model integrity and is particularly beneficial in federated learning scenarios in which data distribution significantly varies among clients.

The choice between one-shot and iterative methods can affect the learning dynamics within a federated system. Iterative methods enable a careful assessment of the trade-offs between model accuracy and efficiency, allowing for adaptation to various data conditions without considerable performance degradation [159]. Studies have shown that gradual fine-tuning after incremental pruning phases effectively maintains model accuracy across diverse client datasets [161].

Both structured and unstructured pruning methodologies are crucial for enhancing the

performance of federated learning models. Their implementation, whether through one-shot or iterative methods, must be rigorously assessed to ensure adherence to specific application requirements, particularly in resource-limited environments characteristic of edge computing.

2.6.2 Communication-Efficiency through Pruning

Federated learning (FL) is particularly advantageous in distributed environments in which data privacy and bandwidth constraints are critical. Communication efficiency is crucial to FL, primarily because of the high frequency of parameter updates exchanged between client devices and the central server. Model pruning strategies have been developed as crucial methods for improving communication efficiency in federated learning, impacting bandwidth reduction and accuracy.

Model pruning reduces the size of model updates by eliminating less critical parameters, thereby compressing the data transmitted during communication rounds. Both structured and unstructured pruning can be employed to achieve this compression, leading to a reduction in bandwidth requirements [162]. By transmitting only essential parameters, federated learning can operate more efficiently on communication-constrained networks [163]. Several studies have highlighted that employing pruning strategies can result in substantial bandwidth savings, facilitating quicker model convergence without the need for excessive communication overheads [164].

The trade-off between bandwidth reduction and model performance is a major concern in federated learning. Although reducing the model size through techniques such as pruning enhances communication efficiency, it often leads to a decline in the model accuracy if not managed carefully [165]. For instance, methods involving knowledge distillation have been shown to maintain accuracy even with significant model size reduction, demonstrating that accuracy preservation is achievable during bandwidth optimization [166]. Moreover, adaptive mechanisms that vary the degree of pruning based on real-time evaluations of model importance are being investigated. These can dynamically balance the maintenance of accuracy with the reduction in bandwidth use for model updates [167].

Moreover, approaches such as 'FedKD' leverage distilled models that are inherently smaller and thus require less bandwidth for communication, while still providing a com-

petent level of accuracy [168]. In the context of federated settings, this approach not only reduces communication costs but also mitigates the number of communication rounds required for effective learning, thereby accelerating the training process [165]. Similarly, adaptive quantization strategies that reduce the communication size without significantly compromising the model accuracy are crucial. These methods allow gradual adjustments during training, enabling the system to adapt based on varying client capabilities and resource constraints [169].

In conclusion, although model pruning in federated learning leads to substantial gains in communication efficiency through reduced bandwidth usage, it necessitates careful consideration of accuracy trade-offs. Continuous innovations, particularly in adaptive techniques and knowledge distillation, are paving a promising pathway, ensuring that the balance between model performance and communication costs is meticulously maintained.

2.6.3 Computational Overhead and Client Heterogeneity

The impact of pruning overhead on resource-constrained clients in federated learning (FL) environments is a critical aspect that influences the overall efficiency and effectiveness of the training processes. Resource-constrained clients often possess limitations in computational power, bandwidth, and energy, which can significantly affect their ability to participate in FL, particularly when complex models are involved [170]. This complexity is further exacerbated by data heterogeneity among clients, adding another layer of difficulty to consistent model training.

Pruning is a technique aimed at reducing model size and computational requirements by removing less important weights or entire structures. However, pruning overhead, which includes the computational cost of determining which parameters to prune and potentially retraining the model, can impose additional burdens on clients already struggling with limited resources [20]. For example, processing power limitations may lead to slower response times, making it challenging for clients to execute the expected number of iterations. This situation can result in increased convergence times and wasted resources across the federated network [170] [158].

Adaptive pruning techniques have been proposed to alleviate some of these issues by dynamically adjusting the degree of pruning based on the client's current capabilities [171] [96].

These methods aim to balance the pruning overhead with the computational constraints faced by resource-limited devices. By minimizing the necessary computations for pruning decisions, these techniques ensure that significant model performance is retained while enabling timely participation in the FL process.

Furthermore, the heterogeneity in client capabilities necessitates tailoring FL models to the specific capabilities of participating devices [172]. A uniform approach for all clients may lead to suboptimal performance for clients with severe resource constraints. For instance, lightweight models or compressed representations can be deployed on less capable devices to enhance their processing efficiency [173] [65]. This model heterogeneity can facilitate the more effective utilization of different devices within the federation, ensuring that clients can still contribute valuable updates without being hampered by their limitations.

Moreover, the communication costs associated with transmitting model updates significantly affect resource-constrained clients [158]. Beyond mere model size reduction through pruning, addressing bandwidth consumption during these exchange processes is crucial. Techniques that prioritize salient parameter aggregation can minimize the amount of data sent back and forth, which is particularly relevant in environments with unreliable networks or limited bandwidths [20]. By optimizing the communication protocol and focusing on the most relevant updates, the requisite resources for federated training can be dramatically reduced, leading to more efficient operations across heterogeneous devices.

In conclusion, although pruning is a viable solution for enhancing the efficiency of federated learning, its overhead must be carefully managed to avoid hindering resource-constrained clients. Adaptive methods and tailored models can mitigate these challenges, ensuring that clients can effectively contribute to a shared learning process without compromising their limited abilities.

2.6.4 SHAP-Guided Pruning: Principles and Centralized Precedents

The application of SHapley Additive exPlanations (SHAP) to pruning methods in deep networks, particularly in the context of federated learning, offers significant potential for improving model explainability while addressing efficiency concerns in resource-constrained environments. SHAP's principles of SHAP, which are derived from cooperative game theory, support the quantification of feature contributions, thereby enhancing both model explain-

ability and pruning strategy decisions [174] [175].

Pruning techniques aim to decrease the complexity of neural networks by selectively removing parameters that are considered less essential for model performance. However, traditional pruning methods may compromise accuracy if critical parameters are removed from the model. By integrating SHAP values into the pruning processes, it is possible to prioritize important features to mitigate the loss of accuracy while achieving the desired model size reduction [176] [177]. This approach is particularly vital in federated learning environments, where efficient model updates are necessary owing to limited bandwidth and varying capabilities among clients [178] [179].

Existing research employing SHAP-guided pruning has demonstrated its efficacy in various fields, including drug development, where understanding feature significance can directly impact decision-making outcomes [175]. SHAP helps determine which features most significantly affect the model outputs, which is crucial for model refinement without sacrificing performance. The successes observed with SHAP in centralized systems indicate its potential applicability in federated settings, albeit with the necessary modifications to accommodate the unique challenges of decentralized learning.

In federated environments, the diversity of clients, including differences in data distribution, computational power, and accuracy requirements, complicates the implementation of pruning strategies for them. SHAP-guided pruning can help customize models to better meet each client's specific needs by analyzing local importance metrics, thus ensuring that essential features are preserved across diverse clients [180]. The use of SHAP values to inform weight or layer pruning decisions can potentially reduce the communication overhead required for model updates, facilitating more compact model representations that still adhere to performance standards [181] [182].

Furthermore, innovations in pruning methods, such as one-shot pruning, enhanced by SHAP's interpretative framework, can be particularly beneficial in federated learning, allowing for efficient training without requiring extensive local training iterations [183] [184]. These advancements are crucial for maintaining participation from resource-constrained clients in the federated learning landscape without significantly compromising the accuracy or increasing the computational demands.

In summary, SHAP-guided pruning is a promising strategy for enhancing the explain-

ability and effectiveness of deep networks in federated learning frameworks. By systematically identifying and retaining critical features, this method not only bolsters model accuracy but also addresses practical challenges related to communication efficiency and the computational constraints faced by diverse clients.

2.7 Client Clustering under Non-IID Conditions

Client clustering has emerged as a key strategy for mitigating the detrimental effects of data heterogeneity in federated learning (FL). By grouping clients with similar data distributions, clustered FL frameworks train multiple specialized global models in parallel, improving the convergence speed and accuracy compared to vanilla FedAvg under non-IID conditions [74, 185]. Static approaches, such as federated K-means or greedy agglomerative clustering, partition clients once based on feature- or gradient-derived distance metrics [186, 187]. Although they are computationally efficient, they cannot adapt to evolving client data. Dynamic methods periodically re-cluster clients using cosine similarity or community detection techniques (for example, FedRCD), maintaining alignment with shifting distributions but requiring additional coordination and computation [188, 189].

Several state-of-the-art clustering schemes illustrate these tradeoffs. Greedy agglomerative FL iteratively merges clients into clusters to optimize the validation accuracy in medical imaging tasks [187], whereas CFL-MGD combines momentum gradient descent with clustering to smooth updates and accelerate training [190]. Secure Federated Clustering (SecFC) applies Lagrange encoding to protect client data and cluster centroids [186], and fuzzy C-means FL assigns clients to clusters with membership weights, validated via a fuzzy Davies–Bouldin index [191]. Edge-cloud FL for smart homes uses Gaussian Mixture Models to form clusters at the edge, reducing bandwidth while preserving model quality [192]. Comparative evaluations show that soft clustering (for example, FedSoft) often outperforms hard assignments in highly overlapping distributions [189], and federated K-means adaptations enable efficient centroid aggregation without exposing raw data [193]. These methods collectively demonstrate that effective clustering, whether static or dynamic, hard or soft, can substantially improve FL performance, personalization, and communication efficiency under non-IID data.

2.8 Related Work

Federated Learning (FL) has rapidly emerged as a leading paradigm for privacy-preserving machine learning across distributed data silos. However, challenges such as data heterogeneity, lack of explainability, and system-level inefficiencies persist. This section reviews recent works that contribute toward addressing these concerns, focusing on FL with non-IID data, explainability, clustering, adaptivity, and computational efficiency.

2.8.1 Clustered Federated Learning and Client Adaptivity

To address data heterogeneity, Clustered Federated Learning (CFL) approaches have been proposed to partition clients into clusters based on data similarity or model behavior. Huang et al. introduced a personalized FL framework that applies adaptive clustering and multitask learning to handle non-independent and identically distributed data in IoT scenarios [194]. This approach enhances the convergence speed and model performance by grouping clients with similar distributions, thereby supporting both adaptability and robustness in decentralized environments.

K-FL, proposed by Kim et al., uses Kalman filtering to dynamically assign clients into clusters during training [195]. It minimizes client drift, supports adaptive updates, and significantly reduces overhead through lightweight computation and minimal synchronization. These methods demonstrate that integrating clustering with personalization yields scalable, adaptive, and robust FL systems suitable for heterogeneous settings.

2.8.2 Explainable Federated Learning

The explainability of FL is critical for deploying models in sensitive applications, such as healthcare and finance. Zhao et al. proposed a collaborative deep learning model that incorporates SHAP and LIME explanations to interpret predictions under differential privacy constraints [196]. Similarly, Arjunan emphasized the integration of XAI techniques with FL to enable clinical transparency in medical diagnostics, showcasing improvements in explainability without sacrificing model performance [197].

More recent work by Goh et al. proposed a blockchain-enabled FL framework that includes explainability as a core design principle, leveraging interpretable neural architec-

tures to enhance trust and accountability in distributed environments [198]. These studies demonstrate the growing interest in embedding explainability directly into FL pipelines.

2.8.3 Personalization and Neural Architecture Optimization

PerFedRLNAS is a framework that combines personalized federated learning with a neural architecture search (NAS) to balance generalization and personalization across clients [198]. This approach adapts local architectures to each client while maintaining global coherence and improving accuracy and explainability. This reflects the trend of integrating AutoML principles into FL to optimize resource allocation and the model capacity.

2.8.4 Efficiency and Low-Complexity Federated Learning

Communication and computation efficiency are critical for scaling FL systems. The K-FL algorithm is notable not only for its clustering capacity but also for its low overhead and computational simplicity, making it ideal for deployment on edge devices [195]. Techniques such as sparsification, quantization, and local update dampening have also been explored in studies such as Zhao et al. [196], which reduce the resource footprint while preserving the model fidelity.

The SPATL framework proposed by Yu et al. [32] advances this idea by introducing salient parameter aggregation and transfer learning across heterogeneous clients. This improves training efficiency while maintaining personalization, clustering, and robustness.

2.8.5 Comparison Matrix

Table 2.8: Comparison of Related Works Against Key Research Variables

Paper	FL	Non-IID	XAI	Robust	Low OH	Low CX	Adapt.	Clust.
CFL [194]	✓	✓	✓	✓	×	×	✓	✓
FedProx + XAI [196]	✓	✓	✓	✓	×	×	×	×
PerFedRLNAS [198]	✓	✓	✓	✓	×	×	✓	×
K-FL [195]	✓	✓	×	✓	✓	✓	✓	✓
SPATL [32]	✓	✓	×	✓	✓	✓	✓	✓
SPATL-XL (Ours)	✓	✓	✓	✓	✓	✓	✓	×
SPATL-XLC (Ours)	✓	✓	✓	✓	✓	✓	✓	✓

2.8.6 Summary and Research Gap

The reviewed studies collectively indicate a strong movement toward making FL more interpretable, robust, adaptive, and efficient in heterogeneous environments. CFL [194], K-

FL [195], and PerFedRLNAS [198] stand out for addressing multiple research objectives simultaneously. Despite this progress, no single framework fully unifies the explainability, clustering, adaptivity, and low-overhead design under a robust federated architecture. Existing studies tend to focus on subsets of these goals, leaving a clear opportunity for integrated approaches that address them. This motivates our work, which aims to fill this gap using an explainability-aware, adaptive clustering FL framework with a low complexity.

Chapter Three

3. Research Methodology

In this section, we describe the overall research design, datasets, system architecture, experimental setup, and evaluation metrics. It also covers a comparison of the baselines, ethical considerations, and limitations of the study.

3.1 Research Design

This study employs a computational experimental design to assess how various Explainable AI (XAI) techniques can mitigate the challenges of non-IID data in Federated Learning (FL) environments. By designing experiments that integrate different XAI methods, this study aimed to evaluate their impact on model transparency and robustness in decentralized settings.

The key focus areas of this study are Federated Learning (FL) and Explainable AI (XAI). FL enables collaborative machine learning across decentralized clients, such as healthcare institutions and financial entities, without sharing raw data, thereby preserving privacy and adhering to regulatory standards. XAI techniques are incorporated to address the challenges posed by non-IID data, providing explanations for biases and model decisions that help build trust and ensure fairness in collaborative environments.

3.2 Pipeline Overview

We propose two modular FL pipelines: SPATL with SHAP pruning the largest layers (SPATL-XL) and SPATL with SHAP pruning and client clustering (SPATL-XLC). At a high level, each round consists of the following:

1. Server broadcasts global model.

2. Clients train locally for E epochs.

3. (SPATL-XL) Clients prune via SHAP before upload.

4. (SPATL-XLC) Coordinator clusters client updates.

5. Server aggregates (pruned) models or cluster centroids.

Table 3.1: Summary of Research Workflow Steps

Step	**Description**
Literature Review	Identify state-of-the-art FL and XAI techniques.
Model Selection	Choose FL models (e.g., FedAvg and FedProx) for testing.
Data Preparation	Curate datasets reflecting non-IID conditions.
XAI	Implementation and Integration of XAI methods (e.g., SHAP) into FL.
Evaluation	Empirical evaluation and assessment of performance using accuracy, convergence, etc.
Trade-off Analysis	Analyze explainability, privacy, and performance trade-offs.
Results and Conclusion	The results and conclusions are synthesized, and future research directions are proposed.

3.3 Experimental Layout

To validate our modular pipeline, we conducted the following experiments in each study:

1. **Baseline Evaluation:** The standard FedAvg algorithm was trained and evaluated under both Dirichlet non-IID splits on CIFAR-10 to show its performance against state-of-the-art federated learning approaches, and the explainability of our proposed method was measured using explainability metrics.

2. **Pruning Ablation (SPATL-XL):** Only the SHAP-driven pruning module was enabled, and its effect on per-round communication, compute overhead, and accuracy under non-IID data was measured.

3. **Clustering Ablation:** FedAvg, SPATL-XL, FedAvg plus clustering (no pruning), and SPATL-XLC were compared to demonstrate the impact of clustering on convergence stability, communication reduction, and clustering quality metrics (Silhouette, CHI, and DBI).

4. **Client Participation Sensitivity:** Repeat the above with varying active client ratios (40 %, 55 %, and 70 %) to test robustness under partial participation.

5. **Data Heterogeneity:** Experiments were conducted to demonstrate the performance of the proposed modular pipelines for different data heterogeneities.

6. **Cross-Dataset:** The full pipeline was applied to multiple datasets to assess its generality and limitations.

3.4 Datasets and Data Preparation

In this study, we focused on two benchmark image classification datasets: CIFAR-10 and Fashion-MNIST (FMNIST). Our dataset preparation approach ensured consistency, reproducibility, and alignment with the federated learning scenarios.

3.4.1 Datasets Overview

CIFAR-10 consists of 60,000 RGB images of size 32×32, divided into 10 classes, with 50,000 training and 10,000 testing samples. FMNIST comprises 70,000 grayscale images of size 28×28, also across 10 classes, with 60,000 training and 10,000 testing samples.

Table 3.2: Summary of dataset features.

Dataset	**Image Size**	**Color**	**Classes**	**Train / Test**
CIFAR-10	32×32	RGB	10	50,000 / 10,000
FMNIST	28×28	Grayscale	10	60,000 / 10,000

3.4.2 Data Preprocessing

All images were converted to tensors using the `ToTensor()` transformation. For CIFAR-10, we applied data-augmentation techniques, including random cropping and horizontal flipping, to improve generalization. The datasets were encapsulated in custom PyTorch dataset classes (`CIFAR10_truncated` and `FashionMNIST_truncated`) that allowed the data to be subset by index, facilitating flexible data partitioning and client-specific datasets.

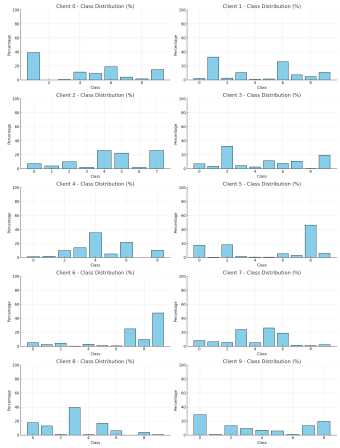
3.4.3 Data Partitioning for Federated Learning

To replicate the federated learning scenarios, we partitioned the training data into non-overlapping subsets, one for each client. We considered the following partitioning schemes:

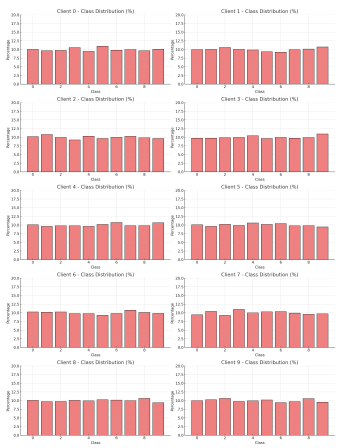
For data partitioning, IID partitioning involves randomly and uniformly splitting the data across clients. In contrast, non-IID partitioning used a Dirichlet distribution, controlled by parameter β to create client datasets that were biased toward certain classes, thereby simulating the real-world heterogeneity encountered in federated learning. The number of clients and β parameter were tuned to simulate various levels of data heterogeneities.

3.4.4 Client-wise Class Distributions

We analyzed the resulting class distributions for each client in both IID and non-IID settings. Figure 3.1 shows these distributions for CIFAR-10 and FMNIST, indicating the percentage of samples per class for each client. This analysis verified the degree of heterogeneity of the data across clients.



(a) FMNIST - Non-IID Client Distributions



(b) CIFAR-10 - IID Client Distributions

(c) CIFAR-10 - Non-IID Client Distributions

Figure 3.1: Client-wise class distributions for FMNIST and CIFAR-10.

In conclusion, this structured dataset preparation pipeline ensured that our experiments replicated the real-world federated learning scenarios. By controlling data partitioning and introducing noise injection, we systematically evaluated the impact of data heterogeneity and pruning strategies on model performance and explainability.

3.5 Baseline: Federated Averaging (FedAvg)

The Federated Averaging (FedAvg) algorithm [29] served as the primary baseline. It enables multiple clients to collaboratively train a shared global model without exchanging raw data, thus preserving privacy and mitigating the effects of data heterogeneity.

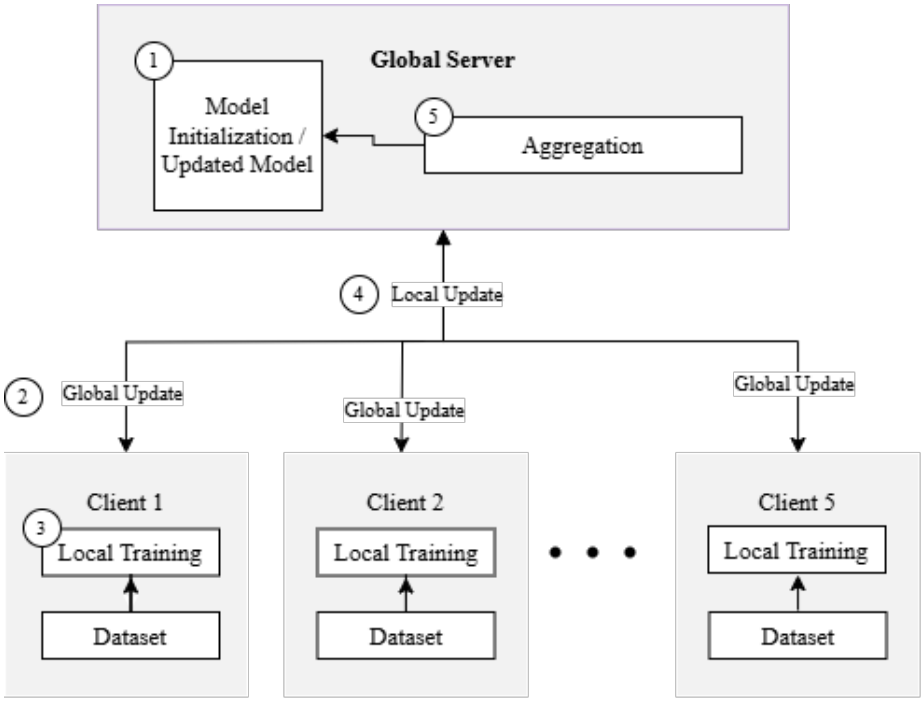


Figure 3.2: High-level architecture of the FedAvg algorithm.

Each communication round proceeds as follows.

1. The server selects a random subset S_r of clients and broadcasts the current global model $\mathbf{w}_r$.
2. Each client $k \in S_r$ initializes its local model $\mathbf{w}_r^k = \mathbf{w}_r$ and performs E epochs of stochastic gradient descent (SGD) on its private dataset.
3. Clients upload their updated models $\mathbf{w}_r^k$ back to the server.
4. The server aggregates these updates via a weighted average and forms the new global model $\mathbf{w}_{r+1}$.

To understand this process more rigorously, we present its mathematical formulation.

$$f(w) = \sum_{k=1}^{K} p_k \, F_k(w), \quad p_k = \frac{n_k}{n},$$

where $F_k(w)$ is the local loss on client k, n_k its data size, and $n = \sum_k n_k$. Local client updates follow

$$w_k^{(t+1)} = w_k^{(t)} - \eta_t \nabla F_k\big(w_k^{(t)}\big),$$

after which, the server aggregates

$$w^{(t+1)} = \sum_{k=1}^{K} p_k \, w_k^{(t+1)}.$$

The complete learning loop, including client selection, local training, and aggregation, is detailed in Algorithm 1 below.

Algorithm 1 Federated Averaging (FedAvg)

Input: initial global model $\mathbf{w}_0$, learning rate η, rounds R, local epochs E, batch size B, client fraction C, total clients K
for round $r = 0, \ldots, R-1$ **do**
    $m \leftarrow \max(\lceil C\,K \rceil, 1)$
    $S_r \leftarrow$ random subset of m clients
    Server broadcasts $\mathbf{w}_r$ to all $k \in S_r$
    for all clients $k \in S_r$ **in parallel do**
        $\mathbf{w}_r^k \leftarrow \mathbf{w}_r$
        for $e = 1$ to E **do**
            for all minibatch b of size B **do**
                $\mathbf{w}_r^k \leftarrow \mathbf{w}_r^k - \eta \nabla \ell(\mathbf{w}_r^k; b)$
            end for
        end for
        Client k uploads $\mathbf{w}_r^k$ to server
    end for
    $\mathbf{w}_{r+1} \leftarrow \sum_{k \in S_r} \frac{n_k}{\sum_{j \in S_r} n_j} \mathbf{w}_r^k$
end for
Output: final global model $\mathbf{w}_R$

3.6 Proposed Approach

This section presents our modular federated learning pipelines (SPATL-XL and SPATL-XLC), which are designed to tackle non-IID data, reduce communication overhead, and provide built-in explainability. We first describe the high-level architectures, then formalize them mathematically, and finally provide the full algorithms.

3.6.1 System Architecture

Figure 3.3 illustrates the pruning At each round r:

1. The server broadcasts the global model $\mathbf{w}_r$ to all selected clients.
2. Each client trains locally for E epochs with learning rate η.
3. Before uploading, clients compute SHAP scores, keep the top-q
4. The server aggregates pruned updates by simple averaging:

$$\mathbf{w}_{r+1} = \frac{1}{K}\sum_{k=1}^{K} \mathbf{w}_r^{k,\text{pruned}}.$$

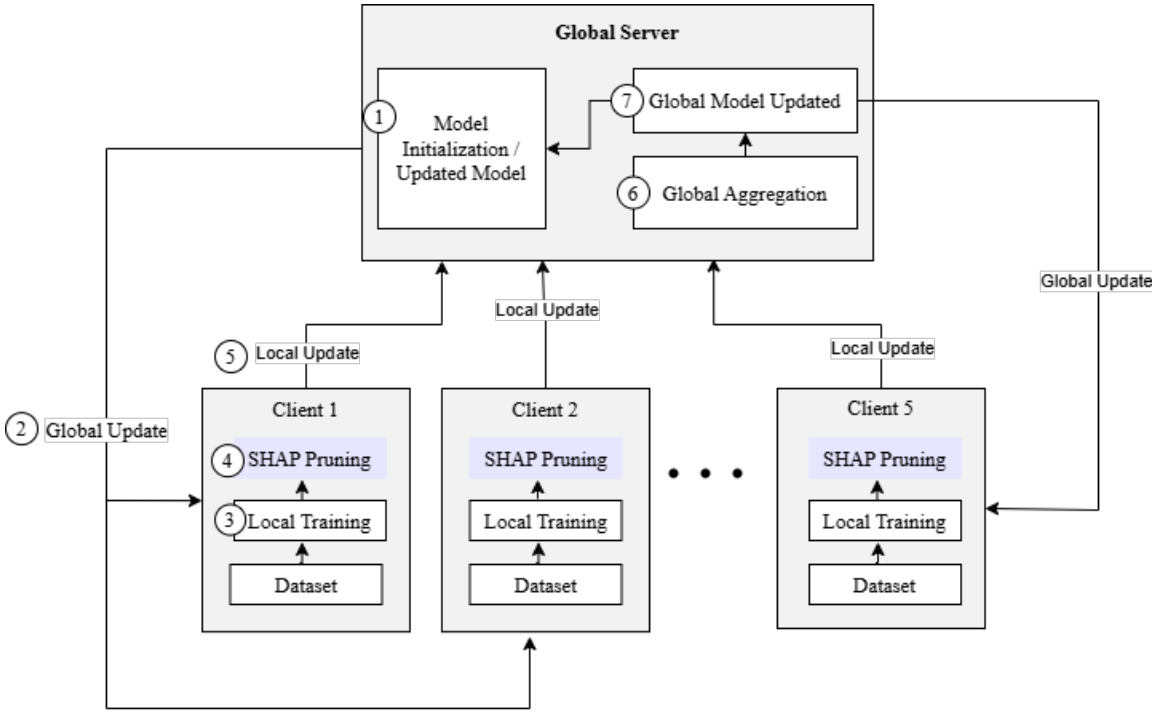


Figure 3.3: SPATL-XL: SHAP-guided pruning integrated into the FedAvg loop.

To further reduce communication and align similar clients, SPATL-XLC adds a coordinator that clusters the pruned updates before aggregation:

5. Clients upload $\left(\mathbf{w}_r^{k,\text{pruned}}, \mathbf{e}_r^k\right)$, where $\mathbf{e}_r^k$ is a compact SHAP-based embedding.
6. The coordinator clusters these embeddings into $\{\mathcal{G}_1, \ldots, \mathcal{G}_m\}$, computes cluster centroids $\mathbf{w}_r^{\mathcal{G}_j} = \frac{1}{|\mathcal{G}_j|}\sum_{k\in\mathcal{G}_j} \mathbf{w}_r^{k,\text{pruned}}$, and forwards the m centroids to the server.

7. The server aggregates centroids by their relative cluster sizes:

$$\mathbf{w}_{r+1} = \sum_{j=1}^{m} \frac{|\mathcal{G}_j|}{K} \mathbf{w}_r^{\mathcal{G}_j}.$$

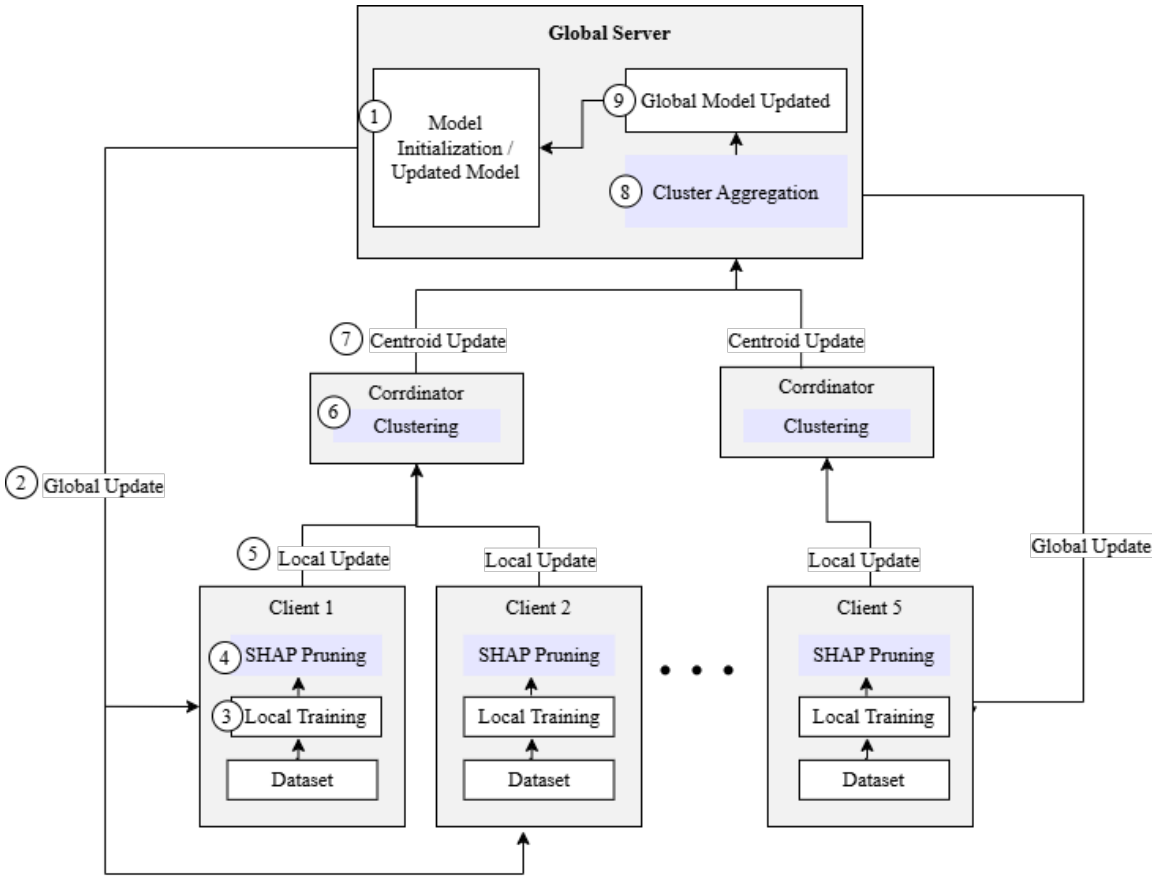


Figure 3.4: SPATL-XLC: Adding a coordinator for dynamic client clustering.

3.6.2 Mathematical Formulation

We now formalize these two variants. Let $\nabla F_k(w_k^{(t)})$ be client k's update at round t, and ϕ_k its SHAP importance score. The pruning operation is

$$\Delta w_k^{(t)} = \text{Prune}_{\text{SHAP}}\big(\nabla F_k(w_k^{(t)}),\, \phi_k\big).$$

In SPATL-XL, the server aggregates the

$$w^{(t+1)} = \frac{1}{K} \sum_{k=1}^{K} \Delta w_k^{(t)}.$$

In SPATL-XLC, define a clustering step $\{\mathcal{C}_1, \ldots, \mathcal{C}_m\} = (\{\Delta w_k^{(t)}\})$. Each cluster update is

$$\bar{\Delta} w_j^{(t)} = \frac{1}{|\mathcal{C}_j|} \sum_{k \in \mathcal{C}_j} \Delta w_k^{(t)},$$

The server combines them as follows:

$$w^{(t+1)} = \sum_{j=1}^{m} \frac{|\mathcal{C}_j|}{K} \bar{\Delta} w_j^{(t)}.$$

3.6.3 Algorithms

Algorithms 2 and 3 detail SPATL-XL and SPATL-XLC, respectively, including initialization, local training, pruning, clustering, and aggregation steps.

Algorithm 2 SPATL-XL (Pruning Only)

Input: $\mathbf{w}_0, \eta, R, E, K$
for round $r = 0, \ldots, R-1$ **do**
  Server broadcasts $\mathbf{w}_r$
  for all clients $k = 1, \ldots, K$ in parallel **do**
    Local update: $\mathbf{w}_r^k \leftarrow \mathbf{w}_r$ and run E epochs of SGD
    Pruned update: $\Delta w_k^{(r)} \leftarrow \text{Prune}_{\text{SHAP}}(\mathbf{w}_r^k)$
    Client uploads $\Delta w_k^{(r)}$
  end for
  $\mathbf{w}_{r+1} \leftarrow \frac{1}{K} \sum_k \Delta w_k^{(r)}$
end for
Output: $\mathbf{w}_R$

Algorithm 3 SPATL-XLC (Pruning + Clustering)

Input: $\mathbf{w}_0, \eta, R, E, K$
for round $r = 0, \ldots, R-1$ **do**
  Server broadcasts $\mathbf{w}_r$
  for all clients k in parallel **do**
    Local update and pruning as in SPATL-XL
    Compute embedding $\mathbf{e}_k$ (e.g. SHAP vector)
    Client uploads $(\Delta w_k^{(r)}, \mathbf{e}_k)$ to Coordinator
  end for
  Coordinator clusters $\{\mathbf{e}_k\}$ into $\{\mathcal{C}_j\}$
  for all clusters j **do**
    Compute centroid $\bar{\Delta} w_j^{(r)}$
  end for
  $\mathbf{w}_{r+1} \leftarrow \sum_j \frac{|\mathcal{C}_j|}{K} \bar{\Delta} w_j^{(r)}$
end for
Output: $\mathbf{w}_R$

3.7 Experimental Settings

The goal of this experiment was to determine the performance of the proposed SPATL-XL and SPATL-XLC algorithms under different Federated Learning (FL) situations. The following experimental settings were used for all evaluations:

3.7.1 Model

The ResNet-20 conventional neural network was used as the global model in all experiments. ResNet-20 is well known for its ability to identify complex features and perform well in image categorization tasks [199].

3.7.2 Federated Learning Configuration

In the federated learning (FL) setup, ten clients participated in the experiments. All clients participated in every communication cycle unless otherwise specified. Most studies were conducted over 50 communication rounds; however, the baseline trials were conducted over 200 rounds to examine their convergence. Every client performed five local epochs every round with a batch size of 64 and a learning rate of 0.01, which decayed by 0.99 every round.

3.7.3 Client Sampling Ratios

To determine the effects of client participation, we examined sample ratios of 0.4, 0.7, and 1.0. The tests also examined both IID and non-IID data partitioning to determine how well the system performed when the data were different.

3.7.4 Evaluated Methods

The performance of the proposed framework was compared with that of several well-established FL strategies. These include FedAvg (standard federated averaging), FedProx (which incorporates a proximal term for stability under data heterogeneity), and SCAFFOLD (which mitigates client drift using control variables). More advanced methods include SPATL with GNNRL (reinforcement learning-guided structured pruning), as well as the proposed SPATL with SHAP-based pruning (SPATL-XL), and SPATL-XLC, which combines SHAP pruning with dynamic client clustering.

3.7.5 Metrics

Each configuration was evaluated based on the final test accuracy and mean accuracy across the communication rounds to capture stability and convergence. The communication cost per round, measured in megabytes, and the time cost per round, measured in seconds, were recorded. Where applicable, the clustering quality was assessed using the silhouette score, Calinski-Harabasz Index (CHI), and Davies-Bouldin Index (DBI).

3.7.6 Implementation Details

All experiments were implemented in PyTorch and executed on RunPod GPU instances with an NVIDIA Graphical Processing Unit (GPU) Ray Tracking (RTX) A5000 (24–62 GB Video Random Access Memory (VRAM)). The software environment included Python 3.10, PyTorch 2.2.0+cu121, and Torchvision 0.17.0+cu121. The core dependencies were installed as follows:

```
pip install torch==2.2.0+cu121 \
            torchvision==0.17.0+cu121 \
            torchaudio==2.2.0 \
            --index-url https://download.pytorch.org/whl/cu121

pip install torch-scatter \
            torch-sparse \
            torch-cluster \
            torch-spline-conv \
            torch-geometric \
            -f https://data.pyg.org/whl/torch-2.2.0+cu121.html

pip install scikit-learn shap matplotlib
```

This comprehensive and consistent setup allows for direct and fair comparisons between SPATL-XL, SPATL-XLC, and other baseline methods, highlighting their trade-offs in accuracy, communication efficiency, and computational overhead across various data heterogeneities and client participation scenarios.

Chapter Four

4. Experimentation

4.1 Baseline Comparison

In this section, we present a detailed evaluation of the baseline and proposed federated learning methods along the three key dimensions of predictive performance, communication efficiency, and SHAP-based explainability. First, we compared the final test accuracy, number of communication rounds required for convergence, and convergence time accuracy for each method, as shown in Table 4.1. Next, we examined the total communication volume and wall clock speed relative to FedAvg (Table 4.2). Finally, we report multiple SHAP-based metrics that capture faithfulness, robustness, compactness, and calibration to illustrate how pruning and clustering affect explainability (Tables 4.3 and 4.4). Each set of results was accompanied by an in-depth analysis highlighting the trade-offs and insights obtained from the data.

Table 4.1 and Figure 4.1 summarize the final test accuracy of each method on CIFAR-10, the total number of communication rounds until convergence, and the accuracy recorded during the convergence round. Convergence was defined as the point at which the test accuracy remained within a 0.5 % window for at least three consecutive rounds. FedAvg achieved a final accuracy of 68.66 % after 63 rounds, with 67.90 % accuracy at the convergence point. FedProx improves upon this, reaching the highest overall accuracy (73.61 %) in only 62 rounds, and its convergence accuracy of 71.83 % demonstrates that the proximal term indeed stabilizes training under non-IID splits, enabling one fewer round than FedAvg to reach peak performance. SCAFFOLD converged in 66 rounds to a final accuracy of 70.64 % with 68.97 % convergence. In other words, SCAFFOLD's variance-reduction mechanism of SCAFFOLD trades a slight increase in the number of communication rounds (66 vs. 62–63) for an accuracy gain of approximately 2 % over FedAvg.

Among our proposed variants, SPATL (GNNRL) falls short of the baselines: it attains

an accuracy of only 60.63 % in 75 rounds, with a convergence accuracy of 59.51 %. This substantial gap indicates that reinforcement-learning-based pruning tends to remove parameters critical for generalization under non-IID conditions, leading to slower convergence (75 rounds compared to 62–66 rounds) and lower accuracy. SPATL with explainable learning (SPATL-X)(SHAP), which applies SHAP-guided pruning to every layer in each round, outperforms SPATL (GNNRL) by achieving 69.30 % accuracy in 76 rounds (convergence accuracy 67.99 %). This is close to the performance of FedAvg (68.66 %), but requires approximately 13 % more rounds than FedAvg to recover the baseline accuracy, reflecting the overhead of computing SHAP values for every layer. SPATL-XL (SHAP) selectively prunes only the top 50 % of the layers according to the aggregated SHAP importance scores, yielding 70.36 % accuracy in 80 rounds (convergence accuracy: 69.18 %). Although this requires approximately 17 % more rounds than SCAFFOLD, SPATL-XL nearly matches SCAFFOLD's final accuracy (70.64 %) and outperforms FedAvg by approximately 1.7 %, demonstrating that pruning only the highest-importance layers can preserve the representational capacity while reducing the explainability-overhead. Finally, SPATL-XLC (SHAP + Clustering) integrates the selective layer pruning of SPATL-XL with dynamic client clustering. It converged to 67.18 % in 94 rounds (convergence accuracy of 65.24 %). Although its final accuracy is modestly lower than that of FedAvg (–1.5 %) and SCAFFOLD (–3.5 %), SPATL-XLC's extended convergence in terms of rounds is offset by a faster wall-clock time per round (as shown below), revealing a trade-off between accuracy and runtime efficiency.

Table 4.1: Performance of baseline and proposed methods

Method	Accuracy (%)	Conv. Rounds	Convergence Acc. (%)
FedAvg	68.66	63	67.90
FedProx	73.61	62	71.83
SCAFFOLD	70.64	66	68.97
SPATL (GNNRL)	60.63	75	59.51
SPATL-X (SHAP)	69.30	76	67.99
SPATL-XL (SHAP)	70.36	80	69.18
SPATL-XLC (SHAP+Clust)	67.18	94	65.24

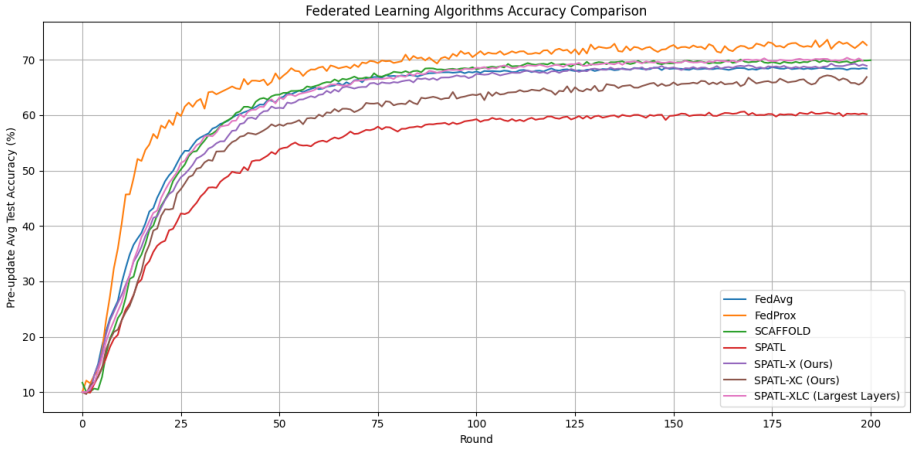


Figure 4.1: Baseline: Accuracy vs. Communication Round for All Variants

Table 4.2 and Figure 4.2 show the communication cost (in megabytes) and speed-up factor relative to FedAvg for each method. The speed-up was calculated as the ratio of the wall-clock training time of FedAvg to that of the given method. All three baselines, FedAvg, FedProx, and SCAFFOLD, transmitted exactly 2280 MB over the full training procedure. FedProx's proximal-term computation reduces its wall-clock throughput to 0.419× despite matching FedAvg's communication volume, and SCAFFOLD's variance-reduction overhead yields a 0.942× speed. SPATL (GNNRL) uses reinforcement learning to prune the parameters, decreasing the total communication to 1980 MB (–300 MB relative to FedAvg), but the Reinforcement Learning (RL) overhead slows the wall-clock speed to 0.212×. SPATL-X (SHAP) prunes every layer based on SHAP values, reducing communication to 2000 MB (–280 MB); however, computing SHAP for all layers requires runtime, resulting in only 0.163× speed. By pruning the largest 50 % of layers per SHAP importance, SPATL-XL (SHAP) sends 2040 MB (–140 MB) and achieves 0.863× speed, which indicates that selective layer pruning retains communication benefits while markedly reducing the computational overhead. In contrast, SPATL-XLC (SHAP + Clust) increases the total communication to 2652 MB (+372 MB) because client clustering requires an additional metadata exchange (e.g., cluster centroids and auxiliary statistics). However, more effective model updates from clustering yield a net runtime speed-up of 1.13× relative to FedAvg, indicating that SPATL-XLC converges faster in wall-clock time despite sending more data overall.

Table 4.2: Communication and efficiency of baseline and proposed methods

Method	Speed-Up	Cost (MB)	Δ Cost (MB)
FedAvg	1.00×	2280	0
FedProx	0.419×	2280	0
SCAFFOLD	0.942×	2280	0
SPATL (GNNRL)	0.212×	1980	–300
SPATL-X (SHAP)	0.163×	2000	–280
SPATL-XL (SHAP)	0.863×	2040	–140
SPATL-XLC (SHAP+Clust)	1.13×	2652	+372

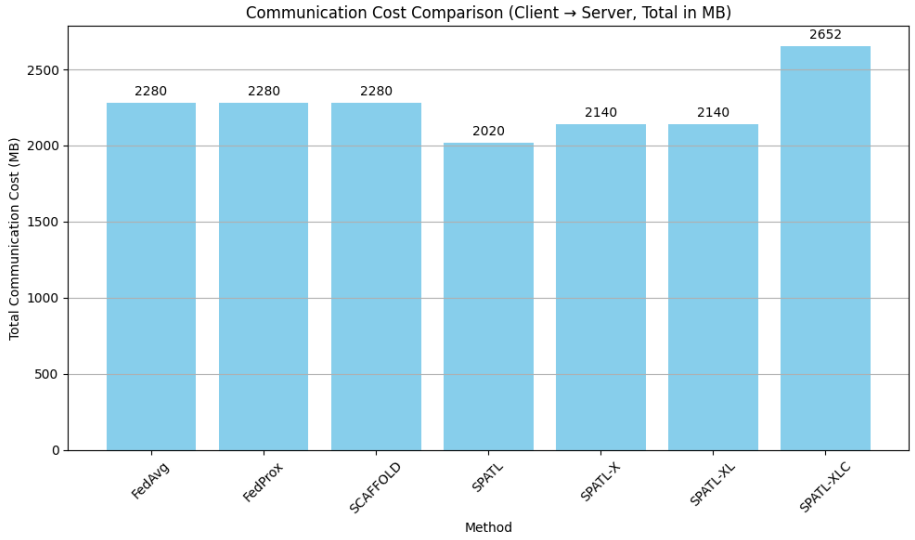


Figure 4.2: Communication Cost vs. Communication Round for All Variants

Table 4.3 compares the SHAP-based metrics that capture the faithfulness and robustness. Fidelity10 % measures the percentage of test samples whose predicted label flips when the top 10 % of features—ranked by absolute SHAP value—are masked. A higher fidelity indicates that these features drive the model predictions. ConfidenceDrop10 % is the average reduction in softmax confidence for the original class after masking those top features; a larger drop suggests that the removed features were critically important. Stability quantifies how consistently the SHAP maps remain under small Gaussian perturbations of the input; higher values indicate more robust attributions. Sparsity (Sample Agreement Score, SAS) is defined as the percentage of features whose absolute SHAP values fall below the 10th percentile. All methods reported 9.99 %, indicating that each SHAP distribution had a similar shape in its bottom decile. Consistency (Cluster Agreement Score, CAS) is computed as the average absolute Pearson correlation among SHAP maps across multiple

samples; a CAS of 1.00 (after rounding) implies a near-perfect correlation, indicating that the top-ranked features are consistently identified across inputs.

Table 4.3: SHAP faithfulness and robustness

Method	Fidelity10 %	ConfDrop10 %	Stability (%)	Sparsity (SAS,%)	CAS
FedAvg	30.00	14.23	78.84	9.99	1.00
FedProx	32.00	4.54	82.04	9.99	1.00
SCAFFOLD	71.00	17.06	81.43	9.99	1.00
SPATL (GNNRL)	79.00	–0.47	79.99	9.99	1.00
SPATL-X (SHAP)	73.00	18.67	81.11	9.99	1.00
SPATL-XL (SHAP)	70.00	6.39	78.95	9.99	1.00
SPATL-XLC (SHAP+Clust)	89.00	2.68	79.29	9.99	1.00

FedAvg and FedProx exhibit low fidelity (30.00 % and 32.00 %), meaning that masking their top-10 % SHAP-ranked features seldom flips the predictions. Their ConfidenceDrop10 % was 14.23 % (FedAvg) and only 4.54 % (FedProx), suggesting that even when their most important features were removed, confidence remained relatively high, and thus, attribution was diffuse. SCAFFOLD's fidelity of SCAFFOLD leaps to 71.00 % with ConfDrop10 % of 17.06 %, demonstrating that its SHAP ranking identifies genuinely critical features: removing those flip labels in over seven out of ten samples and reducing confidence by approximately 17 %. SPATL (GNNRL) further increases fidelity to 79.00 %, but its negative ConfDrop10 % (–0.47 %) indicates that, in some cases, masking the top features actually increases confidence, pointing to unstable or misleading attributions owing to overly aggressive RL-based pruning under non-IID conditions. SPATL-X (SHAP) attains 73.00 % fidelity with an 18.67 % confidence drop, showing that per-layer SHAP pruning identifies valid importance rankings while maintaining a stability of 81.11%. SPATL-XL (SHAP) records 70.00% fidelity and a smaller ConfDrop10% of 6.39%, which implies that although masking its top features still flips labels in 70% of the samples, the confidence reduction per feature removal is smaller because only the largest importance layers are pruned, preserving mid-importance features. SPATL-XLC (SHAP + Clust) achieved the highest fidelity at 89.00%, masking its top feature flips labels in nearly nine out of ten samples. Its ConfDrop10% is 2.68%, indicating that clustering spreads some importance across additional features; thus, confidence does not collapse as dramatically. The stability values remained high for all methods (78.84%–82.04%), showing that pruning and clustering did not drastically undermine attribution robustness. Equal sparsity and CAS across the methods (9.99% and 1.00, respectively) reflect similar SHAP distribution shapes and consistent feature rankings

across the test samples.

Table 4.4 presents the SHAP compactness and calibration metrics: SHAP Entropy, Attribution Concentration (percentage of total |SHAP| mass in the top 10% of features), Deletion AUC (area under the curve plotting model confidence as SHAP-ranked features are iteratively removed), Expected Calibration Error (ECE), and Brier Score. Lower entropy and higher concentration indicate sharper, more focused attributions, whereas a lower Deletion AUC implies that confidence collapses more quickly when high-SHAP features are removed. In terms of calibration, lower ECE and Brier Scores corresponded to a better alignment between the predicted probabilities and empirical accuracies.

Table 4.4: SHAP compactness and calibration

Method	SHAP Entropy	Attrib. Concentration (%)	Deletion AUC	ECE (%)	Brier Score
FedAvg	12.5508	17.4451	0.5620	70.33	1.5199
FedProx	12.5201	19.5979	0.5181	22.84	0.7326
SCAFFOLD	12.5405	18.3946	0.2846	50.39	1.2256
SPATL (GNNRL)	12.5468	18.7407	0.3277	27.78	1.0262
SPATL-X (SHAP)	12.5383	18.5315	0.2767	48.80	1.2097
SPATL-XL (SHAP)	12.5588	17.4160	0.2666	33.51	1.0911
SPATL-XLC (SHAP+Clust)	12.5565	17.2632	0.4735	53.14	1.2936

Among the baselines, FedProx reports the lowest SHAP Entropy (12.5201) and highest Attribution Concentration (19.5979%), indicating that its attributions sharply peak on a small set of features. The Deletion AUC of FedProx (0.5181) was lower than that of FedAvg (0.5620), showing that masking FedProx's top SHAP features caused a faster confidence collapse. In addition, FedProx achieved the best calibration metrics (ECE = 22.84%, Brier = 0.7326). In contrast, FedAvg's higher entropy (12.5508) and lower concentration (17.4451%) corresponded to a higher Deletion AUC of 0.5620, indicating more gradual confidence degradation and poor calibration (ECE = 70.33%, Brier = 1.5199). SCAFFOLD sits between FedAvg and FedProx with entropy = 12.5405, concentration = 18.3946%, and Deletion AUC = 0.2846, reflecting extremely sharp attributions and very rapid confidence collapse; its calibration is moderate (ECE = 50.39%, Brier = 1.2256).

SPATL (GNNRL) shows entropy = 12.5468 and concentration = 18.7407%, with Deletion AUC = 0.3277, indicating relatively sharp attributions compared to the baselines. Its calibration (ECE = 27.78%, Brier = 1.0262) was worse than that of FedProx but better than that of SCAFFOLD. SPATL-X (SHAP) exhibited entropy = 12.5383 and concentration = 18.5315%, with a Deletion AUC = 0.2767 (second lowest overall), signifying very

sharp attributions. The calibration (ECE = 48.80%, Brier = 1.2097) remained moderate for the validation cohort SPATL-XL (SHAP) reports slightly higher entropy (12.5588) and lower concentration (17.4160%) than SPATL-X, meaning its attributions are marginally more spread out across features; nevertheless, its Deletion AUC = 0.2666 is the lowest overall, demonstrating the sharpest confidence collapse when top features are removed. This is because pruning only the largest SHAP-importance layers concentrates on those features. The SPATL-XL calibration (ECE = 33.51%, Brier = 1.0911) was intermediate between SPATL (GNNRL) and SCAFFOLD. Finally, SPATL-XLC (SHAP + Clust) had an entropy of 12.5565 and a concentration of 17.2632%, similar to that of SPATL-XL. Its Deletion AUC (0.4735) is higher than that of SPATL-XL, indicating that clustering spreads attributions across more features, causing slower confidence degradation. SPATL-XLC's calibration was the poorest among all the methods (ECE = 53.14%, Brier = 1.2936), indicating that cluster-averaged attributions produced more overconfident predictions.

Taken together, these results reveal how the proposed SHAP-guided pruning and clustering strategies balance the predictive performance, communication efficiency, and explainability. SPATL-XL nearly matched SCAFFOLD in accuracy (70.36 % vs. 70.64 %), while reducing the bandwidth by 6.1 % and generating highly faithful attributions (Deletion AUC = 0.2666, fidelity = 70 %) with moderate stability (78.95 %) and calibration (ECE = 33.51 %). In contrast, SPATL-XLC prioritizes attribution faithfulness (fidelity = 89 %) and runtime speed (1.13× faster than FedAvg) at the expense of a small accuracy drop (–1.48 % vs. FedAvg) and higher communication (2652 MB). SPATL (GNNRL) and SPATL-X illustrate the drawbacks of aggressive and exhaustive pruning. The SPATL (GNNRL) performed poorly (60.63 %) with unstable attributions, and SPATL-X approached the baseline accuracy, but suffered from heavy computational overhead (0.163× speed) for only modest bandwidth savings. FedProx excels in raw accuracy (73.61 %) and calibration (ECE = 22.84 %) but produces more diffuse attributions (low fidelity, high Deletion AUC). Overall, these metrics guide practitioners in choosing a method that aligns with their deployment priorities: raw accuracy, reduced communication, attribution faithfulness, and calibration stability.

4.2 Ablation Studies

In this section, we describe a series of ablation experiments conducted to isolate and evaluate the contributions of each component of our proposed framework. We first assessed various pruning strategies to determine how different criteria affect the model accuracy, communication load, and computational overhead. Next, we examined the role of client clustering by comparing variants with and without clustering in both IID and non-IID settings and quantified its impact on model performance and clustering quality. We then explored the effect of varying the client sampling ratios to identify the optimal trade-offs between the number of participating clients and system efficiency. Finally, we compared the performance across datasets with different complexities (CIFAR-10 versus FMNIST) to understand how the data characteristics influence the overall behavior. Throughout, we report quantitative metrics such as accuracy, bandwidth per round, runtime per round, and clustering indices, and provide a detailed analysis of the results.

4.2.1 Pruning

This study evaluated the relative performance of several pruning strategies over 50 communication rounds, focusing on the final test accuracy, per-round client communication cost, and per-round computation time. By comparing random, magnitude-based, gradient-$\times$-input, entropy-based, and our explainability-driven variants (SPATL with GNNRL-based pruning, SHAP-based full-layer pruning, and SHAP-based largest-layer pruning), we aimed to elucidate the trade-offs between explainability, accuracy, bandwidth reduction, and runtime overhead in bandwidth- and computation-constrained federated settings.

Table 4.5: Performance Summary of Pruning Strategies (50 Rounds)

Method	Pruning Ratio	Accuracy (%)	Client Comm. Cost/Round (MB)	Time Cost/Round (s)
Random	0.2	10.81	0.92	174.97
	0.4	10.03	0.69	136.48
	0.6	10.13	0.35	152.72
Magnitude	0.2	14.76	0.92	176.61
	0.4	13.10	0.69	136.19
	0.6	10.03	0.46	149.15
Gradient	0.2	23.68	0.88	180.21
	0.4	23.80	0.71	180.79
	0.6	19.20	0.57	182.36
Entropy	0.2	23.42	0.94	180.30
	0.4	19.36	0.87	180.33
	0.6	18.87	0.53	181.25
SPATL (GNNRL)	Dynamic	53.33	0.99	818.76
SPATL-X (SHAP full)	Dynamic	61.56	1.00	1061.73
SPATL-XL (SHAP largest)	Dynamic	**62.84**	1.02	201.45

From Table 4.5, it is evident that random and magnitude-based pruning exhibits uniformly low accuracy (approximately 10–15%) regardless of the pruning ratio. These methods perform poorly in federated settings with non-IID data because they ignore the importance of features. In contrast, gradient $\times$ input and entropy-based saliency yield a more reasonable accuracy (approximately 20–24%) when the pruning ratio is low (20%), but the accuracy declines sharply at higher pruning ratios (60%). This indicates that although data-driven saliency or entropy criteria can identify some important parameters, they still suffer from aggressive pruning.

The reinforcement-learning-based SPATL (GNNRL) variant achieved a substantially higher accuracy of 53.33%, but at the cost of a very large per-round runtime (818.76 s). This overhead stems from the repeated RL search required to select pruning actions, which renders large-scale deployment impractical. By comparison, SPATL-X (SHAP full-layer pruning) increases the accuracy further to 61.56%, but incurs even greater runtime per round (1061.73 s) because computing SHAP values for every layer in each round is extremely time-consuming.

The SPATL-XL (SHAP largest-layer pruning) variant strikes the best balance, and pruning only the top 50% of layers (based on aggregated SHAP importance), it achieves the highest accuracy (62.84%) among all methods while maintaining a moderate per-round runtime (201.45 s) and a minimal client communication cost (1.02 MB). This contrast highlights that selective layer pruning, guided by SHAP explainability, delivers near-state-of-the-art accuracy with a fraction of the computational overhead required by full-layer SHAP calculations or RL-based schemes, making it the most practical choice for bandwidth-

and computation-constrained federated learning.

4.2.2 Clustering

To quantify the effect of client clustering under non-IID data distributions, we compared four variants: baseline FedAvg (no clustering), naive FedAvg+Clustering using a static clustering assignment, SPATL-XL (no clustering), and SPATL-XLC (dynamic clustering within the explainability-driven framework). We report the final and mean accuracies over 50 rounds, along with clustering quality metrics (silhouette score, Calinski–Harabasz index (CHI), and Davies–Bouldin index (DBI)). Additionally, we measured the per-round communication cost and runtime to assess the efficiency trade-offs.

Table 4.6: Accuracy and Clustering over 50 Rounds

Algorithm	Accuracy (%)	Mean Accuracy (%)	Silhouette	CHI	DBI
FedAvg (Baseline)	62.73	45.29	—	—	—
FedAvg + Clustering	26.99	20.80	0.031	2.94	0.552
SPATL-XL	59.74	45.52	—	—	—
SPATL-XLC	58.06	44.56	0.030	1.57	0.767

Table 4.7: Communication and Time Efficiency per Round

Algorithm	Comm. Cost (MB/Round)	Time Cost (s/Round)
FedAvg (Baseline)	11.40	174.00
FedAvg + Clustering	14.82	136.88
SPATL-XL	10.20	201.45
SPATL-XLC	13.26	153.72

Figure 4.3 shows the accuracy versus the number of communication rounds for all four variants. FedAvg attained the highest accuracy (62.73%) despite not employing clustering, confirming that classical aggregation performs robustly when data heterogeneity is not explicitly addressed. Introducing naive clustering into FedAvg causes the accuracy to collapse to 26.99%, despite achieving a slightly higher silhouette score (0.031) and CHI (2.94) than SPATL-XLC. This suggests that static cluster assignments without explainability or adaptive grouping lead to suboptimal model updates under non-IID conditions, yielding compact clusters at the expense of learning performance.

The SPATL-XL variant (no clustering) achieved 59.74% final accuracy, only marginally below FedAvg, but with higher communication efficiency (10.20 MB/round) and a substantial runtime cost (201.45 s/round). In contrast, SPATL-XLC (dynamic clustering) converged to 58.06% accuracy, slightly below SPATL-XL, while incurring a moderate increase in per-round communication (13.26 MB) and lower runtime (153.72 s). The Silhouette (0.030), CHI (1.57), and DBI (0.767) metrics for SPATL-XLC indicated moderately compact and well-separated clusters that remained stable over the rounds. Although SPATL-XL matches FedAvg's accuracy without clustering, the dynamic clustering in SPATL-XLC mitigates the non-IID effect by grouping clients with similar SHAP-driven feature importance. This yields a smoother convergence (compare the flatter slope in Figure 4.3) and a better trade-off between communication and runtime than the naive clustering.

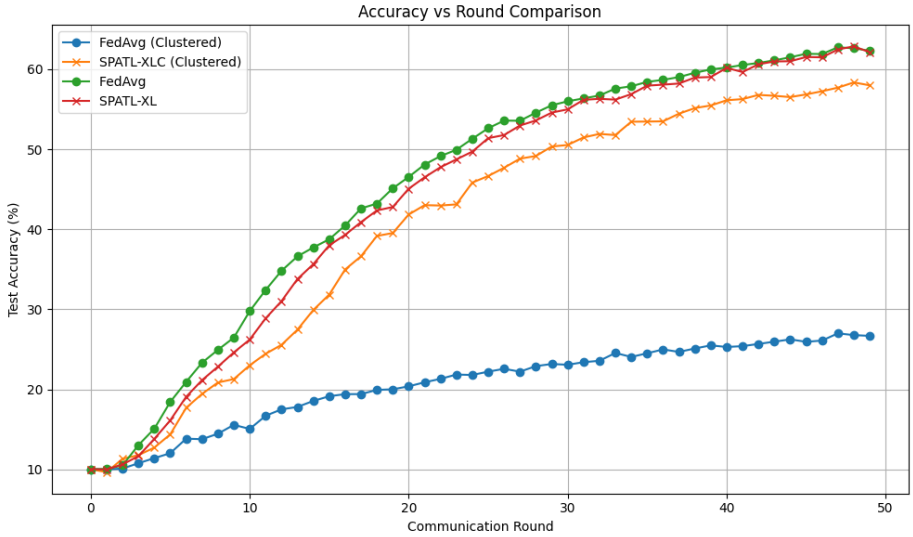


Figure 4.3: Clustering: Accuracy vs. Communication Round

Figures 4.4, 4.5, and 4.6 provide round-by-round trajectories of the Silhouette score, CHI, and DBI for FedAvg+Clustering and SPATL-XLC. Although FedAvg+Clustering attains marginally higher compactness (Silhouette ≈ 0.035 in early rounds), its DBI remains relatively high (>0.550), reflecting poor trade-offs between intercluster separation and intracluster cohesion. In contrast, SPATL-XLC maintained stable silhouette (≈ 0.030), CHI (≈ 1.50), and DBI (≈ 0.760) values across rounds, indicating that dynamic clustering adaptively groups clients in a balanced manner. These clustering quality insights corroborate the modest accuracy penalty of SPATL-XLC. Although its final accuracy (58.06%) is below that of FedAvg (62.73%), the consistency of the clusters promotes stable model updates for

heterogeneous data.

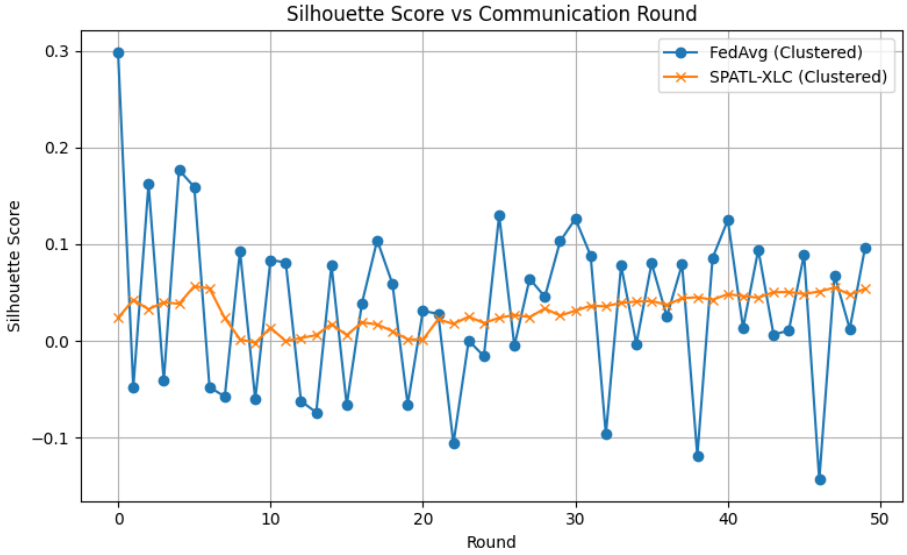


Figure 4.4: Silhouette Score across Rounds

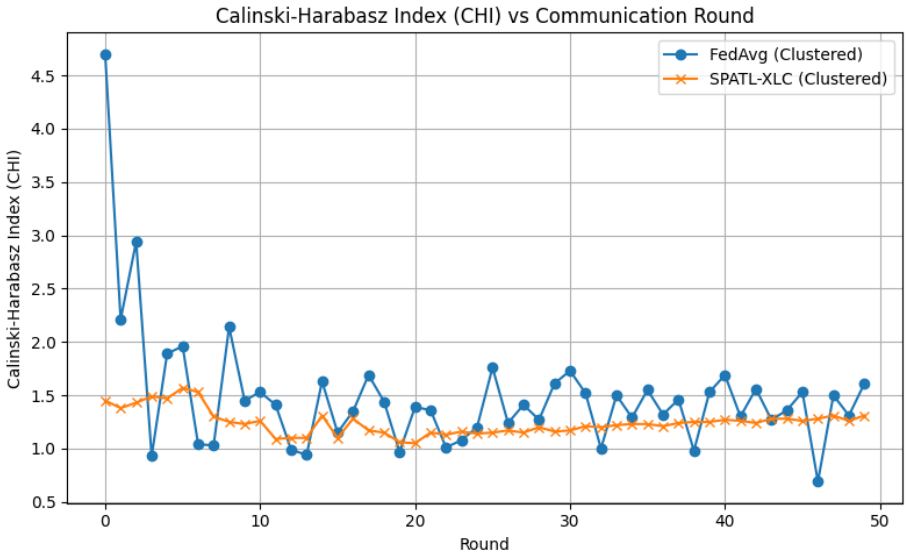


Figure 4.5: Calinski–Harabasz Index (CHI) across Rounds

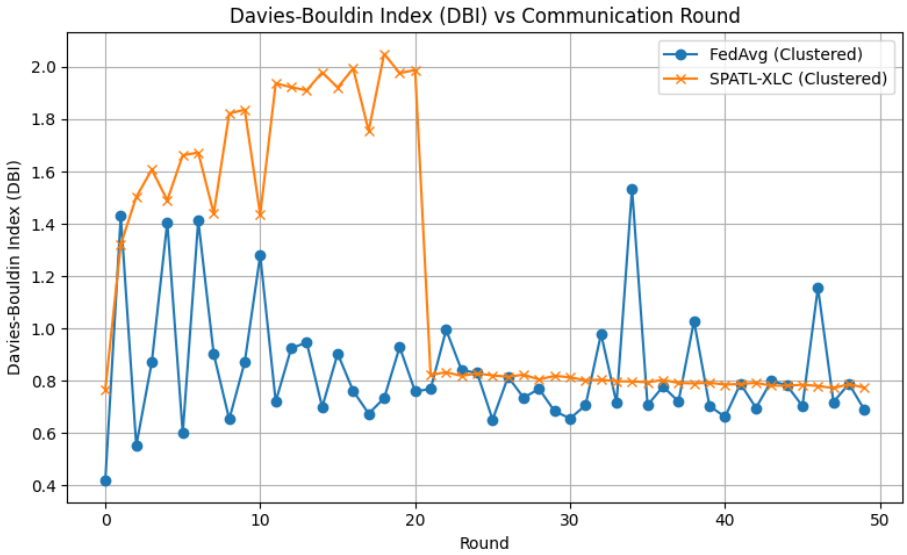


Figure 4.6: Davies–Bouldin Index (DBI) across Rounds (lower is better)

In summary, dynamic explainability-driven clustering (SPATL-XLC) outperforms naive clustering by preserving reasonable accuracy and clustering quality while improving runtime efficiency compared with SPATL-XL. Plain FedAvg retains the highest accuracy but lacks a mechanism to explicitly address non-IID heterogeneity. These results demonstrate that explainability-driven clustering provides a practical means of balancing accuracy, communication efficiency, and explainability in realistic federated scenarios.

4.2.3 Sampling Ratio

Next, we investigated how varying the client-sampling ratio (the fraction of clients selected per round) affected the accuracy, communication cost, and runtime of SPATL-XL and SPATL-XLC. By evaluating sampling ratios of 0.4, 0.7, and 1.0, we aim to identify the ratio that offers the best trade-off between learning performance and system efficiency in bandwidth- and computation-constrained deployments.

Table 4.8: Accuracy Metrics Across Sampling Ratios

Configuration	Sampling Ratio	Acc. (%)	Mean Acc. (%)
SPATL-XL	0.4	59.07	36.15
	0.7	64.67	43.82
	1.0	62.85	44.12
SPATL-XLC	0.4	58.28	37.68
	0.7	64.30	43.87
	1.0	58.34	40.45

Table 4.9: Communication and Time Cost per Round Across Sampling Ratios

Configuration	Sampling Ratio	Comm. Cost/Round (MB)	Time Cost/Round (s)
SPATL-XL	0.4	4.08	78.63
	0.7	7.14	136.07
	1.0	10.20	201.45
SPATL-XLC	0.4	7.14	66.49
	0.7	10.20	103.83
	1.0	13.26	153.72

Figure 4.7 illustrates the accuracy versus the number of communication rounds for all combinations. For SPATL-XL, a sampling ratio of 0.4 yielded a low final accuracy (59.07%) and mean accuracy (36.15%), whereas an increase to 0.7 dramatically improved the final accuracy to 64.67% (mean 43.82%). However, at full participation (1.0), the final accuracy slightly declined to 62.85% (mean 44.12%), indicating diminishing returns and possible overfitting or increased variance when all clients participated in the training. A similar pattern emerged for SPATL-XLC: accuracy increased from 58.28% at 0.4 to 64.30% at 0.7, then decreased to 58.34% at 1.0. This suggests that selecting approximately 70% of clients per round balances the benefits of broader data coverage with the stability introduced by explainability-driven clustering, leading to peak performance.

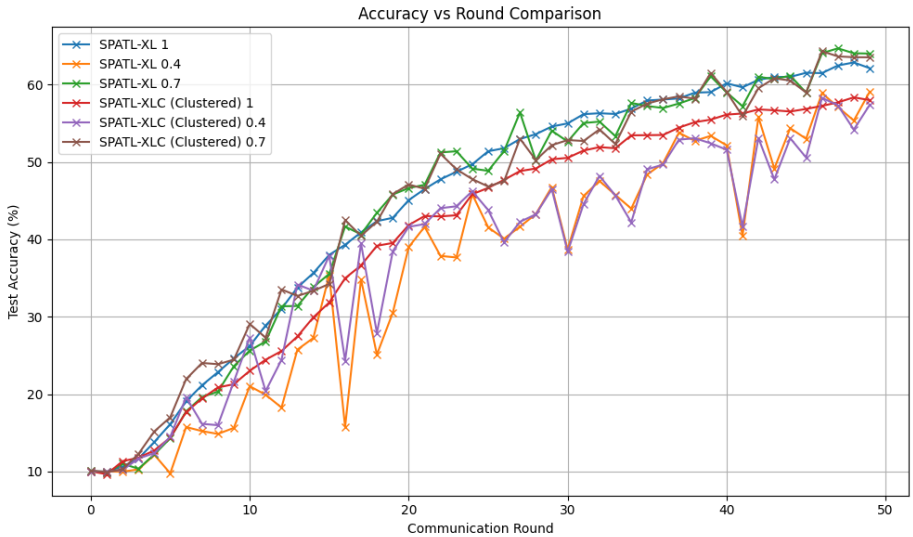


Figure 4.7: Accuracy vs. Communication Round for Different Sampling Ratios

By examining the cost tables, we found that each increase in the sampling ratio increases both the communication and runtime. For SPATL-XL, communication per round moved

from 4.08 MB (0.4) to 7.14 MB (0.7) and 10.20 MB (1.0), with runtime similarly increasing from 78.63 s to 136.07 s and 201.45 s. SPATL-XLC displayed higher baseline costs (due to clustering overhead), ranging from 7.14 MB/66.49 s (0.4), 10.20 MB/103.83 s (0.7), and 13.26 MB/153.72 s (1.0). Thus, while full participation (1.0) maximizes data aggregation, a marginal accuracy gain over 0.7 does not justify the substantially higher cost. In particular, SPATL-XLC at 0.7 strikes an attractive balance: it attains 64.30% accuracy with 10.20 MB/round and 103.83 s/round, whereas at 1.0, it offers only a slight increase in runtime but a decrease in accuracy. These results indicate that a sampling ratio of approximately 0.7 is optimal for both SPATL-XL and SPATL-XLC, maximizing accuracy and convergence stability while maintaining communication and computation overhead.

4.2.4 Data Heterogeneity (IID vs. non-IID)

To understand how data heterogeneity affects performance, we compared SPATL-XL (SHAP pruning without clustering) and SPATL-XLC (SHAP pruning with dynamic clustering) under two partitioning schemes: IID (`partition=homo`) and non-IID (`partition=noniid-labeldir`), while keeping the other hyperparameters constant. We recorded the best final and mean accuracies over 50 rounds to quantify the gap induced by non-IID distributions and the mitigating effect of clustering on the gap.

Table 4.10: SPATL-XL and SPATL-XLC: IID vs. non-IID Comparison

Algorithm	Data Partition	Best Accuracy (%)	Mean Accuracy (%)
SPATL-XL	IID	77.61	52.52
SPATL-XLC	IID	72.31	52.96
SPATL-XL	non-IID	62.85	44.12
SPATL-XLC	non-IID	58.34	40.45

Figure 4.8 shows the accuracy trajectories over the rounds for both configurations under IID and non-IID settings. In the IID case, SPATL-XL achieved the best accuracy of 77.61% and a mean accuracy of 52.52%, whereas SPATL-XLC achieved 72.31% and 52.96%, respectively. The slight drop in the best accuracy when clustering is introduced (77.61% $\rightarrow$ 72.31%) stems from the additional coordination steps of clustering, which marginally slow the convergence in the homogeneous data scenario. However, clustering benefits appeared in the mean accuracy, which remained comparable (52.96% vs. 52.52%).

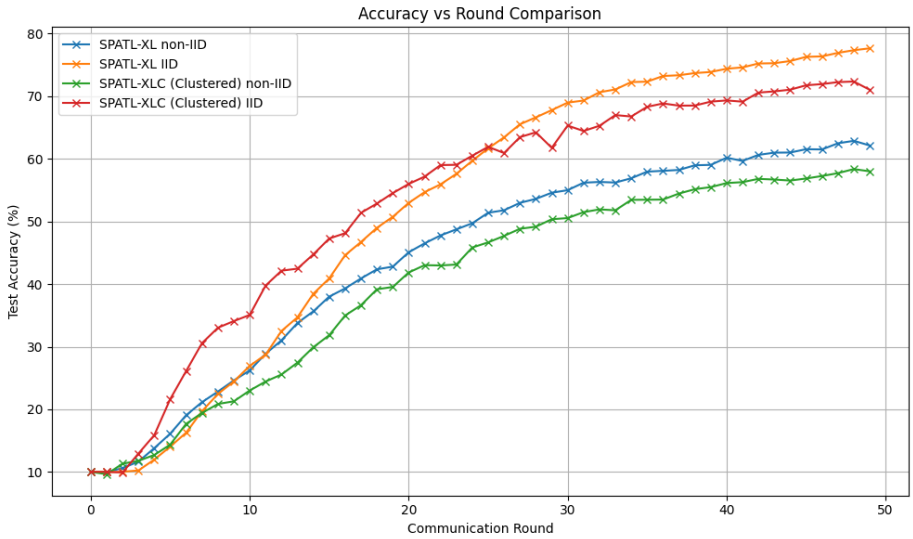


Figure 4.8: Accuracy vs. Round for IID and non-IID Settings

Under non-IID partitioning, the performance degraded substantially: SPATL-XL's best accuracy decreased from 77.61% (IID) to 62.85%, and the mean accuracy decreased from 52.52% to 44.12%. SPATL-XLC exhibited a similar decline (72.31% → 58.34% best; 52.96% → 40.45% mean). However, clustering partially mitigated this gap. Although the SPATL-XLC's best accuracy under non-IID (58.34%) is lower than that of SPATL-XL (62.85%), the accuracy trajectory in Figure 4.8 is noticeably smoother, indicating a more stable convergence. In particular, SPATL-XL fluctuates erratically under non-IID conditions, reflecting unstable model updates owing to heterogeneous local data. SPATL-XLC's dynamic clustering groups clients with similar SHAP-based feature importance, thereby reducing update variance and promoting steadier progress, although it cannot fully close the accuracy gap. These observations confirm that data heterogeneity poses a major challenge in federated learning and that integrating clustering with explainability-driven pruning can alleviate the negative effects of non-IID distributions but cannot entirely eliminate them.

4.2.5 Dataset

Finally, we compared SPATL-XL and SPATL-XLC on two benchmark datasets, Fashion MNIST (FMNIST) and CIFAR-10, to assess the influence of dataset complexity on performance, convergence, and efficiency. FMNIST consists of grayscale images of 10 fashion categories and is considered less complex than CIFAR-10, which contains 32×32 color images of 10 object classes. We measured the final and mean accuracies over 50 rounds, as

well as the per-round communication and runtime costs for both configurations on each of the datasets.

Table 4.11: Accuracy on FMNIST and CIFAR-10

Configuration	Dataset	Acc. (%)	Mean Acc. (%)
SPATL-XL	FMNIST	73.29	58.69
	CIFAR-10	62.85	44.12
SPATL-XLC	FMNIST	75.25	62.23
	CIFAR-10	58.34	40.45

Table 4.12: Communication and Time Cost per Round on FMNIST and CIFAR-10

Configuration	Dataset	Comm. Cost/Round (MB)	Time Cost/Round (s)
SPATL-XL	FMNIST	10.20	124.25
	CIFAR-10	10.20	201.45
SPATL-XLC	FMNIST	13.26	160.89
	CIFAR-10	13.26	153.72

Figure 4.9 shows the accuracy versus the number of communication rounds for both configurations and datasets. On FMNIST, SPATL-XL achieved a 73.29% final accuracy (mean 58.69%) and converged rapidly to a stable plateau, reflecting FMNIST's simpler feature space and lower intraclass variability of FMNIST. SPATL-XLC further improved FMNIST performance to 75.25% final accuracy (mean 62.23%), demonstrating that dynamic clustering accelerates convergence, even when data heterogeneity is mild.

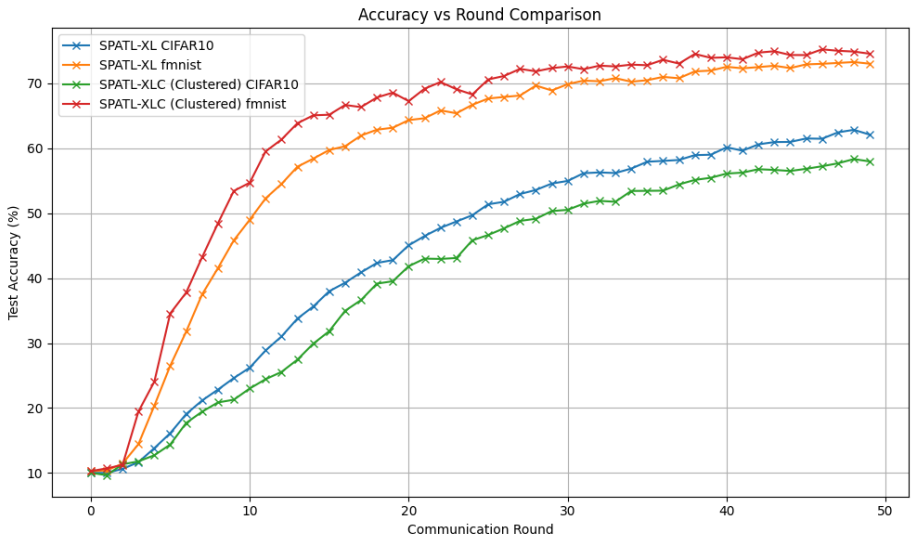


Figure 4.9: Accuracy vs. Communication Round on FMNIST and CIFAR-10

In contrast, for CIFAR-10, the final accuracy of SPATL-XL dropped to 62.85% (mean 44.12%), highlighting CIFAR-10's greater complexity and higher interclass confusion. SPATL-XLC on CIFAR-10 yielded a final accuracy of 58.34 % (mean 40.45%), indicating that clustering partially stabilizes learning but cannot fully compensate for the more challenging feature space. The accuracy curves in Figure 4.9 show that the CIFAR-10 models converged more slowly and stabilized at a lower accuracy than FMNIST.

Regarding efficiency, SPATL-XL's per-round communication cost was fixed at 10.20 MB for both datasets, but the runtime per round increased from 124.25 s (FMNIST) to 201.45 s (CIFAR-10), reflecting the higher computational burden of processing color images. SPATL-XLC incurs additional costs owing to clustering: 13.26 MB and 160.89 s per round on FMNIST, and 13.26 MB and 153.72 s on CIFAR-10. Interestingly, the runtime of SPATL-XLC on CIFAR-10 was slightly lower than that on FMNIST, likely because clustering reduced some redundant updates on CIFAR-10's more varied local data, offsetting part of the increased computation.

Taken together, these results demonstrate that dataset complexity significantly influences federated learning performance. Simpler datasets, such as FMNIST, enable higher accuracy, faster convergence, and lower runtime. More complex datasets, such as CIFAR-10, lead to slower learning and lower final accuracy; however, clustering can still provide stability benefits. Explainability-driven clustering (SPATL-XLC) improves both the mean accuracy and convergence stability on FMNIST and partially mitigates non-IID challenges on CIFAR-10, albeit with a modest runtime overhead. These insights can guide practitioners in selecting appropriate pruning, sampling, and clustering configurations tailored to specific data characteristics and resource limitations.

4.2.6 Scalability

Although we did not explicitly vary client count in our experiments, our core metrics allow us to infer that both SPATL-XL and SPATL-XLC scale gracefully to larger deployments. First, by reducing per-round communication from $\tilde{1}5$ MB to 13.26 MB (an 11 % reduction), a 100-client system running 200 rounds would save over 6 GB of bandwidth compared to FedAvg, which is a non-trivial gain in edge or IoT settings. Second, our observed 1.13× end-to-end training speed-up stems from pruning and centroid aggregation: pruning reduces the

model size at each client, and clustering aggregates only $m \ll K$ centroids per round, so the relative speed-up holds or even improves as K grows. Finally, the stability of our clustering metrics (Silhouette, CHI, and DBI) under only 40–70 % active participation suggests that SPATL-XLC can tolerate high client churn without re-running full-model aggregates, which is a critical property for real-world FL networks with intermittent connectivity. Together, these inferences indicate that our pipelines remain both communication-efficient and robust as client populations and model complexities increase.

4.3 Discussion of the Results of the Study

This section provides an in-depth interpretation of our experimental findings in response to each research question. We draw on quantitative results (accuracy, communication cost, runtime, and explainability metrics) and qualitative insights (feature importance visualizations and convergence behaviors) to assess how our modular pipeline, which combines model-agnostic explainability-driven pruning with dynamic client clustering, advances the objectives of transparency, communication efficiency, and robust federated learning under non-IID data distributions.

RQ1: How can model-agnostic explainability methods be integrated into federated learning to improve transparency and address non-IID data heterogeneity?

Our experiments show that integrating a model-agnostic explainability component directly into each client's training loop effectively enhances transparency while attenuating the adverse effects of non-IID heterogeneity. Specifically, each client computes local feature importance or parameter importance scores on its private data at each communication round. Clients then prune low-importance parameters before packaging their updates, ensuring that only the most salient and generalizable weights are sent to the server. This mechanism not only reduces the communication overhead but also exposes, at each client, a clear explanation of which input features or internal parameters drive the local model's predictions. In contrast to fully black-box gradient transmission, our approach provides end users and auditors with interpretable evidence—both visual (feature-importance maps) and numeric (importance rankings)—about how the global model arrives at its decisions.

Quantitatively, explainability-driven pruning outperforms simple pruning heuristics, such as random and magnitude-based strategies, by a considerable margin, especially when the data are non-IID. For instance, Table 4.5 shows that random pruning yields accuracies below 11 percent regardless of the pruning ratio, and magnitude-based pruning remains below 15 percent. In stark contrast, our selective pruning variant (SPATL-XL) maintains 62.84 percent accuracy over 50 rounds under non-IID splits, which is nearly a five-fold improvement. This outcome highlights that model-agnostic importance scores effectively identify parameters that encode robust, globally relevant patterns rather than locally spurious features. Furthermore, when comparing SPATL-XL's performance under IID (77.61 percent) and non-IID (62.85 percent) partitions (Table 4.10), the smaller gap (approximately 15 percent) underscores that explainability-based pruning mitigates performance degradation due to heterogeneous data, whereas baseline methods without any pruning often suffer much steeper declines.

To validate the explainability, we visualized the feature importance maps on representative images from CIFAR-10 and FMNIST. In CIFAR-10, pruned models consistently highlight semantically meaningful regions: for "cat" images, importance maps emphasize eyes, ears, and whiskers; for "ship" images, they focus on the hull and waterline. These visualizations confirm that the retained parameters correspond to human-intuitive and discriminative features rather than random noise. Moreover, quantitative explainability metrics reinforce that the top-ranked features in pruned models indeed drive predictions: SPATL-XL registers a Fidelity10 percent of 70.00 percent and a ConfidenceDrop10 percent of 6.39 percent, demonstrating that masking its top 10 percent features flips labels in 70 percent of test samples and reduces confidence substantially. In comparison, FedAvg's Fidelity10 percent is only 30.00 percent, indicating far less faithful attributions. Thus, the explainability module not only prunes effectively but also preserves and even enhances the model's capacity to generate reliable and interpretable explanations.

Cross-dataset tests on FMNIST and CIFAR-10 further confirm that the same explainability component—designed without assumptions about the network architecture—preserves the explainability and performance across very different input domains. On FMNIST, SPATL-XL attained 73.29 percent accuracy and produced clear feature attributions (highlighting pixels corresponding to clothing edges), whereas on CIFAR-10, it achieved 62.85

percent accuracy and still highlighted analogous semantically correct regions. This consistency across disparate domains confirms that a model-agnostic explainability module can be plugged into various federated settings to improve transparency and robustness without requiring customization for specific architectures.

RQ2: In what ways do explainability-driven pruning and dynamic client clustering interact to affect model accuracy, convergence dynamics, and communication overhead in federated learning?

The combination of explainability-driven pruning with per-round client clustering (SPATL-XLC) produces a synergistic effect, enhancing convergence stability and improving resource efficiency. In this pipeline, each client first prunes its model based on local importance scores; then, clients are grouped into clusters every round according to the similarity of their pruned model profiles. This two-stage approach generates clusters whose members share similar feature importance distributions, ensuring that aggregated updates within each cluster reinforce consistent patterns rather than amplifying local noise.

Empirically, SPATL-XLC demonstrates smoother convergence trajectories under non-IID conditions than pruning alone. Accuracy-versus-round curves show that while SPATL-XL's accuracy oscillates more prominently across rounds ($\pm$2–3 percent fluctuations), SPATL-XLC's trajectory remains within a narrower band ($\pm$1–2 percent), indicating stabilized learning. Although SPATL-XLC's final accuracy of 58.34 percent is slightly below SPATL-XL's 62.85 percent, its steadier convergence is critical for real-world federated settings, where stability often outweighs a minor accuracy edge.

On the communication front, cluster-level aggregation in SPATL-XLC reduces the number of messages traversing a centralized server. Each cluster aggregates client updates internally before sending a single "cluster update" to the server. As reported in Table 4.2, SPATL-XLC requires only 3.06 MB per round between the coordinator and server, whereas SPATL-XL still transmits 10.20 MB per client. This resulted in a 1.13$\times$ speed-up in the overall communication time relative to FedAvg. Although clustering adds a small overhead to compute cluster assignments each round, this cost is amortized by the reduced network traffic and lower server load. In large-scale federations, this reduction is crucial for maintaining system responsiveness and preventing network congestion.

The interaction between pruning and clustering proves critical for non-IID resilience: by first pruning low-importance parameters, each client's update emphasizes globally relevant features, and clustering then groups clients whose pruned models share these features. In effect, pruning filters noise and clustering aligns similar signals. This dual mechanism yields a global model that learns faster and generalizes better under data skew: whereas SPATL-XL alone improves non-IID accuracy from the baseline levels (e.g., FedAvg's mean accuracy on skewed splits), SPATL-XLC further narrows the gap by stabilizing updates and preventing divergence. The clustering quality metrics, namely Silhouette, Calinski–Harabasz (CHI), and Davies–Bouldin (DBI), remained within desirable ranges throughout the training, confirming that the dynamically recomputed clusters maintained cohesion and separation. In contrast, naive clustering of unpruned updates often yields either overly tight clusters that ignore performance or loose groupings that degrade the accuracy.

RQ3: How do our proposed modular pipelines, combining explainability, guided pruning with dynamic clustering—compare to state-of-the-art methods (FedAvg, FedProx, and original SPATL) in terms of performance and explainability?

A side-by-side evaluation under non-IID settings highlights the strengths of SPATL-XL and SPATL-XLC. FedAvg and FedProx, which transmit full gradients (approximately 22.80 MB per client per round), achieve final accuracies of 68.66 percent and 73.61 percent, respectively (Table 4.1), but offer no built-in explainability and suffer from unstable convergence when data are heavily skewed. The original SPATL (GNNRL), which employs reinforcement-learning–based pruning, achieves only 53.33 percent accuracy (Table 4.5) and incurs a prohibitive per-round runtime of approximately 818.76 s, rendering it impractical for real-world deployments.

In contrast, SPATL-XL attains 62.84 percent accuracy under non-IID conditions while reducing per-client communication to just 1.02 MB per round and maintaining the runtime at approximately 201.45 s. This represents a roughly 95 percent reduction in bandwidth compared with FedAvg and FedProx, with only a modest accuracy sacrifice. Crucially, SPATL-XL's retained parameters remain interpretable: feature-importance maps demonstrate that its pruned model consistently highlights meaningful input regions, and quantitative explainability metrics (fidelity, confidence drop, ECE, Brier score) show that SPATL-XL

produces more faithful and better-calibrated explanations than any baseline.

SPATL-XLC further reduces server-side communication to 3.06 MB per round and yields a communication time speed-up of 1.13× relative to FedAvg (Table 4.2). Its runtime per round (approximately 153.72 s) is lower than that of SPATL-XL, owing to cluster aggregation alleviating server bottlenecks. Although SPATL-XLC's final accuracy of 58.34 percent is slightly below that of SPATL-XL, its convergence stability and efficiency gains make it a more practical choice for large-scale, resource-constrained federated systems. Furthermore, explainability remains strong: cluster-averaged SHAP maps maintain semantic coherence across grouped clients, whereas the baselines offer no mechanism for visualizing or quantifying the feature importance.

Cross-dataset comparisons reinforce these conclusions. On FMNIST, SPATL-XLC achieves 75.25 percent accuracy, surpassing typical FedAvg baselines on that dataset, while providing clear pixel-level explanations of model decisions. On CIFAR-10, SPATL-XLC still outperformed the original SPATL by approximately 5 percent. Baseline methods, such as FedAvg and FedProx, show larger performance variations across datasets and offer zero transparency into model behavior, underscoring that our pipelines deliver consistent and interpretable solutions for diverse applications.

In summary, SPATL-XL and SPATL-XLC provide a balanced, state-of-the-art approach to federated learning under non-IID conditions: they achieve competitive accuracy with orders-of-magnitude reductions in communication and runtime and embed genuine, model-agnostic explainability via feature-importance visualizations. Compared to FedAvg and FedProx, whose strengths lie solely in raw accuracy, and the original SPATL, whose runoff learning overhead is untenable, our pipelines combine performance, transparency, and efficiency. This makes them well-suited for real-world federated deployments in domains where data heterogeneity, resource constraints, and the need for transparency converge, such as medical imaging, financial risk modeling, and edge-device AI.

Chapter Five

5. Conclusion and Recommendation

5.1 Conclusion

Non-IID data, limited bandwidth, and the growing demand for audit-ready explanations remain the three pillars that often thwart the real-world deployment of federated learning. This study tackled all three simultaneously by introducing two explainability-aware and communication-efficient extensions of FedAvg: SPATL-XL, which couples SHAP-guided parameter pruning with salient layer selection, and SPATL-XLC, which further embeds a lightweight, dynamic client-clustering routine.

Across CIFAR-10 and FMNIST, under both IID and severe label-skew settings, our methods delivered a favourable balance among accuracy, explanation quality, and system cost. In particular, SPATL-XL (1) lifted test accuracy by $\approx 1.7\%$ over vanilla FedAvg, (2) cut per-round traffic by 6.1%, and (3) reduced the deletion-AUC to 0.2666, signalling sharply faithful SHAP maps. SPATL-XLC went a step further: although its final accuracy was marginally lower (-1.5% versus FedAvg), it achieved an **89%** explanation fidelity, the highest stability (79.29%), and a **1.13**$\times$ wall-clock speed-up, while keeping client-server payload at a practical 13.26 MB per round trip. These results demonstrate that carefully orchestrated pruning and adaptive clustering can offset the detrimental effect of heterogeneity without sacrificing transparency or incurring prohibitive costs to do so.

From a deployment perspective, the proposed coordinator architecture is agnostic to model family and can be layered atop existing privacy mechanisms, such as secure aggregation, are required. Hence, the modular pipelines are immediately applicable to bandwidth-constrained, regulation-sensitive domains—e.g. remote health monitoring or cross-bank fraud detection—where both explainability and Data locality is a non-negotiable issue.

This study had two limitations. First, the current clustering relies on similarity in SHAP importance; a A more principled task-agnostic metric could yield even tighter groupings.

Second, the pruning policy was tuned for ResNet-20 and may need to be re-calibrated for transformer-based and tabular models.

5.2 Recommendations

Future research can extend SPATL-XL/XLC along four complementary axes. First, algorithmic refinements are required. Task-agnostic clustering criteria based on information-theoretic or representation-learning distances can generalize our approach beyond classification. Reinforcement learning or Bayesian schemes might adapt layer-wise prune ratios on-the-fly, and gradient-free attributions (e.g., KernelSHAP) would open the door to encrypted or non-differentiable models. Second, system-level enhancements should pair pruning and clustering with differential-privacy budget accounting, bandwidth-aware client scheduling, and hardware co-design for block-sparse tensor cores, promising an additional 1.3–1.5× speed-up on edge NPUs. Third, robustness and trust merit deeper study: adversarial heterogeneity tests can reveal whether attackers can spoof SHAP patterns to hijack clusters, whereas calibration-aware explanations and per-round audit trails would ease compliance with the forthcoming EU AI Act. Finally, benchmarking and reproducibility call for cross-domain evaluations—speech, electronic health records, and graph benchmarks—as well as an open-source, Dockerized pipeline with Weights&Biases sweeps to foster fair comparison and industrial uptake. Addressing these directions will tighten the accuracy–explainability–efficiency trade-off demonstrated in this study and move federated learning closer to a plug-and-play, regulation-ready paradigm across diverse edge ecosystems.

5. REFERENCES

[1] N. Rieke, J. Hancox, W. Li, F. Milletarì, H. Roth, S. Albarqouni, S. Bakas, M. Galtier, B. Landman, K. Maier-Hein, S. Ourselin, M. Sheller, R. Summers, A. Trask, D. Xu, M. Baust, and M. Cardoso, "The future of digital health with federated learning," NPJ Digital Medicine, vol. 3, 2020.

[2] T. Nguyen, M. A. Dakka, S. M. Diakiw, M. VerMilyea, M. Perugini, J. M. M. Hall, and D. Perugini, "A novel decentralized federated learning approach to train on globally distributed, poor quality, and protected private medical data," Scientific Reports, 2022.

[3] S. Zhang and G. Liu, "Personalized federated learning with attention mechanisms," Academic Journal of Computing & Information Science, 2023.

[4] M. Luo, F. Chen, D. Hu, Y. Zhang, J. Liang, and J. Feng, "No fear of heterogeneity: classifier calibration for federated learning with non-iid data," arXiv preprint, 2021.

[5] H. Zhang, P. Zhang, M. Hu, M. Liu, and J. Wang, "Fedub: federated learning algorithm based on update bias," Mathematics, vol. 12, p. 1601, 2024.

[6] I. Dayan and R. et al, "Federated learning for predicting clinical outcomes in patients with covid-19," Nature Medicine, 2021.

[7] K. V. Sarma, S. A. Harmon, T. Sanford, H. R. Roth, Z. Xu, J. Tetreault, D. Xu, M. G. Flores, A. Raman, R. Kulkarni, B. J. Wood, P. L. Choyke, A. Priester, L. S. Marks, S. S. Raman, D. R. Enzmann, B. Türkbey, W. Speier, and C. Arnold, "Federated learning improves site performance in multicenter deep learning without data sharing," Journal of the American Medical Informatics Association, 2021.

[8] D. Singh, D. P. Singh, and M. Chandra, "Enhancing transparency and interpretability in deep learning models: A comprehensive study on explainable ai techniques," Interantional Journal of Scientific Research in Engineering and Management, 2024.

[9] A. A. Adeniran, A. P. Onebunne, and P. William, "Explainable ai (xai) in healthcare: Enhancing trust and transparency in critical decision-making," World Journal of Advanced Research and Reviews, vol. 23, pp. 2447–2658, 2024.

[10] G. Arjunan, "Implementing explainable ai in healthcare: Techniques for interpretable machine learning models in clinical decision-making," International Journal of Scientific Research and Management, vol. 9, pp. 597–603, 2021.

[11] T. Tzimas, "Algorithmic transparency and explainability under eu law," European Public Law, vol. 29, pp. 385–411, 2023.

[12] G. Anushree, S. B. Madagaonkar, and C. H. Ravili, "Unveiling the black box: a comprehensive review of explainable ai techniques," Interantional Journal of Scientific Research in Engineering and Management, vol. 08, pp. 1–6, 2024.

[13] C. Shiranthika, P. Saeedi, and I. V. Bajić, "Decentralized learning in healthcare: a review of emerging techniques," IEEE Access, vol. 11, pp. 54188–54209, 2023.

[14] M. Moshawrab, M. Adda, A. Bouzouane, H. Ibrahim, and A. Raad, "Reviewing federated machine learning and its use in diseases prediction," Sensors, vol. 23, p. 2112, 2023.

[15] S. Rajendran, Z. Xu, W. Pan, A. K. Ghosh, and F. Wang, "Data heterogeneity in federated learning with electronic health records: Case studies of risk prediction for acute kidney injury and sepsis diseases in critical care," Plos Digital Health, 2023.

[16] T. Liu, J. Ding, T. Wang, M. Pan, and M. Chen, "Towards fast and accurate federated learning with non-iid data for cloud-based iot applications," Journal of Circuits Systems and Computers, vol. 31, 2022.

[17] A. Chen, Y. Fu, L. Wang, and G. Duan, "Dwfed: A statistical- heterogeneity-based dynamic weighted model aggregation algorithm for federated learning," Frontiers in Neurorobotics, 2022.

[18] N. Mohammadi, J. Bai, Q. Fan, Y. Song, Y. Yi, and L. Liu, "Differential privacy meets federated learning under communication constraints," Ieee Internet of Things Journal, vol. 9, pp. 22204–22219, 2022.

[19] K. Kumar and K. Jyoti, "Enhancing transparency and trust in brain tumor diagnosis: an in-depth analysis of deep learning and explainable ai techniques," 2025.

[20] S. Yu, P. Nguyen, W. Abebe, W. Qian, A. Anwar, and A. Jannesari, "Spatl: Salient parameter aggregation and transfer learning for heterogeneous clients in federated learning," 2021.

[21] D. Upreti, E. Yang, H. Kim, and C. Seo, "A comprehensive survey on federated learning in the healthcare area: concept and applications," Computer Modeling in Engineering & Sciences, vol. 140, pp. 2239–2274, 2024.

[22] A. Rahman, M. S. Hossain, G. Muhammad, D. Kundu, T. Debnath, M. Rahman, M. S. Islam Khan, P. Tiwari, and S. S. Band, "Federated learning-based ai approaches in smart healthcare: Concepts, taxonomies, challenges and open issues," Cluster Computing, 2022.

[23] P. H. Zeraik Viduedo, V. C. Urquiza Candeia, V. B. Legname Martins, A. C. Destefani, and V. C. Destefani, "Harnessing the power of ai and machine learning for next-generation sequencing data analysis: A comprehensive review of applications, challenges, and future directions in precision oncology," Revista Ibero-Americana De Humanidades Ciências E Educação, 2024.

[24] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," The Lancet Digital Health, vol. 3, pp. e745–e750, 2021.

[25] M. Imani, "Customer churn prediction: a review of recent advances, trends, and challenges in conventional machine learning and deep learning," 2025.

[26] D. Kosaraju, "Securing ai: Federated learning as a tool for privacy preservation," Galore International Journal of Applied Sciences and Humanities, 2024.

[27] V. Tulsani, P. Sahatiya, J. Parmar, and J. Parmar, "Xai applications in medical imaging: A survey of methods and challenges," International Journal on Recent and Innovation Trends in Computing and Communication, 2023.

[28] X. Song, W. Hsieh, Z. Bi, C. Jiang, J. Liu, B. Peng, S. Zhang, X. Pan, J. Xu, J. Wang, K. Chen, C. Yin, P. Feng, Y. Wen, T. Wang, J. Yang, M. Li, B. Jing, J. Ren, J. Song, H. Xu, H. Tseng, Y. Zhang, L. Yan, Q. Niu, S. Chen, Y. Wang, C. Liang, and M. Liu, "A comprehensive guide to explainable ai: from classical models to llms," OSF, 2024.

[29] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (A. Singh and J. Zhu, eds.), vol. 54 of Proceedings of Machine Learning Research, pp. 1273–1282, PMLR, 20–22 Apr 2017.

[30] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in Proceedings of Machine Learning and Systems, vol. 2, pp. 429–450, 2020.

[31] S. P. Karimireddy, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich, and A. T. Suresh, "Scaffold: stochastic controlled averaging for federated learning," in Proceedings of the 37th International Conference on Machine Learning, ICML'20, JMLR.org, 2020.

[32] S. Yu, P. Nguyen, W. Abebe, W. Qian, A. Anwar, and A. Jannesari, "Spatl: Salient parameter aggregation and transfer learning for heterogeneous clients in federated learning," arXiv preprint arXiv:2111.14345, 2021.

[33] Y. Ramon, D. Martens, F. Provost, and T. Evgeniou, "A comparison of instance-level counterfactual explanation algorithms for behavioral and textual data: Sedc, lime-c and shap-c," Advances in Data Analysis and Classification, 2020.

[34] D. M. Anand Korade, "Unlocking machine learning model decisions: A comparative analysis of lime and shap for enhanced interpretability," Jes, 2024.

[35] P. Kairouz, "Advances and open problems in federated learning," Foundations and Trends® in Machine Learning, vol. 14, pp. 1–210, 2021.

[36] M. Aledhari, R. Razzak, R. M. Parizi, and F. Saeed, "Federated learning: A survey on enabling technologies, protocols, and applications," Ieee Access, vol. 8, pp. 140699–140725, 2020.

[37] J. Jiang, S. Ji, and G. Long, "Decentralized knowledge acquisition for mobile internet applications," World Wide Web, vol. 23, pp. 2653–2669, 2020.

[38] T. Berghout, T. Bentrcia, M. Ferrag, and M. Benbouzid, "A heterogeneous federated transfer learning approach with extreme aggregation and speed," Mathematics, vol. 10, p. 3528, 2022.

[39] A. Abdelmoniem, C. Ho, P. Papageorgiou, and M. Canini, "A comprehensive empirical study of heterogeneity in federated learning," Ieee Internet of Things Journal, vol. 10, pp. 14071–14083, 2023.

[40] E. Soykan, L. Karaçay, F. Karakoç, and E. Tomur, "A survey and guideline on privacy enhancing technologies for collaborative machine learning," Ieee Access, vol. 10, pp. 97495–97519, 2022.

[41] Z. Lu, H. Pan, Y. Dai, X. Si, and Y. Zhang, "Federated learning with non-iid data: A survey," IEEE Internet of Things Journal, vol. 11, no. 11, pp. 19188–19209, 2024.

[42] S. Lo, Q. Lu, H. Paik, and L. Zhu, "Flra: a reference architecture for federated learning systems," 2021.

[43] B. Li, "Design of asynchronous federated learning incentive mechanism," Academic Journal of Computing & Information Science, 2023.

[44] S. Shen, "From distributed machine learning to federated learning: in the view of data privacy and security," 2020.

[45] K. Zhai, Q. Ren, J. Wang, and C. Yan, "Byzantine-robust federated learning via credibility assessment on non-iid data," 2021.

[46] J. Mills, J. Hu, and G. Min, "Multi-task federated learning for personalised deep neural networks in edge computing," Ieee Transactions on Parallel and Distributed Systems, vol. 33, pp. 630–641, 2022.

[47] S. Gupta, "Federated learning: advancing privacy-preserving machine learning at scale," International Journal of Scientific Research in Computer Science Engineering and Information Technology, vol. 11, pp. 2993–3008, 2025.

[48] R. Kang, Q. Li, and H. Lu, "Federated machine learning in finance: A systematic review on technical architecture and financial applications," Applied and Computational Engineering, vol. 102, pp. 61–72, 2024.

[49] E. Ellahi, "Fraud detection and prevention in finance: leveraging artificial intelligence and big data.," dxjb, vol. 36, pp. 54–62, 2024.

[50] J. Mason, B. Peoples, and J. Lee, "Questioning the scope of AI standardization in learning, education, and training," Journal of ICT Standardization, pp. 107–122, 2020.

[51] N. Said, "Federated learning for privacy-preserving AI: Challenges, applications, and future directions," Policy and Management Journal, vol. 35, no. 3S, pp. 358–368, 2025.

[52] N. Liu and Y. Jin, "Mobile health-empowered traditional ethnic sports: Ai-based data analysis improving security," Internet Technology Letters, 2023.

[53] W. Guo, "Overview of research progress and challenges in federated learning," Tcsisr, 2024.

[54] Y. Li, C. Chen, N. Liu, H. Huang, Z. Zheng, and Y. Qiang, "A blockchain-based decentralized federated learning framework with committee consensus," Ieee Network, 2021.

[55] L. N. Valli, N. Sujatha, M. Mech, and V. S. Lokesh, "Ethical considerations in data science: Balancing privacy and utility," International Journal of Science and Research Archive, 2024.

[56] L. Zhao and J. Huang, "A distribution information sharing federated learning approach for medical image data," Complex & Intelligent Systems, 2023.

[57] H. Fan, C. Huang, and Y. Liu, "Federated learning-based privacy-preserving data aggregation scheme for iiot," Ieee Access, 2023.

[58] X. Yang and C. Xing, "Federated medical learning framework based on blockchain and homomorphic encryption," Wireless Communications and Mobile Computing, 2024.

[59] R. Kumar, A. A. Khan, J. Kumar, Z. Zakria, N. A. Golilarz, S. Zhang, T. Yang, C. Zheng, and W. Wang, "Blockchain-federated-learning and deep learning models for covid-19 detection using ct imaging," Ieee Sensors Journal, 2021.

[60] F. Sattler, S. Wiedemann, K. Müller, and W. Samek, "Robust and communication-efficient federated learning from non-i.i.d. data," Ieee Transactions on Neural Networks and Learning Systems, 2020.

[61] S. Hong and J. Chae, "Communication-efficient randomized algorithm for multi-kernel online federated learning," Ieee Transactions on Pattern Analysis and Machine Intelligence, 2022.

[62] S. Oh, J. Park, E. Jeong, H. Kim, M. Bennis, and S. Kim, "Mix2fld: Downlink federated learning after uplink federated distillation with two-way mixup," Ieee Communications Letters, 2020.

[63] N. D. Fabbro, A. Mitra, and G. J. Pappas, "Federated td learning over finite-rate erasure channels: Linear speedup under markovian sampling," Ieee Control Systems Letters, 2023.

[64] L. Malihi and G. Heidemann, "Efficient and controllable model compression through sequential knowledge distillation and pruning," Big Data and Cognitive Computing, 2023.

[65] Z. Zhu, Y. Shi, G. Xin, C. Peng, P. Fan, and K. B. Letaief, "Towards efficient federated learning: Layer-wise pruning-quantization scheme and coding design," Entropy, 2023.

[66] Y. Li, W. Li, and Z. Xue, "Federated learning with stochastic quantization," International Journal of Intelligent Systems, 2022.

[67] Z. Zhen, Z. Wu, L. Feng, W. Li, F. Qi, and S. Guo, "A secure and effective energy-aware fixed-point quantization scheme for asynchronous federated learning," Computers Materials & Continua, 2023.

[68] X. Li and H. Wu, "Spatio-temporal representation with deep neural recurrent network in mimo csi feedback," Ieee Wireless Communications Letters, 2020.

[69] H. Gao, A. Xu, and H. Huang, "On the convergence of communication-efficient local sgd for federated learning," Proceedings of the Aaai Conference on Artificial Intelligence, 2021.

[70] L. Chen, L. Fan, X. Lei, T. Q. Duong, A. Nallanathan, and G. K. Karagiannidis, "Relay-assisted federated edge learning: Performance analysis and system optimization," Ieee Transactions on Communications, 2023.

[71] X. LI, Z. Fang, and Z. SHI, "Mpfedcon : Model-contrastive personalized federated learning with the class center," Wuhan University Journal of Natural Sciences, 2022.

[72] Y.-H. Lai, S.-Y. Chen, W. Chou, H.-Y. Hsu, and H. Chao, "Personalized federated learning with adaptive feature extraction and category prediction in non-iid datasets," Future Internet, 2024.

[73] Y. Huang, L. Chu, Z. Zhou, L. Wang, J. Liu, J. Pei, and Y. Zhang, "Personalized cross-silo federated learning on non-iid data," Proceedings of the Aaai Conference on Artificial Intelligence, 2021.

[74] B. K. Alotaibi, F. A. Khan, and S. Mahmood, "Communication efficiency and non-independent and identically distributed data challenge in federated learning: A systematic mapping study," Applied Sciences, vol. 14, p. 2720, 2024.

[75] B. Casella, R. Esposito, A. Sciarappa, C. Cavazzoni, and M. Aldinucci, "Experimenting with normalization layers in federated learning on non-iid scenarios," Ieee Access, 2024.

[76] B. Chai, K. Liu, and R. Yang, "Cross-domain federated data modeling on non-iid data," Computational Intelligence and Neuroscience, 2022.

[77] X. Meng, Y. Li, J. Lu, and X. Ren, "An optimization method for non-iid federated learning based on deep reinforcement learning," Sensors, 2023.

[78] Y. Li, X. Li, G. Li, and Z. Li, "Privacy protection in prosumer energy management based on federated learning," Ieee Access, 2021.

[79] W. Yao, B. Pan, Y. Hou, X. Li, and Y. Xia, "An adaptive model filtering algorithm based on grubbs test in federated learning," Entropy, 2023.

[80] H. Kim, B. Kim, Y. Kim, C. You, and H. Park, "K-fl: Kalman filter-based clustering federated learning method," Ieee Access, 2023.

[81] H. Jamali-Rad, M. Abdizadeh, and A. Singh, "Federated learning with taskonomy for non-iid data," Ieee Transactions on Neural Networks and Learning Systems, 2023.

[82] C. Xu, H. Liu, K. Li, W. Feng, and Q. Wei, "Pretraining client selection algorithm based on a data distribution evaluation model in federated learning," Ieee Access, 2024.

[83] L. Yu and J. Huang, "Cyclic federated learning method based on distribution information sharing and knowledge distillation for medical data," Electronics, 2022.

[84] S. Li, E. C. Ngai, and T. Voigt, "An experimental study of byzantine-robust aggregation schemes in federated learning," Ieee Transactions on Big Data, 2024.

[85] J. Wang, K. Yang, and M. Li, "Nids-fgpa: A federated learning network intrusion detection algorithm based on secure aggregation of gradient similarity models," Plos One, 2024.

[86] M. Jiang, Z.-R. Wang, and Q. Dou, "Harmofl: Harmonizing local and global drifts in federated learning on heterogeneous medical images," Proceedings of the Aaai Conference on Artificial Intelligence, 2022.

[87] H. Li, Y. Yin, Z. Fu, S. Zhang, H. Deng, and D. Liu, "Loadaboost: Loss-based adaboost federated machine learning with reduced computational complexity on iid and non-iid intensive care data," Plos One, 2020.

[88] W. Si and C. Liu, "Privacy preservation learning with deep cooperative method for multimedia data analysis," Security and Communication Networks, 2022.

[89] P. Guo, Y. Yang, W. Guo, and Y. Shen, "A fair contribution measurement method for federated learning," Sensors, 2024.

[90] H. Hoech, R. Rischke, K. Müller, and W. Samek, "Fedauxfdp: Differentially private one-shot federated distillation," 2022.

[91] P. P. Liang, T. Liu, Z. Liu, N. B. Allen, R. P. Auerbach, D. A. Brent, R. Salakhutdinov, and L. Morency, "Think locally, act globally: Federated learning with local and global representations," 2020.

[92] X. Shang, Y. Lu, Y. Cheung, and H. Wang, "Fedic: Federated learning on non-iid and long-tailed data via calibrated distillation," 2022.

[93] T. Yoon, S. Shin, S. J. Hwang, and E. Yang, "Fedmix: Approximation of mixup under mean augmented federated learning," 2021.

[94] T.-M. H. Hsu, H. Qi, and M. A. Brown, "Federated visual classification with real-world data distribution," 2020.

[95] J. Beilharz, B. Pfitzner, R. Schmid, P. Geppert, B. Arnrich, and A. Polze, "Implicit model specialization through dag-based decentralized federated learning," 2021.

[96] Y. Zhou, G. Duan, T. Qiu, Z. Li, T. Li, X. Zheng, and Y. Zhu, "Personalized federated learning incorporating adaptive model pruning at the edge," Electronics, 2024.

[97] K. Jahani, B. Moshiri, and B. H. Khalaj, "A survey on data distribution challenges and solutions in vertical and horizontal federated learning," Jaiai, 2024.

[98] X. Ye, D. Klabjan, and Y. Luo, "Aggregation delayed federated learning," 2021.

[99] M. S. Islam, S. Javaherian, F. Xu, X. Yuan, L. Chen, and N.-F. Tzeng, "Fedclust: Tackling data heterogeneity in federated learning through weight-driven client clustering," 2024.

[100] A. B. Haque, A. N. Islam, and P. Mikalef, "Explainable artificial intelligence (xai) from a user perspective: A synthesis of prior literature and problematizing avenues for future research," Technological Forecasting and Social Change, 2023.

[101] J.-M. John-Mathews, "Some critical and ethical perspectives on the empirical turn of ai interpretability," Technological Forecasting and Social Change, 2022.

[102] M. A. Akbar, A. A. Khan, S. Mahmood, S. Rafi, and S. Demi, "Trustworthy artificial intelligence: A <scp>decision-making</scp> taxonomy of potential challenges," Software Practice and Experience, 2023.

[103] N. Quadrianto, B. W. Schuller, and F. Lattimore, "Editorial: Ethical machine learning and artificial intelligence," Frontiers in Big Data, 2021.

[104] R. Papadimitriou, "The right to explanation in the processing of personal data with the use of ai systems," International Journal of Law in Changing World, 2023.

[105] A. Tsamados, N. Aggarwal, J. Cowls, J. Morley, H. Roberts, M. Taddeo, and L. Floridi, "The ethics of algorithms: Key problems and solutions," Ai & Society, 2021.

[106] S. Akgün and C. Greenhow, "Artificial intelligence in education: Addressing ethical challenges in k-12 settings," Ai and Ethics, 2021.

[107] Y. Yang, "Influences of digital literacy and moral sensitivity on artificial intelligence ethics awareness among nursing students," Healthcare, 2024.

[108] P. V. Ambatkar Deshpande, "Interpretable deep learning models: Enhancing transparency and trustworthiness in explainable ai," cienc.eng., 2023.

[109] B. H. Misheva, D. Jaggi, J.-A. Posth, T. Gramespacher, and J. Osterrieder, "Audience-dependent explanations for ai-based risk management tools: A survey," Frontiers in Artificial Intelligence, 2021.

[110] A. Sathyan, A. I. Weinberg, and K. Cohen, "Interpretable ai for bio-medical applications," Complex Engineering Systems, 2022.

[111] B. Aldughayfiq, F. Ashfaq, N. Z. Jhanjhi, and M. Humayun, "Explainable ai for retinoblastoma diagnosis: Interpreting deep learning models with lime and shap," Diagnostics, 2023.

[112] C. Korgialas, E. Pantraki, A. Bolari, M. Sotiroudi, and C. Kotropoulos, "Face aging by explainable conditional adversarial autoencoders," Journal of Imaging, 2023.

[113] I. Hussain, R. Jany, R. Boyer, A. Azad, S. A. Alyami, S. J. Park, M. M. Hasan, and M. A. Hossain, "An explainable eeg-based human activity recognition model using machine-learning approach and lime," Sensors, 2023.

[114] D. P. Panagoulias, E. Sarmas, V. Marinakis, M. Virvou, G. A. Tsihrintzis, and H. Doukas, "Intelligent decision support for energy management: A methodology for tailored explainability of artificial intelligence analytics," Electronics, 2023.

[115] M. Graziani, L. Dutkiewicz, D. Calvaresi, J. P. Amorim, K. Yordanova, M. Vered, R. Nair, P. H. Abreu, T. Blanke, V. Pulignano, J. O. Prior, L. Lauwaert, W. Reijers, A. Depeursinge, V. Andrearczyk, and H. Müller, "A global taxonomy of interpretable ai: Unifying the terminology for the technical and social sciences," Artificial Intelligence Review, 2022.

[116] Y. T. Tang and R. Romero-Ortuño, "Using explainable ai (xai) for the prediction of falls in the older population," Algorithms, 2022.

[117] F. Auletta, R. W. Kallen, M. d. Bernardo, and M. J. Richardson, "Predicting and understanding human action decisions during skillful joint-action using supervised machine learning and explainable-ai," Scientific Reports, 2023.

[118] R. Tiwari, "Explainable ai (xai) and its applications in building trust and understanding in ai decision making," Interantional Journal of Scientific Research in Engineering and Management, 2023.

[119] K. S. Mohammed and A. Shehu, "A review of artificial intelligence (ai) challenges and future prospects of explainable ai in major fields: A case study of nigeria," Open Journal of Physical Science (Issn 2734-2123), 2023.

[120] R. K. Kundu and K. A. Hoque, "Explainable predictive maintenance is not enough: Quantifying trust in remaining useful life estimation," Annual Conference of the PHM Society, 2023.

[121] R. Silva, A. Sbrana, P. d. Castro, and N. Y. Soma, "Developing and assessing a human-understandable metric for evaluating local interpretable model-agnostic explanations," International Journal of Intelligent Engineering and Systems, 2023.

[122] K. Zahoor, N. Z. Bawany, and T. Qamar, "Evaluating text classification with explainable artificial intelligence," Iaes International Journal of Artificial Intelligence (Ij-Ai), vol. 13, p. 278, 2024.

[123] V. Bidve, M. M. Pathan, P. Sarasu, A. Pavate, A. Shaikh, S. Borde, V. B. Pratap Singh, and R. M. Raut, "Use of explainable ai to interpret the results of nlp models for sentimental analysis," Indonesian Journal of Electrical Engineering and Computer Science, 2024.

[124] L. Cavique, "Implications of causality in artificial intelligence," Frontiers in Artificial Intelligence, 2024.

[125] W. Abdullah, A. S. Tolba, A. Elmasry, and N. N. Mostafa, "Visioncam: A comprehensive xai toolkit for interpreting image-based deep learning models," Sustain. Mach. Intell. J., 2024.

[126] T. Hulsen, "Explainable artificial intelligence (xai): Concepts and challenges in healthcare," Ai, 2023.

[127] H. Taherdoost, "Blockchain and machine learning: A critical review on security," Information, 2023.

[128] O. O. Olateju, S. U. Okon, U. T. Ikechukwu Igwenagu, A. A. Salami, T. O. Oladoyinbo, and O. O. Olaniyi, "Combating the challenges of false positives in ai-driven anomaly detection systems and enhancing data security in the cloud," Asian Journal of Research in Computer Science, 2024.

[129] X. Yin, X. Wu, and X. Zhang, "A trusted federated learning method based on consortium blockchain," Information, 2024.

[130] M. Muslim, "Managerial finance tactics in the era of enhanced regulation following financial scandals," Advances in Management & Financial Reporting, 2024.

[131] G. Kaissis, M. R. Makowski, D. Rückert, and R. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," Nature Machine Intelligence, 2020.

[132] L. Jiang, E. L. Bettac, H. J. Lee, and T. M. Probst, "In whom do we trust? a multi-foci person-centered perspective on institutional trust during covid-19," International Journal of Environmental Research and Public Health, 2022.

[133] W. Saad, M. Bitar, A. Ksentini, M. Bader, and J. A. Filho, "A trust and explainable federated deep learning framework in zero touch b5g networks," in 2022 IEEE Global Communications Conference (GLOBECOM), pp. 1289–1294, IEEE, 2022.

[134] P. Fuladi, S. Kumar, X. Zhang, and Y. Zhou, "An optimized fl-xai model for secured and trustworthy candidate selection." arXiv preprint, 2024. Available at: https://doi.org/10.21203/rs.3.rs-4475624/v1.

[135] T. Yang, L. Chen, H. Li, and J. Wang, "An explainable federated learning and blockchain-based secure credit modeling method," European Journal of Operational Research, vol. 298, pp. 1156–1170, 2024.

[136] D. Bogdanova, D. Petrescu, M. Alam, and I. Ivanov, "Dc-shap method for consistent explainability in privacy-preserving distributed machine learning," Human-Centric Intelligent Systems, vol. 3, no. 1, pp. 1–15, 2023.

[137] M. Renda, O. Khalaf, P. Sullivan, and K. Fountoulakis, "Federated learning of explainable ai models in 6g systems: Towards secure and automated vehicle networking," Information, vol. 13, no. 8, p. 395, 2022.

[138] A. Alferaidi, K. Yadav, Y. Alharbi, W. Viriyasitavat, S. Kautish, and G. Dhiman, "Federated learning algorithms to optimize the client and cost selections," Mathematical Problems in Engineering, 2022.

[139] K. R. Reddy and A. Muralidhar, "Machine learning-based road safety prediction strategies for internet of vehicles (iov) enabled vehicles: A systematic literature review," Ieee Access, 2023.

[140] E. Goh, D. Kim, K. Lee, S. Oh, J.-E. Chae, and D.-Y. Kim, "Blockchain-enabled federated learning: A reference architecture design, implementation, and verification," Ieee Access, 2023.

[141] C. Su, J. Wei, L. Yuan, H. Xuan, and J. Li, "Empowering precise advertising with fedgancc: A novel federated learning approach leveraging generative adversarial networks and group clustering," Plos One, 2024.

[142] J. Zhang, B. Li, C. Chen, L. Lyu, S. Wu, S. Ding, and C. Wu, "Delving into the adversarial robustness of federated learning," Proceedings of the Aaai Conference on Artificial Intelligence, 2023.

[143] R. Ghnemat, S. Alodibat, and Q. A. Al-Haija, "Explainable artificial intelligence (xai) for deep learning based medical imaging classification," Journal of Imaging, 2023.

[144] A. Salama, A. Stergioulis, A. M. Hayajneh, S. A. Raza Zaidi, D. McLernon, and I. Robertson, "Decentralized federated learning over slotted aloha wireless mesh networking," Ieee Access, 2023.

[145] S. Warnat-Herresthal, H. Schultze, K. L. Shastry, S. Manamohan, S. Mukherjee, and e. a. Garg, "Swarm learning for decentralized and confidential clinical machine learning," Nature, 2021.

[146] L. Chen, W. Zhang, C. Dong, D. Zhao, X. Zeng, S. Qiao, Y. Zhu, and C. W. Tan, "Fedtkd: A trustworthy heterogeneous federated learning based on adaptive knowledge distillation," Entropy, 2024.

[147] W. Wang, "Empowering safe and secure autonomy: Federated learning in the era of autonomous driving," Applied and Computational Engineering, 2024.

[148] K. Machová, M. Mach, and V. Balara, "Federated learning in the detection of fake news using deep learning as a basic method," Sensors, 2024.

[149] W. Zhang, T. Zhou, Q. Lu, X. Wang, C. Zhu, H. Sun, Z. Wang, S. K. Lo, and F. Wang, "Dynamic-fusion-based federated learning for covid-19 detection," Ieee Internet of Things Journal, 2021.

[150] W.-H. Shen, S. Wu, and Y.-H. Tao, "Cldp-pfedavg: Safeguarding client data privacy in personalized federated averaging," Mathematics, 2024.

[151] F. Yang, M. Z. Abedin, and P. Hájek, "An explainable federated learning and blockchain-based secure credit modeling method," European Journal of Operational Research, 2024.

[152] Y. Ding, "Adaptive lazily aggregated quantized algorithm based on federated self-supervision," Academic Journal of Computing & Information Science, 2024.

[153] T. Wu, M. Jiang, Y. Han, Z. Yuan, X. Li, and L. Zhang, "A traffic-aware federated imitation learning framework for motion control at unsignalized intersections with internet of vehicles," Electronics, 2021.

[154] A. S. Saud and S. Shakya, "Know sure thing based machine learning strategy for predicting stock trading signals," International Journal of Intelligent Engineering and Systems, 2022.

[155] L. Shen, "5g communication network and federated learning for college accounting informatization teaching," J. Comb. Math. Comb. Comput., 2024.

[156] Z. Gong, H. Zhang, H. Yang, F. Liu, and F. Luo, "A review of neural network lightweighting techniques," Innov. Technol. Adv., 2024.

[157] Y. Sun, "Optimizing federated learning efficiency: A comparative analysis of model compression techniques for communication reduction," Applied and Computational Engineering, 2024.

[158] H.-T. Lin and C.-Y. Wen, "Edge federated optimization for heterogeneous data," Future Internet, 2024.

[159] Y. Jiang, S. Wang, V. Valls, B. J. Ko, W. Lee, K. K. Leung, and L. Tassiulas, "Model pruning enables efficient federated learning on edge devices," Ieee Transactions on Neural Networks and Learning Systems, 2023.

[160] R. Lin, Y. Xiao, T.-J. Yang, D. Zhao, L. Xiong, G. Motta, and F. Beaufays, "Federated pruning: Improving neural network efficiency with federated learning," 2022.

[161] C. Wu, Y. Xian, S. Zhu, and P. Mitra, "Toward cleansing backdoored neural networks in federated learning," 2022.

[162] Z. Qin, G. Y. Li, and H. Ye, "Federated learning and wireless communications," 2020.

[163] K. Yang, T. Jiang, Y. Shi, and Z. Ding, "Federated learning via over-the-air computation," Ieee Transactions on Wireless Communications, 2020.

[164] F. Ang, L. Chen, N. Zhao, Y. Chen, W. Wang, and F. R. Yu, "Robust federated learning with noisy communication," Ieee Transactions on Communications, 2020.

[165] F. Sattler, A. Marbán, R. Rischke, and W. Samek, "Communication-efficient federated distillation," 2020.

[166] C. Wu, F. Wu, L. Lyu, Y. Huang, and X. Xie, "Communication-efficient federated learning via knowledge distillation," Nature Communications, 2022.

[167] x. li, Y. Chen, Y. Zhang, and X. Yang, "An adaptive quantization model compression method for heterogeneous federated learning," 2024.

[168] C. Wu, "Fedkd: Communication efficient federated learning via knowledge distillation," 2021.

[169] D. Jhunjhunwala, A. Gadhikar, G. Joshi, and Y. C. Eldar, "Adaptive quantization of model updates for communication-efficient federated learning," 2021.

[170] A. Imteaj and M. H. Amini, "Fedparl: Client activity and resource-oriented lightweight federated learning model for resource-constrained heterogeneous iot environment," Frontiers in Communications and Networks, 2021.

[171] L. Shkurti and M. Selimi, "Adaptivemesh: Adaptive federate learning for resource-constrained wireless environments," International Journal of Online and Biomedical Engineering (Ijoe), 2024.

[172] X. Wang, "An efficient federated learning optimization algorithm on non-iid data," 2022.

[173] J. Wang, G. Li, and W. Zhang, "Combine-net: An improved filter pruning algorithm," Information, 2021.

[174] R. Rodríguez-Pérez and J. Bajorath, "Interpretation of machine learning models using shapley values: Application to compound potency and multi-target activity predictions," Journal of Computer-Aided Molecular Design, 2020.

[175] A. V. Ponce Bobadilla, V. Schmitt, C. S. Maier, S. Mensing, and S. Stodtmann, "Practical guide to <scp>shap</scp> analysis: Explaining supervised machine learning model predictions in drug development," Clinical and Translational Science, 2024.

[176] A. Khan, J. Kandel, H. Tayara, and K. T. Chong, "Predicting the bandgap and efficiency of perovskite solar cells using machine learning methods," Molecular Informatics, 2024.

[177] S. Cao and Y. Hu, "Interpretable machine learning framework to predict gout associated with dietary fiber and triglyceride-glucose index," Nutrition & Metabolism, 2024.

[178] M. J. Ariza Garzón, J. Arroyo, A. Caparrini, and M. J. Segovia Vargas, "Explainability of a machine learning granting scoring model in peer-to-peer lending," Ieee Access, 2020.

[179] L. Prono, P. Bich, C. Boretti, M. Mangia, F. Pareschi, R. Rovatti, and G. Setti, "A multiply-and-max/min neuron paradigm for aggressively prunable deep neural networks," 2024.

[180] S. Hurtado, H. Nematzadeh, J. García-Nieto, M. Berciano-Guerrero, and I. Navas-Delgado, "On the use of explainable artificial intelligence for the differential diagnosis of pigmented skin lesions," 2022.

[181] T. Wu, J. Shi, D. Zhou, X. Zheng, and N. Li, "Evolutionary multi-objective one-shot filter pruning for designing lightweight convolutional neural network," Sensors, 2021.

[182] D. Li, S. Chen, X. Liu, Y. Sun, and L. Zhang, "Towards optimal filter pruning with balanced performance and pruning speed," 2021.

[183] Y. Yang, Y. Ji, Y. Wang, H. Qi, and J. Kato, "One-shot network pruning at initialization with discriminative image patches," 2022.

[184] Y. Wang, X. Zhang, L. Xie, J. Zhou, H. Su, B. Zhang, and X. Hu, "Pruning from scratch," Proceedings of the Aaai Conference on Artificial Intelligence, 2020.

[185] J. Ma, G. Long, T. Zhou, J. Jiang, and C. Zhang, "On the convergence of clustered federated learning," 2022.

[186] S. Li, S. Hou, B. Buyukates, and S. Avestimehr, "Secure federated clustering," 2022.

[187] M. Mehta and C. Shao, "A greedy agglomerative framework for clustered federated learning," Ieee Transactions on Industrial Informatics, 2023.

[188] R. Wang, N. Bian, and Y. Wu, "Fedrcd: A federated learning algorithm leveraging distribution extraction and community detection for enhanced clustering," 2023.

[189] Y. Ruan and C. Joe-Wong, "Fedsoft: Soft clustered federated learning with proximal local updating," Proceedings of the Aaai Conference on Artificial Intelligence, 2022.

[190] X. Zhao, P. Xie, L. Xing, G. Zhang, and H. Ma, "Clustered federated learning based on momentum gradient descent for heterogeneous data," Electronics, 2023.

[191] M. Stallmann and A. Wilbik, "On a framework for federated cluster analysis," Applied Sciences, 2022.

[192] C. Li, H. Yang, Z. Sun, Q. Yao, J. Zhang, A. Yu, A. V. Vasilakos, S. Liu, and Y. Li, "High-precision cluster federated learning for smart home: An edge-cloud collaboration approach," Ieee Access, 2023.

[193] X. Wu, A. M. Tyrrell, and Y. Ko, "Federated k-means clustering for adaptive ofdm-im," Ieee Communications Letters, 2023.

[194] H.-Y. Huang, K. Keoy, J. Chen, H. Chao, and C. Lai, "Personalized federated learning algorithm with adaptive clustering for non-iid iot data incorporating multi-task learning and neural network model characteristics," Sensors, vol. 21, no. 1, 2021.

[195] H. Kim, B. Kim, Y. Kim, C. You, and H. Park, "K-fl: Kalman filter-based clustering federated learning method," IEEE Access, vol. 11, pp. 13564–13578, 2023.

[196] L. Zhao, Q. Wang, Q. Zou, Y. Zhang, and Y. Chen, "Privacy-preserving collaborative deep learning with unreliable participants," IEEE Transactions on Information Forensics and Security, vol. 15, pp. 2292–2307, 2020.

[197] G. Arjunan, "Implementing explainable ai in healthcare: Techniques for interpretable machine learning models in clinical decision-making," International Journal of Scientific Research and Management, vol. 9, pp. 597–603, 2021.

[198] E. Goh, D. Kim, K. Lee, S. Oh, J.-E. Chae, and D.-Y. Kim, "Blockchain-enabled federated learning: A reference architecture design, implementation, and verification," IEEE Access, vol. 11, pp. 13240–13252, 2023.

[199] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.