ALFadapt: Mitigating Catastrophic Forgetting in Federated Learning

John Lucas R. P. de Sousa[1], **Eduardo Ribeiro**[1], **Lucas Bastos**[1], **Denis Rosário**[1], **Eduardo Cerqueira**[1]

[1] Federal University of Pará (UFPA)

{john.sousa, lucas.bastos, eduardo.ribeiro}@itec.ufpa.br

{denis, cerqueira}@ufpa.br

Abstract. *Catastrophic forgetting, driven by concept drift, challenges federated learning (FL), especially in dynamic environments. Existing layer-freezing methods focus on communication efficiency but often overlook concept drift's impact on individual neural network layers. We propose ALFadapt (Automatic and Adaptive Layer Freezing) to mitigate catastrophic forgetting by dynamically freezing stable layers and allowing trainable layers to adapt to evolving data. Experimental results show that ALFadapt significantly reduces accuracy loss during scenario transitions and improves model resilience when revisiting previous scenarios. This method offers a robust solution for environments characterized by concept drift and non-IID data distributions.*

1. Introduction

The rapid expansion of FL and intelligent distributed systems has created a pressing need for Machine Learning (ML) models capable of adapting to diverse and dynamic environments [Kairouz et al. 2021]. In this context, FL has emerged as a promising solution for collaborative learning from distributed data while preserving privacy [Li et al. 2024]. In the FL paradigm, clients train local models on their data and periodically share model updates with a central server, creating a globally optimized model. Recent studies, such as [Chellapandi et al. 2024] and [Malan et al. 2024], have demonstrated the efficacy of FL in the Internet of Things (IoT) and edge environments, highlighting its potential to optimize distributed systems. However, applying FL to broader distributed settings introduces unique challenges, since these environments are highly heterogeneous and dynamic, differing significantly from more stationary IoT settings.

A key challenge in FL is catastrophic forgetting, which arises primarily from concept drift. It is a phenomenon where the underlying data distribution evolves as clients encounter various conditions, such as, different usage patterns, temporal changes, and various contextual factors [Chow et al. 2023]. Unlike some traditional centralized systems that data distributions often exhibit slower shifts, FL face rapid and unpredictable changes that lead to non-independent and identically distributed (non-IID) data distribution [Babendererde et al. 2023]. This non-IID nature combined with bandwidth constraints exacerbates the difficulty of maintaining the adaptability of ML model to new conditions, while retaining critical knowledge of past scenarios [Yang et al. 2024]. These constraints limit the frequency and size of updates between clients and the central server, further amplifying the challenge [Babakniya et al. 2023].

Given these challenges, it is crucial to design a solution to effectively address these challenges, which need adaptability and the preservation of past knowledge [Ding et al. 2022]. One promising approach is selective layer freezing, which aims to balance these competing requirements. Recent study, such as [Sorrenti et al. 2023], explored selective layer freezing as a potential solution in edge-based IoT systems, demonstrating its ability to reduce communication overhead and preserve knowledge in continual learning. Selective layer freezing involves freezing layers in a ML model that capture generalizable features across scenarios, while leaving other layers trainable to adapt to new environments. However, these techniques have not been extensively applied in broader FL contexts, where dynamic and heterogeneous conditions demand more nuanced solutions [Bai et al. 2022].

In this paper, we introduce ALFadapt, an adaptive selective layer freezing method for federated learning (FL) that mitigates concept drift in decentralized, resource-constrained environments. In contrast to existing methods like FedAvg, which primarily focus on communication efficiency, ALFadapt introduces a dynamic, stability-based layer freezing mechanism that effectively mitigates catastrophic forgetting while maintaining model adaptability. By leveraging activation map similarity and gradient analysis, our method dynamically freezes ML layers to balance preserving prior knowledge and adapting to novel data patterns. Simulation results demonstrate ALFadapt's effectiveness, significantly reducing forgetting rates compared to FedAvg. While FedAvg experienced accuracy drops of 3-5% after distribution shifts, ALFadapt limited accuracy losses to 1-2% and enabled faster performance recovery, showcasing its superior adaptability in dynamic federated learning environments. Conversely, ALFadapt maintained a more stable trajectory, demonstrating superior adaptability and knowledge retention. The framework's effectiveness is particularly evident in its ability to accelerate accuracy recovery, outperforming the baseline in later training rounds and reinforcing its suitability for dynamic federated learning environments.

The remainder of this paper is structured as follows. Section 2 introduces related works that address concept drift and catastrophic forgetting in federated learning environments. Section 3 describes the fundamental theory behind this work. Section 4 details the proposed ALFadapt method. Section 5 presents a comparative study of the proposed ALFadapt method under various drift scenarios. Finally, Section 6 provides the conclusions of this work and outlines potential directions for future research.

2. Related Works

Several works have explored techniques to address catastrophic forgetting challenges, such as selective freezing for communication reduction or adaptive model updates to improve performance under drift conditions. [Yang et al. 2023] conducted a systematic study on concept drift in FL, categorizing its impact through several characteristics, such as speed, severity, and synchronism. While they highlighted the challenges posed by non-IID and dynamic data distributions, their work did not propose practical strategies to mitigate catastrophic forgetting, which is addressed by approaches like selective layer freezing.

[Jothimurugesan et al. 2023] introduced FedDrift and FedDrift-Eager algorithms to manage multiple models using hierarchical clustering and local drift detection. These

algorithms excel at handling distributed concept drift by clustering clients with similar drift patterns. However, their reliance on computationally intensive clustering algorithms limits scalability in large-scale FL deployments, making them less suitable for resource-constrained environments.

[Malan et al. 2024] proposed Automatic Layer Freezing (ALF) to optimize communication efficiency by progressively freezing stable layers during training. While ALF reduced communication overhead by up to 83.91%, its primary focus on efficiency overlooked catastrophic forgetting, which is a critical issue in heterogeneous environments where drift disrupts learned knowledge. While Malan et al. focus on communication efficiency, our approach addresses both communication efficiency and catastrophic forgetting.

[Sorrenti et al. 2023] tackled catastrophic forgetting and computational overhead in continuous learning scenarios by selectively freezing layers based on stability, effectively preserving knowledge while adapting to new tasks. However, their approach is limited to static IoT environments, and not address the unique challenges of federated learning with distributed and heterogeneous data.

[Yang et al. 2024] proposed the long-term online learning (LSTOL) framework, combining short-term learners for rapid adaptation with a long-term probabilistic controller to prevent catastrophic forgetting. While LSTOL mitigated forgetting, its reliance on ensemble-based methods increased computational overhead, making it less suitable for resource-constrained FL environments.

Based on the state-of-the-art analysis, we conclude that it is important to mitigate concept drift and preserve previously learned knowledge in a decentralized and resource-constrained environment. Table 1 summarizes the analyzed works in terms of drift analysis, layer freezing, catastrophic forgetting and dynamic FL, our framework bridges the gap between these approaches by balancing knowledge retention, adaptability, and efficiency in federated learning environments. For instance, communication-focused strategies, such as ALF, prioritize minimizing update sizes to optimize bandwidth usage. Although effective in reducing communication overhead, these methods often neglect the nuanced effects of concept drift on model performance. To the best of our knowledge, only AL-Fadapt leverages a stability-index-based selective layer freezing mechanism to directly address catastrophic forgetting while retaining secondary benefits, such as communication efficiency.

Table 1. Comparison of Related Works

Work	Drift Analysis	Layer Freezing	Catastrophic Forgetting	Dyn. FL
Yang et al., 2023	✓			
Jothimurugesan et al., 2023	✓		✓	✓
Malan et al., 2024		✓		
Sorrenti et al., 2023		✓	✓	
Yang et al., 2024	✓		✓	
Proposed Work	✓	✓	✓	✓

3. Background

FL faces significant challenges due to concept drift and catastrophic forgetting, particularly in dynamic environments where data distributions evolve over time. This section provides the concepts required to understand ALFadapt method. First, we introduce concept drift, explaining how changes in data distributions impact FL models. Afterwards, we discuss layer freezing, which is a technique to mitigate catastrophic forgetting while reducing communication overhead. These concepts lay the groundwork for ALFadapt, which builds on these principles to achieve a balance between adaptability and stability in FL settings.

3.1. Concept Drift

The concept drift in FL occurs when the relationship between input features and labels changes over time. Figure 1 illustrates the way concept drift works, starting with a certain data distribution and varying overtime. This concept drift effect can happen in two main ways: label shift and domain shift. Label shift happens when the distribution of labels $P(y)$ changes, while domain shift occurs when the distribution of input features $P(x)$ changes. The key difference is that label shift only affects how input features relate to labels, without changing the features themselves, whereas domain shift changes the actual features [Babendererde et al. 2023]. Both types of drift are important in FL environments, where clients work in diverse and constantly changing contexts, leading to data distributions that differ across clients and complicate the global model's ability to generalize effectively [Malan et al. 2023].

Label shift happens when factors like user behavior, population demographics, or changes in the environment affect the prevalence of certain classes in the data. For instance, in a mobile keyboard prediction system, the introduction of new slang terms or shifts in language use can cause some words to become more frequent, creating an imbalance in the data. These shifts lead to problems like reduced model accuracy, biased predictions favoring the more frequent classes, difficulties with model convergence, and the need for more frequent retraining. Additionally, the model may experience catastrophic forgetting, meaning it might adapt to new label distributions at the cost of forgetting previously learned distributions.

On the other hand, domain shift occurs when the input features themselves change over time, making the learned patterns less applicable. This type of shift is just as important for FL, as it can result in the model no longer being able to correctly interpret the features, even if the labels are stable.

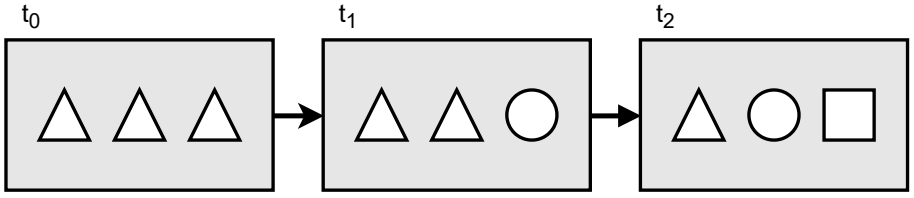

Figure 1. Concept Drift Representation

Let $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ represent the full dataset, where x_i are the input features and y_i are the corresponding labels. To study these shifts, the framework

separates the situation into two phases: a stable environment and an environment where concept drift occurs. In the initial Scenario A, the dataset $D_A \subset D$ has a predefined distribution, meaning all classes are arranged accordingly to a certain Dirichlet distribution coefficient. The data follows a stable distribution $P_A(x, y)$, where the label proportions stay the same. The FL model M is trained on this data with the aim to minimize the loss function, which ensures the model learns to classify data in a way that represents all classes equally.

The concept drift occurs when the system transitions from Scenario A to Scenario B, where the distribution of labels or features changes. This shift can happen abruptly or gradually. In the case of an abrupt label shift, the label distribution suddenly changes at a specified training round. This simulates situations where certain classes experience unexpected changes, such as a sudden surge in popularity. Mathematically, this is represented by updating the label distribution $P(y)$ to a new distribution $P'_B(y)$ using a Dirichlet distribution with a different concentration parameter α', which causes an immediate shift, as shown in Equation 1. This results in a new joint distribution $P'_B(x, y) = P'_B(y)P_A(x|y)$, which differs from the original distribution.

$$P'_B(y) = \text{Dirichlet}(\alpha') \quad \text{where} \quad P'_B(x|y) = P_A(x|y). \tag{1}$$

Gradual label shift involves a smooth and gradual change in the label distribution over several training rounds. This reflects situations like seasonal trends or changes in user preferences, where the distribution of labels slowly shifts over time. The label distribution $P_t(y)$ at time t gradually moves from $P_A(y)$ to $P_B(y)$ as a function of time t, as shown in Equation 2. In this sense, $\alpha(t)$ denotes a function that increases over time. This allows for a smooth transition from the initial to the target label distribution, helping the model adapt more gradually.

$$P_t(y) = (1 - \alpha(t))P_A(y) + \alpha(t)P_B(y), \tag{2}$$

To better analyze the impact of concept drift in FL, both abrupt and gradual label shifts need to be simulated to ensure a more accurate representation of real-world scenarios and a better understanding of their effects on model performance. This controlled approach provides insights into how different types of drift affect model performance, allowing for a structured evaluation of adaptation mechanisms. As the global model adapts to new label distributions, its performance on the original, stable distributions may decrease, leading to biased predictions that favor more frequent classes. This undermines the model's overall accuracy and raises fairness concerns, especially in cases where minority classes are important. Additionally, the dynamic nature of these shifts can create instability in the training process, making convergence more difficult and requiring frequent model updates. Hence, the risk of catastrophic forgetting becomes more pronounced in such environments, as the model must continually balance adapting to new patterns and retaining old knowledge.

3.2. Layer Freezing

Layer freezing is a strategy for FL with the aim to mitigate the concept drift in dynamic and heterogeneous environments. This approach, represented in Figure 2, freezes

the lower layers of a neural network, which capture general, scenario-independent features, while keeping the higher layers—those reflecting more specific, scenario-dependent patterns—trainable. By doing so, the model maintains stability in its core knowledge and adapts effectively to new data patterns, addressing both catastrophic forgetting and communication efficiency challenges inherent in non-stationary FL settings [Malan et al. 2024].

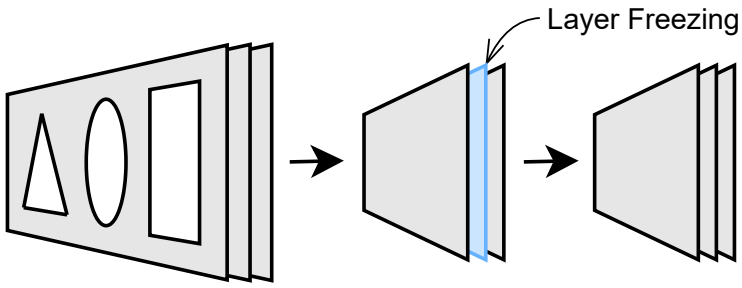

Figure 2. Layer Freeze Representation

The technique leverages the hierarchical nature of neural networks: lower layers typically learn fundamental features (*e.g.*, edges, textures, or basic shapes) that remain consistent across different environments, whereas higher layers extract more complex, context-dependent features that vary with environmental factors. For example, features learned in lower layers may be invariant between urban and rural driving, while higher layers capture more dynamic patterns, such as those influenced by weather conditions or road types.

By analyzing activation maps—representations of the neural network's output at different layers—the selective freezing process identifies which layers have consistent responses across various conditions. These stable layers, which capture generalizable features, are then frozen to maintain the model's stability, reducing unnecessary updates. Meanwhile, the upper layers remain flexible, allowing them to adapt to new, scenario-specific conditions.

This approach offers advantages over other concept drift methods by operating within a single model, avoiding the computational and storage overhead associated with ensemble techniques. By integrating selective freezing with existing FL strategies, such as FedAvg, it is possible to dynamically balance stability and adaptability. This extension of FedAvg targets both catastrophic forgetting and communication efficiency, as fewer parameter updates are required, reducing the need for frequent communication between clients and the server. In heterogeneous FL environments, selective freezing ensures robust model performance by preserving shared knowledge in the frozen layers while allowing the trainable layers to adapt to local variations. This helps prevent overfitting to dominant patterns and supports minority scenarios, improving the model's performance across a range of clients.

The efficiency gains are substantial: freezing layers reduces the number of updated parameters, thereby decreasing bandwidth usage and computational requirements, which is especially valuable in resource-constrained FL environments. However, challenges such as premature layer freezing can arise, potentially limiting the model's adapt-

ability. To successfully implement selective freezing, it is essential to carefully calibrate the freezing criteria, ensuring that the frozen layers capture stable, generalizable features. Dynamic adjustments to the freezing mechanism are necessary to strike the right balance between stability and adaptability, maintaining the model's flexibility over time.

4. ALFadapt Design

To address the critical challenges in dynamic and heterogeneous environments in FL, we introduce ALFadapt as an adaptive selective layer freezing technique designed to balance knowledge retention and adaptability while improving communication efficiency. ALFadapt incorporates a stability-based freezing mechanism, ensuring that stable layers remain unchanged while allowing trainable layers to adjust effectively to new distributions. This work is an adaptation of the technique proposed by [Malan et al. 2024], which originally focused on improving communication efficiency in FL by freezing layers. However, this adaptation extends the approach to also address catastrophic forgetting, offering a broader solution for managing both communication overhead and model adaptability in dynamic environments.

The ALFadapt framework, represented in Figure 3 and detailed in Algorithm 1, operates by dynamically evaluating the stability of each neural network layer throughout the training process. At the beginning of each round, the global model is distributed to participating clients (4), where local training is performed using their respective datasets (17-24). After training, the stability of each layer is assessed based on its historical weight changes (9), which are tracked using an Exponential Moving Average (EMA). Layers exhibiting minimal fluctuations over consecutive rounds are identified as stable and subsequently frozen to preserve previously acquired knowledge. In contrast, those experiencing substantial variations remain trainable, facilitating adaptation to new data distributions.

The stability index S_l for each layer l at a given round is determined by an EMA-based formulation, where past updates contribute to the current evaluation. The model employs a dynamically adjusted stability threshold, which serves as the decision boundary between frozen and trainable layers. This threshold is influenced by the maximum observed stability index across layers, modulated by predefined hyperparameters that control sensitivity to parameter changes. By implementing this adaptive strategy, ALFadapt effectively prevents catastrophic forgetting while ensuring that the model remains flexible enough to incorporate new knowledge.

The stability threshold τ is dynamically adjusted using the following equation, presented in 3:

$$\tau = \alpha \cdot \max(S_l) + \beta \tag{3}$$

Where the α hyperparameter control the sensitivity of layer freezing. A higher value of α increases the threshold for freezing layers, meaning layers with more fluctuation in their weights across training rounds will remain trainable and β represents the lower bound for layer freezing. It ensures that only those layers whose stability index exceeds the value are kept trainable, helping the system to avoid prematurely freezing layers.

Following local training, clients transmit only the updated parameters of trainable layers back to the central server, significantly reducing the communication overhead compared to conventional FL approaches. Upon aggregation, the server integrates these updates using the FedAvg method and recalculates the stability indices to refine the freezing criteria for subsequent rounds. This iterative process allows the framework to dynamically adjust layer freezing decisions in response to evolving data distributions, striking a balance between preserving prior knowledge and accommodating new patterns.

By freezing stable layers, ALFadapt optimizes communication efficiency without compromising the model's adaptability. The approach is particularly beneficial in bandwidth-constrained environments, where frequent updates to all layers would impose excessive transmission costs. Moreover, the method enhances model resilience by mitigating the adverse effects of concept drift, ensuring that previously learned representations are retained while still allowing for controlled adaptation.

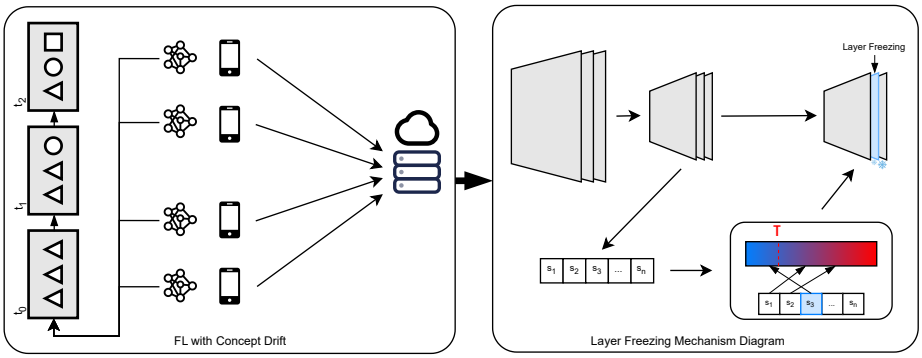

Figure 3. Framework Representation

5. Evaluation

This section presents the simulation setup employed to evaluate the performance and efficiency of ALFadapt compared to baseline approaches. We first describe the simulated FL scenario, including the underlying framework, database characteristics, and simulation parameters. We also present the obtained results, focusing on the metrics of accuracy, loss, and forgetting rate.

5.1. Setup

To evaluate the effectiveness of ALFadapt in mitigating catastrophic forgetting, we conducted simulations in a FL environment using a framework based in Pytorch that is an adaptation of previous work for [Malan et al. 2024]. Our experiments involved 100 clients, each assigned a non-IID subset of the CIFAR-10 dataset, partitioned using a Dirichlet distribution to mimic real-world federated learning settings with heterogeneous data distributions. The key parameters of our simulation setup, including the number of clients, training rounds, and data distribution strategies, are detailed in Table 2.

In each communication round, 10% of the clients participated in training. Each client performed one local epoch per round to maintain consistency in local updates. The FL training process ran for 400 rounds, incorporating domain shifts at predefined intervals (e.g., rounds 50, 100, 200, 300, and 400), introducing different levels of concept drift.

Algorithm 1 ALFadapt Overview

Require: Dataset D, Number of clients N, Global model G, Number of rounds R, Stability threshold parameters α, β, Shift intervals $\{T_1, T_2, \dots\}$, Shift transforms $\{T_1', T_2', \dots\}$, Participation rate p

Ensure: Trained global model G, Log of training metrics

Data Partitioning
Partition D into subsets $\{D_1, D_2, \dots, D_N\}$ using Dirichlet distribution for non-IID split
Assign each client i data subset D_i from Scenario A (Shift T_1')
Initialize global model G and distribute it to all clients
for each round $t = 1$ to R **do**
  if $t \in \{T_1, T_2, \dots\}$ **(Shift) then**
    Update each client's dataset to Scenario B by applying Shift transform T_k'
    for each layer l in G **do**
      Compute stability index S_l for layer l
      if $S_l < \alpha$ **then**
        Mark layer l as *frozen*
      else if $S_l > \beta$ **then**
        Mark layer l as *trainable*
      end if
    end for
  end if
  Sample $p \times N$ clients for participation
  for each participating client i **in parallel do**
    Receive global model G
    Train local model G_i on local dataset D_i
    Exclude updates for *frozen* layers
    Update only *trainable* layers using local data
    Send updated G_i to server
  end for
  Aggregate local models: $G \leftarrow \text{Aggregate}(\{G_1, G_2, \dots, G_N\})$
  Evaluate global model G on test set: Compute loss and accuracy
  Log round t metrics: Train loss, Test loss, Test accuracy, Frozen layers
end for
return Trained global model G

The data partitioning follows two different distribution strategies: a Dirichlet distribution with different concentration parameters varying at predefined intervals, simulating varying levels of class imbalance across clients, and a pathological distribution, representing an extreme case where some clients receive only a subset of the dataset classes. These distributions introduce different levels of data heterogeneity, challenging the model's ability to generalize across clients. For the domain shift part, the distributions are characterized by distinct transformation pipelines—such as standard augmentations, intense color jittering, and grayscale conversion—to the entire dataset. These transformations alter the input data distribution without changing the underlying class labels, thereby simulating covariate shifts that models might encounter in dynamic environments.

ALFadapt is evaluated against the FedAvg baseline, focusing on accuracy, forgetting rate, and model adaptability. The stability index, which determines the freezing of layers, is computed using an EMA of layer parameter changes. This strategy allows for a dynamic adaptation of the model to different levels of concept drift while preserving previously learned representations. To assess the impact of ALFadapt in comparison to FedAvg, we analyze the model's accuracy throughout training, measuring the extent to which it maintains performance despite the presence of drift. The forgetting rate is computed by comparing accuracy before and after distribution shifts, as shown in Equation 4. A positive forgetting rate indicates a decline in accuracy after a shift, signifying knowledge degradation, whereas a negative forgetting rate suggests that the model has improved post-shift.

$$\delta = \text{acc}(M_A, D_A) - \text{acc}(M_A, D_B), \tag{4}$$

Table 2. Simulation Parameters

Description	Values
Number of Clients	100 Clients
Client Participation	10%
Local Epochs per Round	1
Total Training Rounds	400
Dataset	CIFAR-10
Initial Data Distribution	Dirichlet (α=0.2)
Concept Drift Occurrence	Rounds 50, 100, 200, 300, 400
Freezing Strategy	ALFadapt
Baseline	FedAvg
Performance Metrics	Accuracy, Forgetting Rate

5.2. Label Shift Concept Drift Analysis

In the label shift scenario, the distribution of class labels was dynamically altered over time, adhering to a sequence of Dirichlet distribution parameters [0.3, 1.0, 0.1, 0.2, 0.4] at drift intervals occurring in rounds 50, 100, 200, 300, and 400. This manipulation induced varying degrees of class imbalance, forcing the model to adapt to fluctuating class prevalences.

Figure 4 illustrates the trajectory of test accuracy over the 400 training rounds under label shift conditions. The FedAvg baseline exhibited significant declines in accuracy following each drift event, underscoring the model's susceptibility to catastrophic forgetting—a phenomenon where the model loses previously acquired knowledge upon learning new information. Specifically, after each drift, FedAvg's accuracy dropped by approximately 3-5%, indicating a marked deterioration in performance on classes that were previously well-represented.

In contrast, the ALFadapt approach demonstrated remarkable resilience against such degradations. By strategically freezing layers based on entropy-derived stability

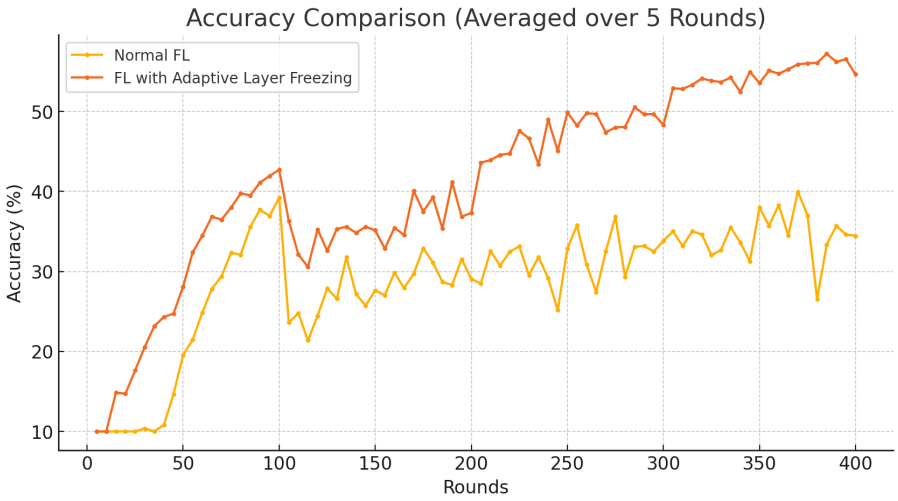

Figure 4. Accuracy under Label Shift Concept Drift

indices, ALFadapt preserved the representations learned from earlier class distributions while allowing subsequent layers to adapt to new class imbalances. This mechanism resulted in smaller accuracy drops than baseline FL post-drift, especially at round 100, effectively halving the forgetting rate in comparison, as observed in Table 3. Furthermore, ALFadapt facilitated a faster recovery in accuracy post-drift, with metrics rebounding to near-pre-shift levels faster than baseline. This swift recovery underscores ALFadapt's potential in maintaining knowledge retention without hindering the model's capacity to learn from evolving data distributions.

However, it is noteworthy that abrupt label shifts—particularly those with extreme distributional changes—still posed challenges. While ALFadapt mitigated significant accuracy losses, the model required additional rounds to stabilize and fully adapt to the new class distributions. This observation suggests that while ALFadapt enhances resilience, it may benefit from further refinements to handle sudden and drastic shifts more seamlessly.

5.3. Domain Shift Concept Drift Analysis

The domain shift scenario altered the input feature space rather than the label distribution by applying transformations in this order: color jittering, grayscale conversion, and noise addition. These modifications simulated real-world variations in data acquisition, requiring the model to generalize across different feature representations. These transformations collectively challenge the model's robustness by altering the pixel-level information while preserving the semantic content of the images. As shown in 5, both FedAvg and ALFadapt demonstrated a steady evolution in accuracy over time, adapting progressively to new distributions. However, a clear distinction emerged after the first domain shift at round 100, where ALFadapt quickly surpassed the baseline and maintained superior accuracy. At the third shift (round 200), both methods suffered a significant accuracy drop, but unexpectedly, ALFadapt experienced a sharper decline than FedAvg, as highlighted in Table 3. This result suggests that, while selective layer freezing improves adaptation, certain abrupt feature-space changes may still introduce temporary instabilities.

This temporary performance dip is likely a result of ALFadapt's dynamic stability threshold mechanism freezing layers that were not yet fully adapted to the new domain.

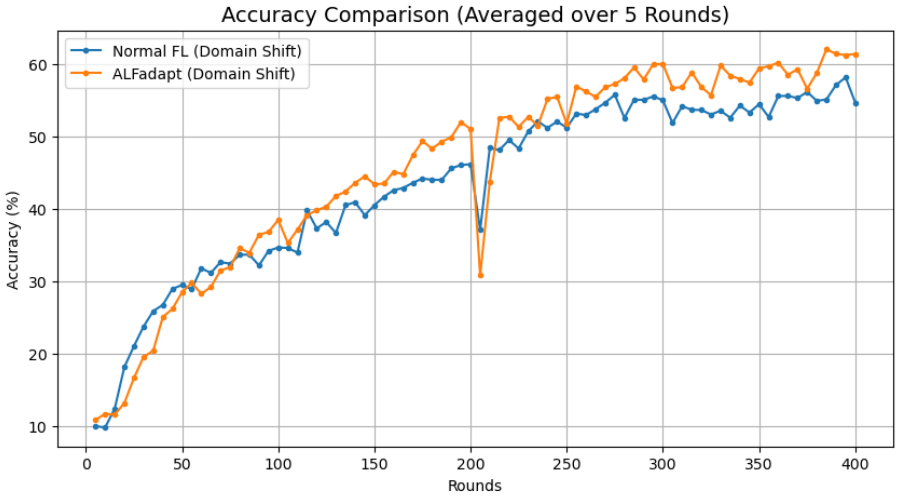

Figure 5. Accuracy under Domain Shift Concept Drift

The layer freezing mechanism in ALFadapt can cause a brief reduction in model performance during extreme domain shifts, as it may freeze layers prematurely, leading to an initial drop in accuracy. This behavior is expected, as the model needs time to unfreeze or adjust the layers to adapt to the new data distribution. Despite this unexpected performance dip, ALFadapt showcased strong resilience, recovering its accuracy within 10-15 rounds and returning to pre-shift levels faster than FedAvg. From that point onward, ALFadapt continued to outperform the baseline, maintaining higher accuracy for the remainder of the training process and finishing the 400 rounds with superior performance. Although ALFadapt was more impacted by the third shift, its ability to recover quickly and sustain a higher accuracy trajectory reinforces its effectiveness in handling domain shifts. This behavior highlights an important characteristic of the method: while short-term disruptions can occur, ALFadapt's adaptive freezing mechanism preserves knowledge effectively, allowing for faster stabilization after abrupt feature variations.

File	Round 50	Round 100	Round 200	Round 300
Label Shift Normal FL	-2.54%	16.78%	0.50%	-7.45%
Label Shift ALFadapt	-5.32%	6.49%	-5.21%	-5.38%
Domain Shift Normal FL	-1.29%	3.19%	20.19%	3.32%
Domain Shift ALFadapt	0.58%	0.05%	9.06%	3.13%

Table 3. Forgetting Rate Measured at Different Distribution Shifts

The experimental results highlight the effectiveness of ALFadapt in mitigating concept drift in federated learning. Under label shift conditions, ALFadapt consistently reduced the forgetting rate, ensuring better knowledge retention and faster recovery. Despite a more pronounced accuracy drop during the third domain shift, ALFadapt quickly rebounded and outperformed other methods in later rounds. Unlike FedAvg, ALFadapt utilizes a dynamic stability threshold to freeze layers selectively, preserving prior knowledge while adapting to evolving data. These advantages make ALFadapt more robust in dynamic FL environments, especially those with recurrent data shifts.

6. Conclusion

We presented an adaptive selective layer freezing framework for federated learning, designed to mitigate the effects of catastrophic forgetting caused by concept drift. The proposed framework dynamically evaluates layer stability using an EMA-based stability index, enabling the freezing and unfreezing of stable layers while allowing trainable layers to adapt to new data distributions. This approach demonstrates improved resilience against catastrophic forgetting by preserving previously acquired knowledge without significantly compromising model adaptability.

Our results indicate that the framework is particularly effective in scenarios with gradual or moderate concept drift, demonstrating reduced forgetting and improved knowledge retention. While it consistently outperformed the FedAvg baseline, challenges remain in handling abrupt distribution shifts, as observed in domain shift scenarios. In these cases, the framework exhibited a temporary performance drop but rapidly recovered, reinforcing its adaptability and resilience. This suggests that further refinements, such as dynamic threshold adjustments or hybrid freezing strategies, could enhance its ability to mitigate sudden changes more effectively. Overall, the proposed approach provides a robust solution for federated learning in dynamic and heterogeneous environments, where maintaining long-term model stability is crucial.

As future work, we plan to extend the framework by integrating mobility features to evaluate its effectiveness in vehicular environments. Additionally, we aim to explore hybrid approaches that combine entropy with other stability metrics to further enhance its ability to handle abrupt drift scenarios.

7. Acknowlegements

This research was partially sponsored by CNPq grant 404186/2021-1, CAPES, the Brazilian Ministry of Science, Technology, and Innovations, with resources from Law nº 8,248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex, and published under Arquitetura Cognitiva (Phase 3), DOU 01245.003479/2024-10

References

Babakniya, S., Fabian, Z., He, C., Soltanolkotabi, M., and Avestimehr, S. (2023). A Data-Free Approach to Mitigate Catastrophic Forgetting in Federated Class Incremental Learning for Vision Tasks. *Advances in Neural Information Processing Systems*, 36(NeurIPS 2023).

Babendererde, N., Fuchs, M., Gonzalez, C., Tolkach, Y., and Mukhopadhyay, A. (2023). Jointly Exploring Client Drift and Catastrophic Forgetting in Dynamic Learning.

Bai, J., Sajjanhar, A., Xiang, Y., Tong, X., and Zeng, S. (2022). FedEWA: Federated Learning with Elastic Weighted Averaging. *Proceedings of the International Joint Conference on Neural Networks*, 2022-July:1–8.

Chellapandi, V. P., Yuan, L., Brinton, C. G., Zak, S. H., and Wang, Z. (2024). Federated Learning for Connected and Automated Vehicles: A Survey of Existing Approaches and Challenges. *IEEE Transactions on Intelligent Vehicles*, 9(1):119–137.

Chow, T., Raza, U., Mavromatis, I., and Khan, A. (2023). FLARE: Detection and Mitigation of Concept Drift for Federated Learning based IoT Deployments. *2023 International Wireless Communications and Mobile Computing, IWCMC 2023*, pages 989–995.

Ding, J., Tramel, E., Sahu, A. K., Wu, S., Avestimehr, S., and Zhang, T. (2022). Federated learning challenges and opportunities: An outlook. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8752–8756.

Jothimurugesan, E., Hsieh, K., Wang, J., Joshi, G., and Gibbons, P. B. (2023). Federated Learning under Distributed Concept Drift. *Proceedings of Machine Learning Research*, 206:5834–5853.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., Hea, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konečný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Raykova, M., Qi, H., Ramage, D., Raskar, R., Song, D., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., and Zhao, S. (2021). Advances and open problems in federated learning.

Li, M., Avdiukhin, D., Shahout, R., Ivkin, N., Braverman, V., and Yu, M. (2024). Federated learning clients clustering with adaptation to data drifts.

Malan, E., Peluso, V., Calimera, A., and MacIi, E. (2023). Communication-Efficient Federated Learning With Gradual Layer Freezing. *IEEE Embedded Systems Letters*, 15(1):25–28.

Malan, E., Peluso, V., Calimera, A., Macii, E., and Montuschi, P. (2024). Automatic Layer Freezing for Communication Efficiency in Cross-Device Federated Learning. *IEEE Internet of Things Journal*, 11(4):6072–6083.

Sorrenti, A., Bellitto, G., Salanitri, F. P., Pennisi, M., Spampinato, C., and Palazzo, S. (2023). Selective Freezing for Efficient Continual Learning. *Proceedings - 2023 IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2023*, pages 3542–3551.

Yang, G., Chen, X., Zhang, T., Wang, S., and Yang, Y. (2023). An Impact Study of Concept Drift in Federated Learning. *Proceedings - IEEE International Conference on Data Mining, ICDM*, (Icdm):1457–1462.

Yang, R., Yang, T., Yan, Z., Krajnik, T., and Ruichek, Y. (2024). Preventing Catastrophic Forgetting in Continuous Online Learning for Autonomous Driving. *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5505–5512.