# Adaptive Federated Learning in Conceptually Drifting Environments

**Oguzhan Baser** [1] **Alice Zhang** [2] **Gautham Krishna Gudur** [3] **Manisha Bandi** [4]
**ob3942** [1] **az8797** [2] **gg32849** [3] **mb65958** [4]

## Abstract

In the domain of Federated Learning (FL), the nuanced challenges arising from Concept Drift (CD) within dynamic environments, exemplified by real-world healthcare systems, present a compelling technical quandary. FL, positioned at the intersection of privacy-preserving machine learning and decentralized data collaboration, grapples with the ever-changing data distributions inherent in applications like hospital databases. The existing paradigms fall short in their ability to seamlessly detect and adapt to CD in the federated context, impeding FL's capacity to sustain model accuracy amid evolving medical practices, patient demographics, and diagnostic technologies within hospital datasets. This lacuna necessitates cutting-edge methodologies that not only facilitate secure collaboration among decentralized learners but also enable their agile adaptation to the shifting data dynamics endemic to individual local datasets. This paper embarks on an exploration of pioneering techniques, aiming to fortify FL against CD in real-world applications, thereby contributing to the advancement of adaptive and privacy-preserving machine learning, a quintessential pursuit in data science.

## 1. Introduction

In the dynamic landscape of machine learning, Federated Learning (FL) takes center stage, seamlessly integrating privacy-preserving techniques with decentralized data collaboration. This paper addresses the intricate challenges posed by Concept Drift (CD) within real-world applications, spanning domains such as healthcare systems and voice command recognition. CD, a phenomenon in which the statistical characteristics of a target variable change over time, introduces a complex layer of dynamism to machine learning models. Traditional FL methodologies, rooted in the assumption of static data distributions, face substantial hurdles when confronted with the dynamic label distributions intrinsic to evolving datasets experiencing CD. The ability to adapt to these shifting data dynamics is crucial for sustaining model accuracy amid changes in medical practices, patient demographics, and diagnostic technologies within hospital databases or fluctuations in voice command patterns in voice recognition systems.

In response to these challenges, our work introduces innovative techniques tailored to fortify FL against CD, emphasizing not only the imperative for secure collaboration among decentralized learners but also their agile adaptation to the shifting label dynamics within individual local datasets. The proposed streamlined architecture for 2D data feature extraction and classification, coupled with an adaptive learning rate mechanism, emerges as a robust solution to enhance model adaptability in the face of CD. Experimental validations on datasets like Google Speech Commands and CIFAR-10 underscore the efficacy of our methodology, marking significant strides in advancing adaptive and privacy-preserving machine learning amidst the complexities introduced by CD.

## 2. Background

In the statement of the problem (Sec. 3), we expound upon two fundamental machine learning paradigms, which constitute the foundational framework underpinning our endeavor. In this section, we explain these two concepts.

### 2.1. Federated Learning

FL is a privacy-preserving distributed machine learning scheme introduced by (McMahan et al., 2017). In the FL pipeline, an edge device $e_i$ has its own private dataset $\mathcal{D}_{e_i}$ consisting of $m_{e_i}$ number of data samples $x_k$ (images or text) and corresponding labels $y_k$, as formulated in:

$$(x_k, y_k) \in \mathcal{D}_{e_i}, \quad \forall k \in \{1, 2, ..., m_{e_i}\}. \tag{1}$$

In communication round $t$, the cloud $c$ distributes the current model weights $w_c^t$ to a total of $N_e$ collaborative edge devices. Each edge device $e_i$ has a classification model $\hat{y}_k = f(x_k, w_c^t; g_{e_i}^t)$, initialized with the cloud's weight update $w_c^t$ for round $t$. The model $f$ takes data sample $x_k$, and produces a prediction $\hat{y}_k$ based on its weights $w_c^t - g_{e_i}^t$. At round $t$, each edge device $e_i$ estimates the optimal gradient update $g_{e_i}^{t*}$ that minimizes a pre-determined loss function $\mathcal{L}$ defined by:

$$\mathcal{L}(\mathcal{D}_{e_i}, w_c^t; g_{e_i}^t) = \frac{1}{m_{e_i}} \sum_{k=1}^{m_{e_i}} \mathcal{L}(y_k, f(x_k, w_c^t; g_{e_i}^t)). \tag{2}$$

This loss function is typically calculated by averaging the cross-entropy loss over each label and corresponding prediction pair $(y_k, \hat{y}_k)$ in the private dataset $\mathcal{D}_{e_i}$. With its private dataset, each edge device $e_i$ trains the distributed model by:

$$g_{e_i}^{t*} = \arg\min_{g_{e_i}^{t}} \mathcal{L}(\mathcal{D}_{e_i}, w_c^t; g_{e_i}^t). \tag{3}$$

Then, all edge devices $e_i$ upload their optimized gradients $g_{e_i}^{t*}$ to the cloud $c$. The cloud $c$ aggregates all edge gradients $g_{e_i}^{t*}$ and updates its old weights $w_c^t$ with a certain learning rate $\eta$ in order to get the next round's central model $w_c^{t+1}$ by:

$$w_c^{t+1} = w_c^t - \frac{\eta}{N_e} \sum_{i=1}^{N_e} g_{e_i}^{t*}. \tag{4}$$

This iteration continues until accomplishing a convergence on the model performance.

### 2.2. Concept Drift

In various FL contexts (Ma et al., 2022; Kairouz et al., 2021), the definition of concept and CD can differ. Here, we focus on a definition based on an FL system with $N$ edge clients collaborating with the cloud over communication rounds $t$.

*Definition: The concept for client $c$ at time $t$ with the private dataset $\mathcal{D}_c^t$ comprising data $x$ and label $y$ pairs is the distribution denoted as:*

$$\mathcal{P}_c^t(x, y) : \mathcal{D}_c^t \sim \mathcal{P}_c^t(x, y) \tag{5}$$

*Definition: CD occurs for FL client $c$ at communication round $t + 1$ if:*

$$\mathcal{P}_c^t(x, y) \neq \mathcal{P}_c^{t+1}(x, y) \tag{6}$$

CD is the occurrence where the statistical characteristics of a target variable trained with a given machine learning model, change over time. It implies that the association between input features and the target variable is not constant but evolves. This can lead to a decline in the performance of machine learning models trained on historical data due to changing assumptions. Factors contributing to CD include alterations in data distribution, label/target distribution, shifts in feature-target relationships, and variations in the data generation process. Addressing CD is crucial in real-world applications, demanding model adaptation to sustain predictive accuracy amid evolving data. In this work, we address the challenge of label drift, which is the change in the distribution of dependent/target variables and has been relatively less explored compared to covariate/data shifts and shifts in feature-target relationships.

CD in FL occurs when participating clients alter their distributions (label distributions in our case), driven by multiple phenomena pertaining to shifts in users' habits. A model with new input data points which are conceptually drifting from the original input data is typically trained across multiple epochs. This process is repeated across multiple FL iterations (communication rounds) with label drift occurring across multiple user clients.

## 3. Problem Statement

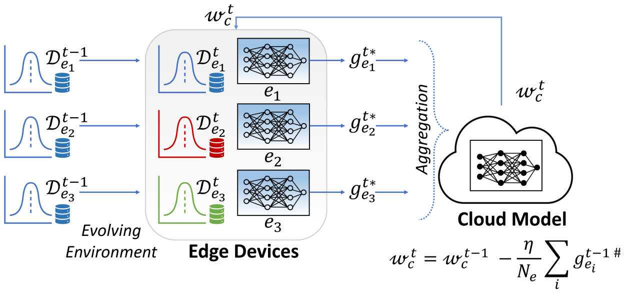

*Figure 1.* Dynamic FL System in CD Environments

Figure 1 illustrates the dynamic nature of our FL system. The cloud model sends model weights $w_c^t$ for round $t$ to edge devices $e_i$. Each edge device $e_i$ trains its model initialized with $w_c^t$ on its training dataset $\mathcal{D}_{e_i}^t$ and provides gradient updates $g_{e_i}^{t*}$ to be aggregated in the cloud for future rounds. Unlike traditional FL approaches that assume fixed data distributions, our work explores the system's response to changing data distributions (represented by the transition from blue to red and green) across evolving FL iterations and examines the system's adaptability to these changes.

In the dynamic landscape of FL, CD presents a significant challenge due to diverse and evolving data sources across decentralized devices. Shifts in local environments, user behaviors, or device-specific factors cause subtle yet impactful changes in data properties (Kairouz et al., 2021). Detecting and addressing CD is essential for FL systems to maintain model accuracy amidst evolving data patterns, so monitoring gradient update variance is crucial in that regard. A sharp increase signals a significant deviation in local data patterns among devices, possibly indicating CD. Devices experiencing data distribution shifts may produce high-variance gradients, showcasing divergence from the established global model. Additionally, covariance between gradient updates is vital. Positive covariance indicates similar trends, reinforcing the model. Negative covariance suggests diverse learning patterns among devices, possibly indicating CD. The second-order moments of gradient updates is defined as following.

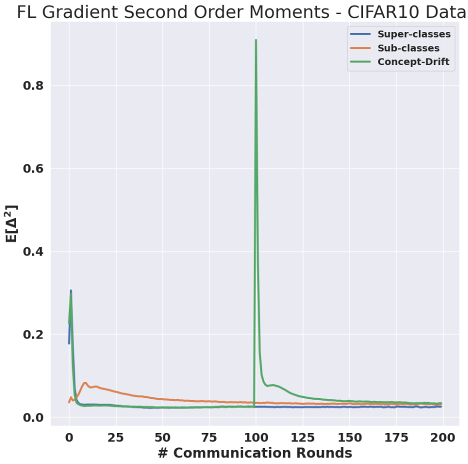

*Figure 2.* The Second Order Moments of CIFAR-10 Gradient Updates

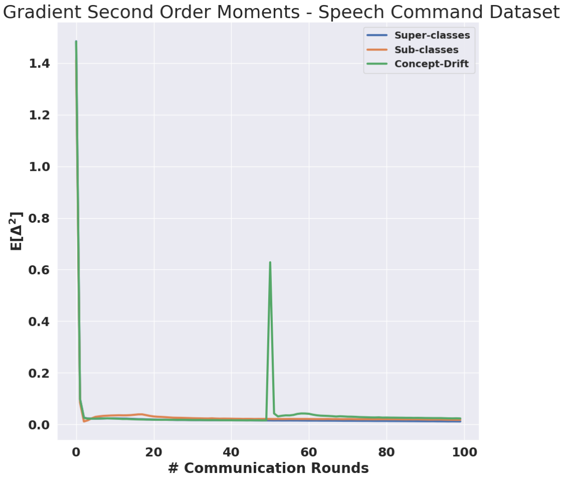

*Figure 3.* The Second Order Moments of Google Speech Command Updates

## 4. Methodology:

From the cloud's perspective, there are limited observations and corresponding measures that can be taken. For instance, we lack access to the data itself and, consequently, cannot monitor changes in its distributions. Our analysis is restricted to the aggregated gradient updates originating from the edge devices. As these gradients are random vectors for each communication round, our analysis is conducted in terms of expectations.

*Definition: Second order moment for the gradient update $\Delta_c^t$ of the client $c$ at the communication round $t$ is calculated as:*

$$\mathbb{E}[\|(\Delta_c^t)^2\|]. \tag{7}$$

We conducted an analysis of the second-order moments of gradient updates originating from edge devices in three settings: super-class-only training, subclass-only training, and CD. The results for the CIFAR-10 dataset with CD at communication round 100 are illustrated in Figure 2, while the results for the Speech Command dataset with CD occurring at communication round 50 are depicted in Figure 3. In both cases, a significant increase in the second-order moments of the gradients is observed at the communication round where CD occurs, unlike the settings of superclasses-only training and subclasses-only training. We intend to leverage this observation to implement a mitigation mechanism by adjusting the learning rate accordingly. Additionally, there are high values of second-order moments at the initial rounds of FL, but these are ignored as they are mainly a result of the initiation of FL, during which gradients deviate from random initialization.

Integrating the second-order moments of gradient updates (Panchal et al., 2023) enables FL systems to adapt to evolving data environments. Algorithms can trigger retraining when variance or covariance exceeds thresholds. These moments also influence gradient update weighting during aggregation, ensuring CD-affected devices are appropriately accounted for in the global model update. This approach not only boosts FL model accuracy and adaptability but also enables autonomous recognition of data pattern shifts, facilitating timely responses to CD in decentralized and evolving data landscapes.

*Definition: The adaptive learning rate $\eta^t$ for communication round $t$ is determined by the second-order moments for the gradient updates $\mathbb{E}[\vec{\Delta}^2]$ and is parameterized by the weighting value between the previous learning rate and the second-order moments, denoted as $\beta$, and the normalization value of the moments, denoted as $\gamma$. The formulation is as follows:*

$$\eta^t = \eta(\mathbb{E}[\vec{\Delta}^2]; t, \beta\gamma) = \beta\eta^{t-1} + \frac{(1-\beta)\mathbb{E}[\|(\Delta_c^t)^2\|]}{\gamma}. \tag{8}$$

This adapted learning rate enables the cloud aggregator of the FL system to increase its learning rate in response to a significant increase in the second-order moments of the aggregated gradient updates and, conversely, decrease the learning rate as needed.

## 5. Data Collection and Description

Since the main goal of this project is to demonstrate CD in FL in real-world scenarios, we established the following criteria to select datasets for this project.

1. *Rooted in Reality:* The selected dataset should have real-world applications, augmenting its practical sig-

nificance.

2. *Preserving Privacy:* There should be existing concerns regarding the privacy of the sensors used to collect the data, which motivates the use of FL over centralized training.
3. *Hierarchical Structures:* We want to logically construct larger classes, which we call *superclasses*, from the existing class labels in the dataset. It is essential to construct appropriate CD simulations based on CD definitions.

We selected the `Google Speech Commands` dataset following these criteria in addition to the `CIFAR-10` as a baseline dataset to realize our experiments.

**Speech Commands Dataset** (Warden, 2018) This dataset contains 105,829 utterances of 35 words by 2,618 different speakers, originally intended for the development and evaluation of keyword spotting systems. We only use the ten keywords that can be used as commands in robotics or activity assistants including ”Yes”, ”No”, ”Up”, ”Down”, ”Left”, ”Right”, ”On”, ”Off”, ”Stop”, and ”Go.” Plots of samples of raw audio signals from this dataset are shown in Figure 4. *Rooted in Reality*: Keyword spotting such as ”Hey Google” is commonly used to wake and trigger voice assistants. Furthermore, the speech comes from a wide range of speakers representative of a diverse population. *Preserving Privacy*: Speaker identity and biometrics can be determined through features extracted from spoken utterances (Nautsch et al., 2019). Commonly, voice activity assistants are connected to a cloud server that processes received audio commands, which exposes the speaker's voice and spoken content to potential adversaries who can extract the speaker's identity for malicious purposes. Thus, training keyword recognition systems on local edge devices instead of a centralized server can mitigate potential privacy leakages of speaker identity and provides a motivation for FL on this dataset. *Hierarchical Structures*: The labels of this dataset can be grouped together to form superclasses. For instance, a category of directional commands can be constructed from utterances related to directions such as ”Up”, ”Down”, ”Left”, and ”Right.”

**CIFAR-10** (Krizhevsky et al.) This dataset contains 60,000 images split evenly into 10 classes including ”Airplane”, ”Automobile”, ”Bird”, ”Cat”, ”Deer”, ”Dog”, ”Frog”, ”Horse”, ”Ship”, and ”Truck.” *Rooted in Reality*: This dataset contains 10 classes of natural images, making it representative of real-world image recognition tasks in comparison to other datasets like MNIST, which only contains images of digits (Deng, 2012). *Preserving Privacy*: Camera privacy is a real-world concern due to the amount of information a user can discretely capture with a camera. *Hierarchical Structures*: These class labels can be grouped into larger superclasses. For instance, a mammal superclass can be constructed from images in the ”Cat”, ”Deer”, ”Dog”, and ”Horse” classes. Overall, CIFAR-10 is commonly used for evaluation and comparison of results in computer vision, and similarly, we use the dataset to evaluate our system on one of the most popular image datasets.

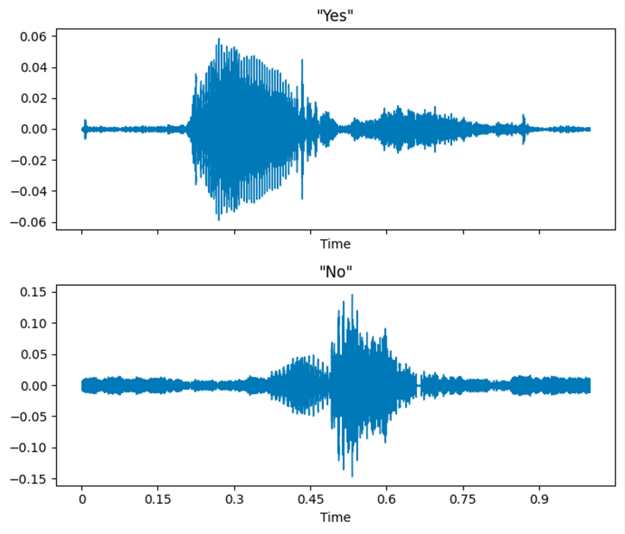

*Figure 4.* Raw waveform of the utterances “Yes” and “No.”

## 6. Data Pre-Processing and Exploration

### 6.1. Feature Engineering

We extracted log-Mel power spectrograms from the raw waveforms as input features to our classification model. Log-Mel power spectrograms are like spectrograms in that they indicate when certain frequencies and their magnitudes are present in a signal. However, they are different from normal spectrograms in that they also take into consideration the human perception of pitch in which humans are more sensitive to changes in low frequency sounds than high frequency sounds.

The calculation of log-Mel power spectrograms is illustrated in Figure 5 and is as follows: To prepare for computation of log-Mel power spectrograms from the input audio, utterances that are less than one second in duration are zero padded until their duration is one second. From the raw audio, we compute the short-time Fourier Transform (STFT) with a window length of 25ms and a hop size of 1ms to localize in time the frequency components of the input speech. We calculate the magnitude of the STFT and then multiply the power spectrogram by Mel filter banks, which incorporates human perception of pitch into the spectrogram. Low frequencies are exaggerated and high frequencies are compressed to reflect humans' greater sensitivity to low frequencies compared to high frequencies. Taking the log of the Mel power spectrogram yields the log-Mel power spectrogram. An example of log-Mel power spectrograms calculated for ”Yes” and ”No” utterances are shown in Figure 6.

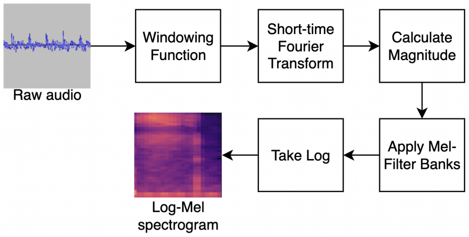

*Figure 5.* Computational pipeline to obtain log-Mel spectrograms.

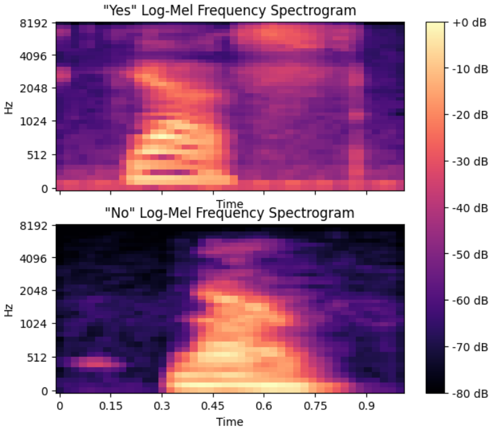

*Figure 6.* Log-Mel spectrograms of the utterances “Yes” and “No.” Generally, low frequency content corresponds to vowels, while high frequency content corresponds to consonants.

# 7. Learning/Modeling

## 7.1. Neural Network Architecture

We tailored a PyTorch architecture for 2D data feature extraction and classification tasks, embodying a streamlined architecture with two convolutional layers and two fully-connected layers. The initial convolutional layer, `conv1`, utilizes a 3x3 kernel to extract 32 feature maps, followed by max-pooling (`pool`) with a 2x2 kernel for spatial dimension reduction. The subsequent convolutional layer, `conv2`, employs 64 filters with a 3x3 kernel for further feature refinement. Flattening the feature maps, the first fully-connected layer `fc1` with 64 neurons transforms spatial features into a compact representation. The final fully-connected layer `fc2` produces the model output, a vector of size `output size` representing class probabilities. Throughout the architecture, Rectified Linear Units (`ReLU`) introduce non-linearity. This design strikes a balance between model expressiveness and computational efficiency, making well-suited for diverse 2D classification tasks.

## 7.2. CD on Real Data:

In our experiments, we use super classes and sub-classes to showcase the effect of label drift (CD). We group the sub-classes into various logical super classes so that the label distribution of multiple sub-classes changes. We initially start our model training with the super classes for the first 50 communication rounds, and then the sub classes for the rest of the communication rounds, thereby emulating CD.

In the Google Speech Commands Dataset, we use ‘Binary Commands’, ‘Directions’ and ‘Commands’ as super classes, and ‘On’, ‘Off’, ‘Yes’, ‘No’, ‘Left’, ‘Right’, ‘Up’, ‘Down’, ‘Stop’, ‘Go’ as the sub-classes/original classes. Similarly in CIFAR-10 dataset, we use [$< 3$, $3 <= 5$, and $> 5$] as super classes, and [0,1,2,3,4,5,6,7,8,9] as the original sub-classes.

We use a 2-layered CNN in our models since it is widely used across literature for FL keyword spotting tasks, and also for FL image tasks. It also enables the model to not overfit to the initial super classes and train with sub-classes with CD across multiple epochs and FL iterations/communication rounds.

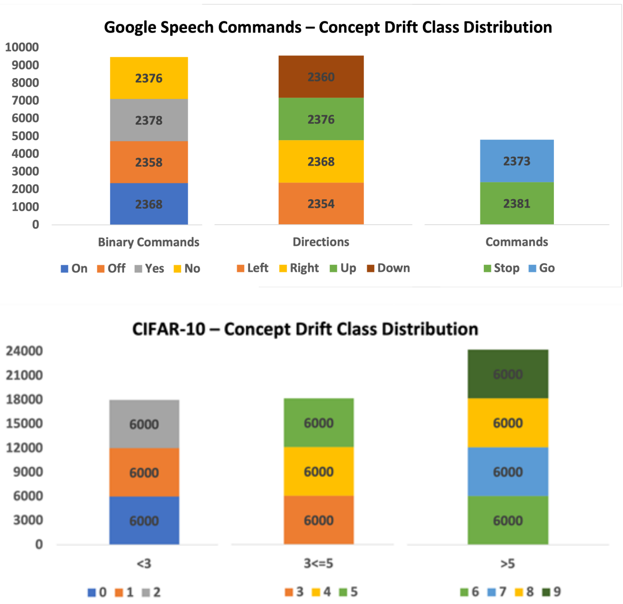

*Figure 7.* The Sub-class&Super-class Histogram of Google Speech Commands Dataset and CIFAR-10 Dataset

# 8. Experimental Results

## 8.1. How well Adaptive FL solves the problem:

In our study, we conducted a series of simulations to investigate various aspects of FL. Specifically, we explored scenarios involving super-classes exclusively, subclasses exclusively, CD from super-classes to subclasses, and adaptive FL configurations. These experiments were performed on both the CIFAR-10 dataset and the Google Speech Commands dataset, as illustrated in Figures 8 and 9.

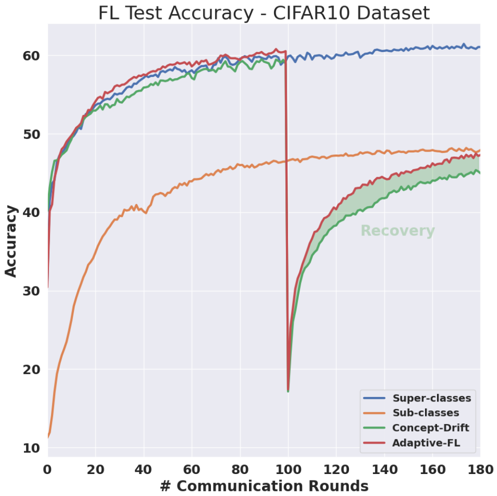

*Figure 8.* The Improvements on CIFAR-10 CD

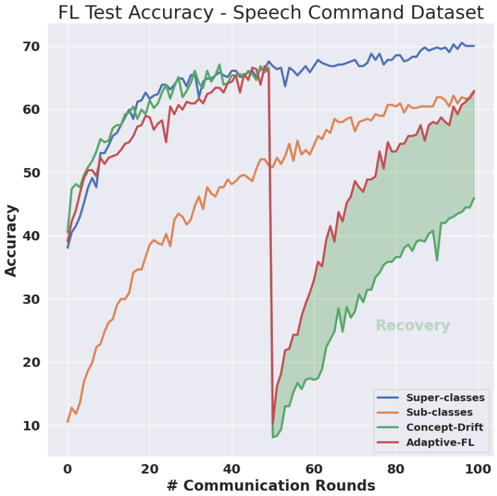

*Figure 9.* The Improvements on Google's Speech Command CD

A noteworthy observation pertains to instances of CD, where a transition occurs from super-classes to sub-classes features. Typically, the final converged accuracy is expected to align closely with the super-classes only case. However, our findings reveal that, in the presence of CD, this alignment is not achieved. To address this challenge, our proposed method incorporates an adaptive mechanism that increases the learning rate during global updates to accommodate drifting concepts, as depicted by the red curve. This approach leads to significant enhancements in performance following CD (green region) and attains a sub-class level of performance upon convergence, as indicated by the convergence of the red and orange curves. These results underscore the effectiveness of our proposed method, showcasing its ability to adapt to CD and generalize across diverse datasets.

# 9. Conclusion

In this study, we delved into the realm of FL with a specific focus on addressing the challenges posed by CD in practical scenarios. Our investigation has unveiled the pivotal role of gradients as effective indicators of changes in data distributions, particularly in the context of CD. Leveraging this insight, we introduced a streamlined 2D data feature extraction and classification architecture within the FL framework. Augmented by an adaptive learning rate mechanism, our methodology proves robust in enhancing model adaptability to dynamic shifts in underlying data, as exemplified by CD. One can regenerate the results or analyze the hyperparameters at `https://github.com/oguzhan-baser/adaptiveFL`.

The key takeaways underscore the significance of gradients, the pragmatic use of increased learning rates for adaptation, and the generalizability of our results across diverse datasets and models. Gradients, as reliable indicators of CD, provide valuable insights for detecting and adapting to evolving data dynamics within FL. The computational efficiency of increasing learning rates, while suboptimal, emerges as an effective and practical strategy for addressing CD.

However, critical lessons and avenues for future exploration have also surfaced. Increasing the number of edge users emerges as a crucial factor for refining FL behavior and understanding adaptivity. Future work should focus on scalability to larger and more diverse FL settings. Additionally, while our results showcase promising mitigations, avenues for achieving enhanced solutions remain an open question. Balancing the trade-off between acquiring more accurate second-order moments and minimizing waiting time is crucial. Future research should optimize this balance, refine our methodology, and explore novel techniques to improve its effectiveness in addressing CD within FL systems.

In conclusion, our work contributes to advancing adaptive and privacy-preserving machine learning. The technical insights gained and future exploration directions highlight the evolving nature of FL research, providing a foundation for resilient and adaptable FL systems in dynamic real-world scenarios.

## References

Deng, L. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

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., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konecný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Qi, H., Ramage, D., Raskar, R., Raykova, M., 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. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021. ISSN 1935-8237. doi: 10.1561/2200000083. URL http://dx.doi.org/10.1561/2200000083.

Krizhevsky, A., Nair, V., and Hinton, G. Cifar-10 (canadian institute for advanced research). URL http://www.cs.toronto.edu/~kriz/cifar.html.

Ma, X., Zhu, J., and Blaschko, M. B. Tackling personalized federated learning with label concept drift via hierarchical bayesian modeling. In *Workshop on Federated Learning: Recent Advances and New Challenges (in Conjunction with NeurIPS 2022)*, 2022. URL https://openreview.net/forum?id=RBPvr4Ehojh.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and Arcas, B. A. y. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Singh, A. and Zhu, J. (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Nautsch, A., Jiménez, A., Treiber, A., Kolberg, J., Jasserand, C., Kindt, E., Delgado, H., Todisco, M., Hmani, M., Mtibaa, A., Abdelraheem, M. A., Abad, A., Teixeira, F., Driss, M., Gomez-Barrero, M., Petrovska-Delacrétaz, D., Chollet, G., Evans, N., Schneider, T., and Busch, C. Preserving privacy in speaker and speech characterisation. *Computer Speech  Language*, 58, 11 2019. doi: 10.1016/j.csl.2019.06.001.

Panchal, K., Choudhary, S., Mitra, S., Mukherjee, K., Sarkhel, S., Mitra, S., and Guan, H. Flash: Concept drift adaptation in federated learning. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26931–26962. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/panchal23a.html.

Warden, P. Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition. *ArXiv e-prints*, April 2018. URL https://arxiv.org/abs/1804.03209.